
LLMs are not stochastic parrots: Evidence for meaning-mediated abstraction from conlang-like tasks

Julia Witte Zimmerman^{*,†,‡}

Calla G. Beauregard^{*,‡}

Tabia Tanzin Prama^{*}

Parisa Suchdev[†]

Kathryn Cramer^{*}

Elisabeth Kollrack^{*}

^{*}Computational Story Lab

[†]Computational Ethics Lab

[‡]Equal contribution (first project members)

Vermont Complex Systems Institute

University of Vermont

Burlington, VT 05405, USA

Abstract

The strong version of the *stochastic parrot* argument claims that, although large language models (LLMs) may exceed rote regurgitation, they cannot move beyond statistical pattern matching into abstraction or reasoning, remaining ontologically near the lower bound of pattern reuse despite producing alluringly fluent text. We test this hypothesis using conlang-like tasks. Several LLMs are given only natural-language descriptions of fictional languages that subvert prominent superficial patterns in training data by combining statistically uncommon and unattested features. Crucially, no example outputs are given. We argue that if the models exhibit rule-following behaviour, they cannot be relying solely on superficial statistical patterns; such patterns often work against the correct output. Instead, successful performance requires representations of the constraints specified in the prompt. Across three complementary task families, models systematically move in the meaning-predicted direction: they distinguish prompt exposure from instructed use, alter semantic relationships in response to novel constraints, and sometimes produce exact matches to complex translation answer keys. Although performance varies across the spectrum of models used, these results provide evidence for meaning-mediated abstraction in LLMs and refute the strong stochastic parrot hypothesis. Our work shows that, under appropriate architectural and contextual constraints, statistical learning can produce meaning-mediated abstractions, although generation remains strongly constrained by superficial plausibility. We discuss implications for model development and for understanding how increasingly abstract representations may emerge from plausible-text-generation objectives.

^{*}julia.zimmerman@uvm.edu

1 Introduction

This pre-print describes the initial results of an empirical probe of a widespread claim about LLM limitations. It has been created to provide access to our findings to participants at ISC Summer School 2026, where we presented this project as a poster, as well as IC2S2 2026 and SLSA/ 4S 2026, where we presented² related works [Institut des sciences cognitives, Université du Québec à Montréal, 2026, 12th International Conference on Computational Social Science, 2026, Society for Literature, Society, and the Arts, 2026, Society for Social Studies of Science, 2026].

In this work, we test the strong version of the *stochastic parrot* hypothesis using conlang-like tasks³ where following the meaning of an instruction conflicts with familiar surface distributions. This project is currently underway with a suite of 14 conlang-like rules (17 sets of prompts, with 2 non-conlang controls and 1 correction), but we present our initial results focused on a subset of these rules and a subset of our planned analyses as part 1 of 2.

A brief summary of the **key points** in this pre-print:

- Statistical plausibility and meaning-mediated plausibility can be experimentally dissociated. Our three conlang families dissociate plausibilities in different ways: in brief, prompt exposure versus instructed use, semantic similarity versus instructed irrelevance, and relational preservation across radically different surface forms.
- Across three independent rule families LLMs moved in the meaning-predicted direction, showing that superficial statistical patterns are not a sufficient explanation for their behavior even though such patterns remain powerful constraints. For every rule, at least one model shows clear evidence of rule-following; for most rules, most models do.
- The strong stochastic parrot hypothesis is behaviorally inadequate, even for base-model LLMs, because successful performance on these tasks is better explained by meaning-mediated abstraction than by superficial statistical regularities.
- We provide an existence proof of meaning-mediated abstraction implicated through rule-following behaviour, not evidence of human-equivalent understanding.
- Plausible output alone does not reveal its mechanism: Benchmark success can come without construct validity.
- Next-token prediction may give models access to meaning while simultaneously anchoring generation disproportionately to its most superficial and fungible forms. The bottleneck may not be knowledge acquisition, but control over which kind of plausibility governs generation.

1.1 Why this mental model matters

As with the introduction of prior information technologies – photography, printing, even writing – the reaction to Generative AI (Gen AI) is divergent and contentious due to its destabilizing effects amongst a complex milieu of motivations, beliefs, and constraints [Yates, 1966, Zimmerman, 2025]. To contend with these disruptions, people develop mental models (including specific phrases and associations) which they draw on when answering questions as to Gen AI’s behavior, origin, and capacity, in order to meaningfully interpret it [Rapp et al., 2025, Grice, 1989, Agha, 2003, Eckert, 2008, Silverstein, 2003, Peirce, 1931, Gibson, 1979].⁴ A similarly complex milieu shapes how the components of those mental models develop and endure [Irvine and Gal, 2000]. We believe good mental models are a necessary precursor to effectively navigating this fraught landscape (independent of stance).⁵

Regardless of the original intent behind it, the term *stochastic parrot* has been widely and variously adopted since its introduction in Bender et al. [2021]. It is now a popular mental model for Gen AI, and is frequently applied even to models that differ substantially from those discussed in the original paper.

²Or will present, in the case of SLSA/ 4S 2026.

³Constructed languages, or conlangs, are communication systems intentionally designed by someone (or something, or someones). For brevity, we refer to our conlang-like tasks as conlangs, tasks, or rules.

⁴Although not the subject of this paper, we have two in-progress works related to this topic [Zimmerman et al., 2026, Morgenstern et al., 2026].

⁵Bad mental models have multiple harms. They can lead to incorrect predictions, act as thought-terminating clichés, undermine an otherwise compelling argument in the eyes of the intended audience, and cause disagreement where none was necessitated.

In this work, we are not concerned with the term itself, but with empirically evaluating a claim it is often taken to imply: that LLMs trained on text are fundamentally limited to relatively superficial statistical regularities and therefore incapable of forming the kinds of abstractions associated with meaning-like or thought-like behavior.⁶ For our purposes, the strong stochastic parrot hypothesis (SSP) is the view that model behaviour can ultimately be explained in terms of token co-occurrence statistics, surface-level syntactic patterns, and similar symbolic patterns. By contrast, we argue that, given the nature of the correct outputs in our tasks, explanations invoking less superficial aspects of meaning are substantially more parsimonious.

Rather than treating statistical learning and abstraction as mutually exclusive possibilities, we ask under what conditions statistical learning can give rise to increasingly abstract behavior.⁷ We argue that understanding how statistical learning scaffolds transitions along the concrete-abstract pattern spectrum is important for understanding AI systems as well as human cognition.⁸ Dismissing statistical learning as inherently incapable risks obscuring mechanisms through which meaning can emerge.

We contend that if models systematically move in the direction predicted by our tasks, then explanations invoking only superficial symbolic regularities become increasingly difficult to maintain.

1.2 What is plausibility?

Despite significant variation, LLMs are fundamentally built around plausible text generating architectures, comprising plausible text generating objectives (such as masked language models like RoBERTa [Liu et al., 2019] and autoregressive transformers like ChatGPT [OpenAI et al., 2024]) and token-centric design choices. Specifically, next-token-prediction (NTP) is the popular architectural choice underlying the models we utilize here. For concision, we’re referring back to the family of neural architectures united by this fundamental learning strategy as plausible-text-generating-architectures (PTGA).

These models optimize for plausibility by repeatedly guessing what token comes next and being rewarded for good guesses, based on a large corpus of mostly human-generated text. In that operationalization of plausibility, everything that goes into making someone likely to continue the utterance in that particular way is collapsed into a single score.⁹

Speaking approximately, many determinants of plausibility fall under what we might call the *meaning* of the utterance. Some are more directly perceptible via statistical learning than others, with the most accessible to an LLM being superficial statistical regularities such as token co-occurrence. Other determinants may require increasingly abstract representations, which may themselves be assembled from lower-level patterns. Humans appear capable of learning across this spectrum, both from concrete statistical regularities and from contextually-mediated abstractions.¹⁰

Plausibility alone is lossy. When an LLM generates plausible text, it is unclear whether that is a result of the model’s understanding of the relevant concepts – any part of what might be called the *meaning*

⁶This mental model is prevalent. Beyond popular descriptions such as “spicy autocorrect,” we have encountered similar claims in academic settings, e.g. a comparison of deep learning to photographs that assemble meaningful patterns via photons without interacting with that meaning beyond its most superficial manifestation.

⁷And potentially increasingly thought-like, connecting meaning construction across architectures and scales, drawing on the Distributional Hypothesis (DH) as a partial explanation for similar processes in people. Although LLM architecture is extremely different from ours, the DH is explicitly the inspiration for the distributional-semantics-related aspects of LLM architecture as well [Sahlgren, 2008, Harris, 1954, 1968]. In this work, our aim is not to demonstrate equivalence between human and machine cognition – or to argue that machines can “think” – but to evaluate a more specific claim about the limits of statistical learning.

⁸Language operates across multiple levels of abstraction, from highly fungible surface forms to deeply contextual and socially grounded interpretations. See discussion of the pattern-to-thought transition and the fungibility-groundedness spectrum in Zimmerman [2025]. See also upcoming works with Grayson Wycliffe Storer, Alice Patania, and Juniper L. Lovato.

⁹Determinants such as world knowledge, mood, past experience, history with the speaker, situation, pragmatics, prosody, semantics, tone, location, physical ability (for example, ability to hear in a noisy room), propositional content, implications, consequences, belief, truth, commonsense, motivation, personality, attentiveness, familiarity, memory, etc. On the fungibility-groundedness spectrum, the most personal of these factors are unique, undetectable as a pattern in the text itself, the stigmergic trace of the communication. The more fungible these factors are, the more likely they are to be shared due to commonalities like human embodiment and culture, the more likely their signal is to be perceptible to a statistical learner. Note that the specifics of this are clearly embodiment-/ architecture-dependent!

¹⁰More discussion of this topic is available in Zimmerman [2025]. For an example of how medium and message are inherently intertwined and both decoding and encoding are architecture-dependent, see Pera et al. [2026].

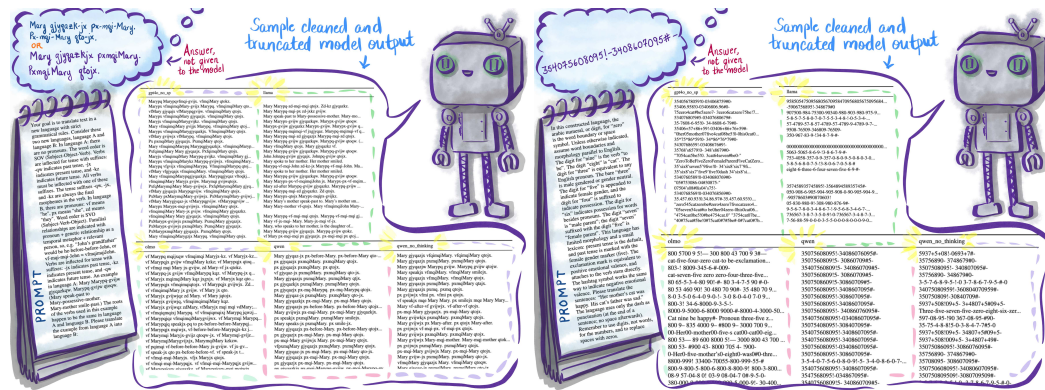


Figure 1: This figure shows prompts describing the translation family of rules (translate-1 and dan-1), their answer keys, and sample outputs from each model, to illustrate what we mean by separating meaning-mediated plausibility from statistical plausibility (although we do that in different ways across the rule families). It is apparent from these responses that these tasks are hard, that some models do much better than others, and that dan-1 is more difficult than translate-1. However, it is also apparent that models can produce outputs exactly matching or near to the answer keys, behaviour which is better explained by meaning plausibility than statistical plausibility.

of the prompt (beyond the most explicitly statistical components) – or a result of successful pattern recognition and regurgitation. For convenience, we’ll call these meaning-mediated plausibility or meaning plausible and statistically-mediated plausibility or statistically plausible, respectively.¹¹

The fact that superficial statistical likelihood often¹² aligns with the meaning of an utterance is what makes plausible text generation so powerful. Under the right conditions, statistical learning may access traces of meaning, not just the most superficial patterns visible in the training data. Meaning-mediated abstractions therefore need not be seen as alternatives to statistical learning. They may emerge because they are useful compressions for solving statistical prediction problems, both *in silico* and *in vivo* [Greco et al., 2024, Stöckl et al., 2024].

The challenge is that plausible outputs do not reveal which process produced them. Both superficial statistical regularities and meaning-mediated abstractions can generate plausible text, obscuring inference of the underlying mechanism from the response alone. This creates a construct-validity problem for many benchmark-style evaluations (see Section 4.2). We design our experiment to address this challenge, by forcing the relatively concrete plausibilities of token statistics to contradict the relatively abstract plausibilities of meaning.

1.3 Prior work and superficial patterns

Our argument is not that LLMs don’t rely on patterns. Quite the opposite: the existence of learned patterns is what makes these systems interesting. If we allow patterns at any level of abstraction, essentially everything discussed in this paper can be described as pattern matching.¹³ The question is therefore not whether models learn and use patterns, but what kinds of patterns they learn.¹⁴ The strong stochastic parrot hypothesis (SSP) says that LLM behaviour can be explained by relatively concrete regularities: this symbol follows this symbol, these symbols tend to occur near those symbols, and so on.¹⁵ See Fig. 29.

¹¹However, keep in mind these are conceptually along the same concrete-abstract, fungibility-groundedness spectrum of meaning, with the superficial statistical patterns (symbol co-occurrence, some morphology and syntax) on the more concrete, immediately apparent side and semantic, pragmatic, register, and discourse-level content higher on the abstract side. There is no bright line between them! We want to probe the abstract end and discredit the most concrete end (including some low-level abstractions) as sufficient explanation for behaviour.

¹²And perhaps specifically, often but not always. We will return to this topic in Section 4.3.

¹³Even as statistical pattern matching.

¹⁴And when they use them.

¹⁵By contrast, people appear capable of learning across a wide spectrum, from concrete regularities to more abstract principles that can be flexibly expressed through different symbolic systems, without a discrete boundary where pattern matching becomes abstraction.

Several prior works suggest that LLMs are not restricted solely to superficial patterns. [Digutsch and Kosinski \[2023\]](#) find that semantic similarity predicts GPT-3 activations better than word co-occurrence. [Hewitt et al. \[2025\]](#) show that models can learn and self-verbalize meanings associated with newly introduced embeddings, while [Hewitt et al. \[2024\]](#) find evidence of instruction-following behaviour prior to explicit instruction tuning. [Cui et al. \[2024\]](#) identify the possibility of a transition from positional to semantic pattern matching.¹⁶

Taken together, these observations establish antecedent plausibility for meaning-mediated structure in LLMs, but they do not cleanly dissociate statistical and meaning-based explanations. Our contribution is to probe that distinction explicitly through conlang-like tasks designed to decouple meaning-mediated plausibility from familiar surface distributions.

The crux of our argument therefore depends on constructing tasks whose correct responses are statistically implausible based on natural language corpora, yet which are learnable from natural-language instruction. We use “meaning-mediated plausibility” to refer to outputs better explained by the content of an instruction than by its component surface distributions.¹⁷

2 Methodology

Our research question is, **can a model generate an answer that is licensed by a natural-language instruction (plausible via meaning) but disfavored by superficial¹⁸ distributions (statistically implausible)?**

Our argument relies on three steps. First, we construct conlang-like tasks in which statistical plausibility and meaning-mediated plausibility point in different directions. See Figure 1 for examples. Second, we formulate explicit hypotheses about what successful rule-following should look like and measure movement in those directions. Third, we ask whether the resulting behaviour is better explained by the content of the instructions than by familiar surface distributions alone.¹⁹

We are not trying to show that LLMs are general reasoners or that they learn, create, or use abstractions in a human-like way. Our argument posits the existence of some non-trivial meaning-mediated abstractions, accessible under some conditions, but we do not identify their specific nature.

2.1 Model selection

Because our rules vary in difficulty and operate near the frontier of model failure, we balance architectural simplicity against performance. Our strongest argument would be made with a purely autoregressive LLM, so the more “pure” a model is, the more imperfect task performance we are willing to tolerate. Based on preliminary testing, we selected models that span next-token-prediction purity, instruction tuning, reasoning-oriented training, openness and training-data transparency, interpretability, and sufficient performance to produce detectable output-level evidence. See Appendix A.3 and Table 16 for details.

In short, GPT-4o and Qwen were selected for expected performance, OLMo for openness, and Llama as a base model. We treat Qwen with and without thinking enabled separately because we expect CoT-style (chain-of-thought style) fine-tuning to modulate the influence of lower-level statistical plausibility (Section 4.3). As a base model, Llama’s weaker directional shifts remain theoretically interesting (à la [Hewitt et al. \[2024\]](#)), especially when considered as part of this spectrum of models. These models were chosen to support interpretation of the present experiment rather than to enable a controlled architecture or scaling comparison, so although our results motivate exploratory hypotheses, they are not probative on those questions.

¹⁶Further discussion of the pattern-to-thought transition in machines in [Zimmerman \[2025\]](#) and [Zimmerman et al. \[2024b\]](#), although our views continue to evolve. The current work is also inspired by [Diamond \[2023\]](#), [Huang et al. \[2023\]](#), [Kallini et al. \[2024\]](#), [Mahowald et al. \[2024\]](#), [Zimmerman et al. \[2024a\]](#).

¹⁷And “statistical plausibility” to refer to outputs that can be explained by their component surface distributions, whether they can be explained by other kinds of plausibility or not.

¹⁸At or near surface-level.

¹⁹We do not view “instruction following” and “abstraction” as competing explanations. Successfully following a sufficiently novel instruction requires some representation of the instruction’s content. Our argument is not that the representation must be human-like, but that explanations based solely on superficial symbol co-occurrence are inadequate. See Figure 29.

2.2 Experimental design

We use five models: GPT-4o, Qwen (thinking, with CoT-style-prompting enabled), Qwen (no thinking, with CoT-style-prompting disabled), OLMo, and Llama.²⁰ More detailed descriptions of each model are provided in Table 16.

For each model and rule, we ran 50 independent outputs. Each rule – except sidequest-1²¹ – has 3 prompt variations with slight differences in style, formatting, and wording, but interchangeable propositional content, assigned randomly across runs. Since small variations in wording can cause notable changes in model behaviour [Sclar et al., 2024], we included these variants to help control for sensitivity to surface phrasing without venturing into prompt engineering or optimization. Across all models, we used fixed parameters (temperature of 0.7, a token budget of 4,096, and approximately 20 sentences of requested output where possible). Since our goal is an existence argument rather than performance optimization, we did not perform parameter sweeps. Outputs were generated independently, without in-context accumulation (but see Appendix A.2). For additional details, see Appendix A.3.

We frame our argument around existence because that is the logical structure required to refute the SSP while sidestepping common interpretive pitfalls: existence lets us ask a specific and tractable question without placing undue weight on complex unknowns such as consistency across contexts (see Appendix A.3).²²

2.3 Identifying and measuring rule-following behaviour

Outputs were cleaned through a combination of manual effort and automation. To enhance intercleaner reliability, we discussed cleaning criteria as a team, reviewed examples together for consistency, and spot-checked 3 randomly chosen computationally-cleaned runs per rule per model to refine the process. Where possible, we applied metrics to raw outputs as well as cleaned outputs, to sanity-check that cleaning choices were not driving results.

During the preliminary testing phase, we developed automatable tests for specific aspects of rule-following – essentially unit tests for the output – building measurability into the rules themselves. Each test captures only some aspects of compliance, making them a lower bound on rule-following; underestimating is preferable to overestimating, since overestimating could lead to a false positive. Tests range from open-ended (story rules, double verb rules) to exact (translation rules, which can be checked against an answer key), and applied to both cleaned and raw outputs where possible. We conducted both model-level and pooled analyses, with permutation testing and corrections where warranted. See Appendix A.3 for details.

Some behavior may be visibly rule-directed without being fully captured by automated checks, but for the most part, that analysis will be done in part 2. See Appendix A.11 for some preliminary observations.

2.4 Rule design

The rules must satisfy two requirements. First, correct responses should be difficult to explain using superficial statistical regularities. Second, the tasks should remain close to linguistic competence rather than depend primarily on unrelated capacities such as mathematics, counting, very long-distance memory, or sensory experience.

We therefore aim for a theoretical “sweet spot”: tasks reliant on structures that models could arguably have learned given their architecture and training data, yet whose correct outputs diverge from familiar natural-language distributions. Neural network architectures can represent data structures underlying approximately context-free and slightly context-sensitive languages, formal classes that align in

²⁰There are some minor variations in how we refer to these models, but they all refer to the same five: GPT-4o, Qwen (with thinking enabled), Qwen (without thinking enabled), OLMo, and Llama. GPT-4o is sometimes described as “no system prompt”, because it has a slightly different expected prompt format than the others. We treated all models as identically as possible, so we did not provide any additional system prompt to GPT-4o, even though that is a common application.

²¹Due to a copy-paste error, the sidequest-1 runs analyzed in part 1 use 2/3 identical prompts and 1/3 a second variant. Corrected data (sidequest-1-fix) will be analyzed in part 2.

²²However, the degree of alignment between our hypotheses and results is stronger than that framing alone suggests. Obviously, that is a topic for Results 3 and Discussion 4, but we felt earlier signposting might be helpful.

theory with natural language [Wilcox et al., 2019]. Since LLMs aren’t people, we don’t exactly know what they could have learned while acquiring the level of linguistic competence they exhibit, but we can consider people as a kind of outer bound: tasks that no human acquires through language exposure alone are unlikely to have been learned by PTGA models. Our conlangs therefore remain within the approximate domain of language learning while being unusual enough that explanations based solely on superficial pattern matching are not merely less parsimonious, but increasingly hard to sustain.²³

To undermine statistical plausibility at concrete levels, we introduce linguistic features that are rare to unattested in natural language, including unusual word orders, rare English bigrams, word-order parity rules, inserted verbs, and a translation system in which prompts and correct answers share no character-level overlap. These manipulations make correct outputs atypical under familiar surface distributions while preserving the overall structure of a language task. At the same time, we avoid completely disrupting information locality, requiring mathematical reasoning, or relying on embodiment [Zimmerman, 2025, Lovato et al., 2024, Pfeifer and Bongard, 2006, Hahn et al., 2020, Mayzner and Tresselt, 1965, Norvig, 2013, Liu et al., 2024b, Kallini et al., 2024, Futrell et al., 2020b, Futrell and Hahn, 2024, Gibson et al., 2019, Harris, 1954, 1968, Sahlgren, 2008, Zipf, 1965, Futrell et al., 2020a]. For example, we do not test counting, because linguistic competence does not require counting ability.

In addition to designing rules that are theoretically unlike the training data, we perform contamination checks for potential training-data exposure (Section 2.5; Appendix A.4). Because this is an existence argument, success should be understood as systematic movement in the predicted direction rather than perfect rule execution. These are intentionally difficult tasks, and imperfect performance does not imply that the rule was not represented at all. We will discuss the exploratory testing in which we developed the conlangs more in part 2.

2.5 Training data exposure

We performed complementary checks for training data exposure. While these checks cannot prove the absence of exposure of these models to these rules, they can dramatically reduce how much likelihood we allocate to that possibility.

The results of these checks (as well as the number of models and runs per rule, the design and diversity of the rules, and the inclusion of prompt variations) make it implausible that any model we tested could have succeeded by direct recall of training data. We searched for 19 representative strings contained in our prompt and answer pairs in 8 major training corpora including OLMo, DCLM-baseline, and Dolma-v1.7; only one string, matching “zero is the word boundary OR zero is the space symbol”, appeared in some of the training corpora (ranging from 1-4 instances). These appear to refer to wholly dissimilar contexts (see Table 13). We used the InfiniGram N-gram LM to test whether the same training corpora assigned the highest probability to a token that was part of the correct continuation for dan-1 and translate-1 prompt variants; of 12 variations, only one condensed form of the dan-1 prompt yielded a correct prediction by any training corpus, and none when we took (fuzzy) order into account [Liu et al., 2024a]. See Table 7, Table 8, and Table 13. For additional details on these checks and the additional checks we performed, see Appendix A.4.

2.6 Conlang families and hypotheses

In part 1, we focus on three rule families that provide complementary evidence for our core argument. In part 2, we extend the discussion to the full set of rules, including additional evidence regarding rule-following behaviour, potential cognitive tradeoffs and a more detailed exploration of output modes.²⁴

²³A caveat: As children, speakers learn tasks that look like aspects of the conlangs we use. But they do so with varying degrees of conscious accessibility. Our tasks foreground linguistic skills as conscious reasoning, which is a non-trivial difference. Furthermore, first language acquisition and second language acquisition are different for people, and are not uniform across language-related skills [Malaia et al., 2020].

²⁴As the paper is already somewhat lengthy, we split the minimal core argument from the broader experimental results and exploratory analysis.

- The story family (comparison-1, sidequest-1, comparison-2)²⁵ probes whether unusual output can be explained by prompt exposure alone, using sustained generation of unusual lexical distributions over approximately 20 sentences.
- The double verb family (doublev-1, doublev-2) probes whether models can override syntactic and semantic plausibility priors through the insertion of additional irrelevant verbs.²⁶
- The translation family (translate-1, dan-1)²⁷ probes composed symbolic mappings involving lexical, morphological, semantic, and word-order constraints. Dan-1 mitigates the role of character-level overlap between prompt and answer.²⁸

Summaries of the key hypotheses for each family of rules are:

- Story family: comparison-1 < sidequest-1 < comparison-2 in rates of rare bigram usage²⁹, and comparison-2 should be distinguishable from both comparison-1 and sidequest-1.³⁰
- Double verb family: production of double verbs; high semantic similarity across paired verbs in doublev-1, obviously lower similarity in doublev-2.
- Translation family: outputs should be substantially closer to the answer keys than matched random strings, with weaker performance as the representational regime becomes less familiar (moving from translate-1 to dan-1).

Together, these families span lexical, semantic, morphological, syntactic, word-order, metaphorical, discourse-level, and arguably pragmatic rule application. While any individual rule could admit rule-specific alternative explanations, the convergence of evidence across multiple qualitatively different dissociations between meaning-mediated and statistical plausibility strengthens the overall argument. Taken together, the results allow us to probe not only whether models move beyond verbatim regurgitation, but whether the patterns they rely on extend substantially beyond superficial distributional patterns. For readability, all rules are available in Appendix A.5.

3 Results

Our hypotheses were specific and directional, and the observed outcomes generally match those predictions. All rules elicit rule-following behaviour from some models. All models exhibit some evidence of rule-following on all rules except the most challenging (dan-1), and the predicted ordering holds, with effect sizes frequently near ceilings, and the same qualitative pattern across cleaned and uncleaned outputs. The weakest evidence, where results are not statistically significant, comes from Llama, which is the model we expected *a priori* to show the weakest signal. Table 1 summarizes our results. Each family of rules is described in more detail below (Sections 3.1, 3.2, 3.3), and discussed in Section 4.

We observe that model responses can be multimodal. Anecdotally, we observe some failures that metaphorically evoke attractors in dynamical systems: initial trajectories of varying durations, eventually collapsing into degenerate cycles. Some initial observations as to such output modes are given in Appendix A.11, though we mostly defer to part 2.

3.1 Story rules

This family controls for the possibility that unusual output is caused by exposure to unusual prompt content rather than by rule-following. We distinguish three regimes:

- Baseline: In comparison-1, we request ordinary English story generation of about 20 sentences.

²⁵And sidequest-1-fix in part 2.

²⁶Doublev-2 arguably also conflicts with dependency locality, since the irrelevant verb is not near anything it modifies. The rules also require some overriding of lexical and morphological priors, to mark the irrelevant verb with a specific rare bigram.

²⁷And translate-2 in part 2.

²⁸Additionally, translate-1 requires a metaphor map, and dan-1 uses chiasm (“Her mother’s cat... His cat’s father...”) to further subvert statistical plausibility.

²⁹In terms of tokens, not types, in the type-token distinction. Rare bigram ratio was computed as the proportion of all output bigrams that belonged to the predefined rare bigram inventory.

³⁰Additionally, idealized rule-following would see comparison-1 and sidequest-1 be indistinguishable, but PTGA constraints should prevent models from achieving that. In practice, we expect the best rule-following to show up as comparison-1 and sidequest-1 looking much more like each other than either does to comparison-2.

Table 1: Tested conlangs grouped by family, with hypotheses, tests, and summarized findings.

Rule(s)	Description	Hypothesis	Test	Y/N
<i>Family: Story rules</i>				
sidequest-1, comparison-1, comparison-2	comparison-2 should show a preponderance of rare bigrams versus the other two rules, ruling out prompt-driven echoing as a sufficient explanation.	Rare bigram usage should be higher for comparison-2 than sidequest-1 and comparison-1 (<i>use</i> versus <i>mention</i> of rare bigrams should yield distinguishable output); ordering should be comparison-1 < sidequest-1 < comparison-2; idealized rule-following should see comparison-1 and sidequest-1 rare bigram usage be indistinguishable, but PTGA should prevent this in practice.	Rate of rare bigram usage comparison across conditions; see Figure 2. Additionally: Fig. 19, Tables 2, 3, 4, 5, 17, 18, 19, 20, 21, 22.	Y (for all cleaned outputs: Qwen, GPT-4o, OLMo; Y for some models for uncleaned outputs)
<i>Family: Double verb rules</i>				
doublev-1, doublev-2	Models must use a novel double-verb structure; doublev-2 additionally emphasizes that the second verb should be semantically unrelated to the first.	(1) Production of double verbs; (2) Ratio of verbs marked with <i>-qv</i> to total verbs ≈ 0.5 ; (3) verb pairs in doublev-1 have higher cosine similarity than in doublev-2.	(1 and 2) Count; (3) KS test and Wilcoxon Rank Test for doublev-1 > doublev-2 cosine similarity. See Tables 27, 28, Additionally: Figures 4, 5, Tables 23, 24, 25, 26, Figures 3, 20, 21.	(1) Y for all models; (2) Y for some models; (3) Y for all models
<i>Family: Translation rules</i>				
translate-1	Complex novel translation task.	(1) String distance measures – output Levenshtein distances, normalized Levenshtein, Jaccard, and cosine similarity – should be closer to the answer key than randomized strings (select-character and all-character). (2) Exact matches to answer key.	KS test and Wilcoxon Rank-Sum on cosine string similarity vs. random strings; string distance comparisons; regex. See: Figures 6, 25, Tables 31, 32.	(1) Y for all models ; (2) Y for some models
dan-1	Complex translation task. Uses no characters from the prompt, controlling for token exposure.	(1) Output closer to correct answer than random strings; (2) Exact matches.	KS test and Wilcoxon Rank-Sum on cosine similarity vs. random strings; string distance comparisons; regex. See: Figures 7, 24, Tables 33, 34.	(1) Y for some models; (2) Y for some models
translate-1, dan-1	Supplementary check to support the above results via established machine translation metrics (BLEU, chrF, TER, METEOR).	Similar cosine similarity findings across both rules; translation metric scores approach optimal (BLEU = 100, chrF = 100, TER = 0, METEOR = 99.6).	Cross-rule comparison of cosine similarity; translation benchmarks against reference. See Figures 22, 23.	Y for some models

- Prompt exposure: In sidequest-1, rare bigrams [Norvig, 2013, Mayzner and Tresselt, 1965] are presented in the prompt, but the model is instructed to write an ordinary English story.
- Rule-following: In comparison-2, the same rare bigrams are described as the morphemes of a new language and the model is instructed to produce a story of about 20 sentences in that language. Both sidequest-1 and comparison-2 contain the same list of rare bigrams in the prompt.

An alternative explanation to rule-following is that symbols introduced in a prompt tend to recur in subsequent output, independent of their intended role.³¹ If prompt exposure alone were sufficient to explain conlang performance, sidequest-1 and comparison-2 should produce similarly elevated rates of rare-bigram use.

Because of competing plausibilities, we expect the ordering of rare bigram use to be (1) comparison-1 < sidequest-1 < comparison-2. Rule-following behaviour should at least satisfy: (2) rare bigram usage in comparison-1 and comparison-2 are distinguishable, and (3) rare bigram usage in sidequest-1 and comparison-2 are distinguishable. An idealized account of rule-following – in which only

³¹For example, introducing a topic such as baseball tends to increase the frequency of baseball-related terminology later in the conversation. If a person is introduced with a male pronoun, male pronouns are likely to recur. Such regularities may help bridge syntax, semantics, and pragmatics.

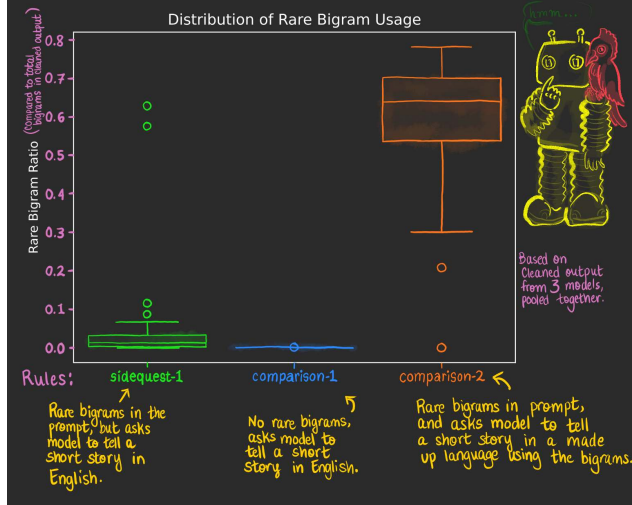


Figure 2: Comparison of rare bigram usage across conditions, for the 3 models (Qwen (with thinking), GPT-4o, OLMo) with cleaned output.

Table 2: Fisher-combined p -values across models (Qwen (T), Olmo, GPT-4o (no SP)) for each contrast; cleaned outputs.

Comparison	Fisher statistic (X^2)	$p_{combined}$	k
comparison-2 vs sidequest-1	226.794	3.700e-46	3
sidequest-1 vs comparison-1	187.022	1.093e-37	3
comparison-2 vs comparison-1	243.473	1.017e-49	3

instructed use matters and prompt exposure has no effect (which we don’t actually expect to see, since we think in practice the plausibilities compete) – would additionally satisfy (4) sidequest-1 and comparison-1 are indistinguishable, since merely encountering the rare bigrams would have no effect. The cleaned results should be more dramatic than the uncleaned results (since any additional English text generated reduces the ratio of rare bigrams).

All three models (GPT-4o, Qwen (thinking), OLMo) with cleaned outputs satisfy the predicted ordering (1) with comparison-2 strongly distinguishable from both controls (2 and 3). The same holds with the uncleaned outputs for four of the five models; Llama’s rule-following (comparison-2) outputs differ from baseline (comparison-1) in the expected direction but are not reliably distinguishable from its prompt exposure (sidequest-1) outputs. As expected, no model exhibits idealized rule-following (4): prompt exposure has a detectable effect in every model. However, the size of that effect varies substantially. GPT-4o comes closest to the ideal pattern, with sidequest-1 remaining much closer to baseline than to rule-following output.

That inclusion in the prompt is not generally sufficient to cause inclusion in the output weakens prompt exposure as a general alternative explanation for rule-directed behaviour. Although the prompts are otherwise similar, not identical, our results suggest the interpretation that the same surface forms – the lists of rare bigrams – can produce different behaviour in LLMs depending on their communicative role within the prompt. Because the data contain many tied and near-zero values, we use permutation tests to estimate combined p -values while minimizing distributional assumptions. Combined results are reported in Tables 2, 3. Additional results are reported in Appendix A.7.

3.2 Double verb rules

The double verb family instructs models to add a second, irrelevant verb to each main verb and mark it with the affix -qw. In doublev-1, the added verb is just described as irrelevant. In doublev-2, we additionally specify that it should be semantically unrelated to the first verb and feel random relative to the rest of the sentence. Otherwise, the rules are the same. Because verbs occurring near

Table 3: Fisher-combined p-values across all models for each contrast (uncleaned output/ all output).

Comparison	Fisher statistic (X^2)	$p_{combined}$	k
comparison-2 vs sidequest-1	303.771	2.479e-59	5
sidequest-1 vs comparison-1	290.608	1.501e-56	5
comparison-2 vs comparison-1	349.397	5.350e-69	5

Table 4: **Mann-Whitney U tests (i.e. Wilcoxon Rank-Sum) for cleaned model outputs for sidequest-1, comparison-1, and comparison-2 (cleaned outputs available for GPT-4o, Qwen (thinking), and OLMO).** Note that GPT-4o shows the closest to ideal performance here (sidequest-1 and comparison-1 outputs clearly more similar to each other than comparison-2 output is to either).

Model	Comparison	U	p	p (BH)	Cliff’s Delta	Median A	Median B	Median Diff
gpt4o (no SP)	sidequest-1 vs comparison-1	1980	3.05e-08	3.05e-08	0.584	0.001982	0.000000	0.001982
gpt4o (no SP)	comparison-2 vs comparison-1	2500	5.42e-19	3.94e-18	1.000	0.565855	0.000000	0.565855
gpt4o (no SP)	comparison-2 vs sidequest-1	2500	6.06e-18	1.36e-17	1.000	0.565855	0.001982	0.563873
olmo	sidequest-1 vs comparison-1	2406	6.11e-16	6.88e-16	0.925	0.005576	0.000000	0.005576
olmo	comparison-2 vs comparison-1	2465	1.90e-17	3.41e-17	0.972	0.706837	0.000000	0.706837
olmo	comparison-2 vs sidequest-1	2450	1.31e-16	1.68e-16	0.960	0.706837	0.005576	0.701260
qwen (T)	sidequest-1 vs comparison-1	2500	1.31e-18	3.94e-18	1.000	0.020997	0.000000	0.020997
qwen (T)	comparison-2 vs comparison-1	2500	1.31e-18	3.94e-18	1.000	0.629056	0.000000	0.629056
qwen (T)	comparison-2 vs sidequest-1	2461	7.12e-17	1.07e-16	0.969	0.629056	0.020997	0.608059

one another in English are typically semantically related, and at least related to the words around them, the second rule pushes against strong plausibility priors (both syntactic and semantic) learned from natural-language data, and violates the principle of least effort [Zipf, 1965, Gibson et al., 2019, Futrell et al., 2020b].³² Comparing cosine similarity between verb pairs from doublev-1 to doublev-2 therefore provides a direct test of whether models alter generation in the requested semantic direction.

Rule-following should result in the production of additional verbs, about half of which should be marked with the instructed rare bigram morpheme.³³ The relationships between the pairs of verbs should further satisfy specific hypotheses. First, the cosine similarities between paired verbs should be substantially lower in doublev-2 than in doublev-1. Second, our PTGA account predicts that doublev-1 should often exhibit very high cosine similarities, because local plausibility priors strongly

³²Doublev-2 may also conflict with dependency locality, since especially when the second verb is irrelevant in the sense specified in doublev-2, it is arguably not near anything it modifies.

³³Successful rule-following output should yield a total count of verbs per run that is twice the number of verbs marked with the -qw morpheme. Since doublev-2 emphasizes the irrelevance of the added verb, we expect that cosine similarity between verbs should be a lower value (e.g. closer to 0) for the verbs produced by this prompt than for doublev-1. Since we think doublev-2 is harder, we expect the ratio of -qw verbs to go down, and the total number of verbs produced to go down, compared to doublev-1. In part 1 we have not tracked whether -qw marks the *better* choice of “irrelevant” verb, where present. We may explore that in part 2.

Table 5: Mann-Whitney U comparisons of rare bigram usage medians for all sidequest-1, comparison-1, comparison-2 output (uncleaned outputs, available for all 5 models).

model	comparison (A-B)	U	p	p_{bh}	Cliff’s δ	Median A	Median B	$\bar{x}_A - \bar{x}_B$
gpt4o (no SP)	sidequest-1 vs comparison-1	1980	3.05e-08	3.81e-08	0.584	0.001982	0.000000	0.001982
gpt4o (no SP)	comparison-2 vs comparison-1	2500	5.42e-19	7.48e-18	1.000	0.565855	0.000000	0.565855
gpt4o (no SP)	comparison-2 vs sidequest-1	2500	6.06e-18	1.14e-17	1.000	0.565855	0.001982	0.563873
llama	sidequest-1 vs comparison-1	1649	1.26e-03	1.35e-03	0.319	0.000000	0.000000	0.000000
llama	comparison-2 vs comparison-1	1812	1.21e-05	1.40e-05	0.450	0.000417	0.000000	0.000417
llama	comparison-2 vs sidequest-1	1414	2.34e-01	2.34e-01	0.131	0.000417	0.000000	0.000417
olmo	sidequest-1 vs comparison-1	2500	2.49e-18	7.48e-18	1.000	0.030387	0.000000	0.030387
olmo	comparison-2 vs comparison-1	2500	2.49e-18	7.48e-18	1.000	0.328400	0.000000	0.328400
olmo	comparison-2 vs sidequest-1	2495	9.54e-18	1.59e-17	0.996	0.328400	0.030387	0.298013
qwen (T)	sidequest-1 vs comparison-1	2500	4.72e-18	1.01e-17	1.000	0.030265	0.000254	0.030011
qwen (T)	comparison-2 vs comparison-1	2500	4.72e-18	1.01e-17	1.000	0.075579	0.000254	0.075325
qwen (T)	comparison-2 vs sidequest-1	2424	5.97e-16	8.14e-16	0.939	0.075579	0.030265	0.045314
qwen (no T)	sidequest-1 vs comparison-1	2500	1.74e-18	7.48e-18	1.000	0.041571	0.000000	0.041571
qwen (no T)	comparison-2 vs comparison-1	2500	1.74e-18	7.48e-18	1.000	0.395845	0.000000	0.395845
qwen (no T)	comparison-2 vs sidequest-1	2450	1.35e-16	2.02e-16	0.960	0.395845	0.041571	0.354274

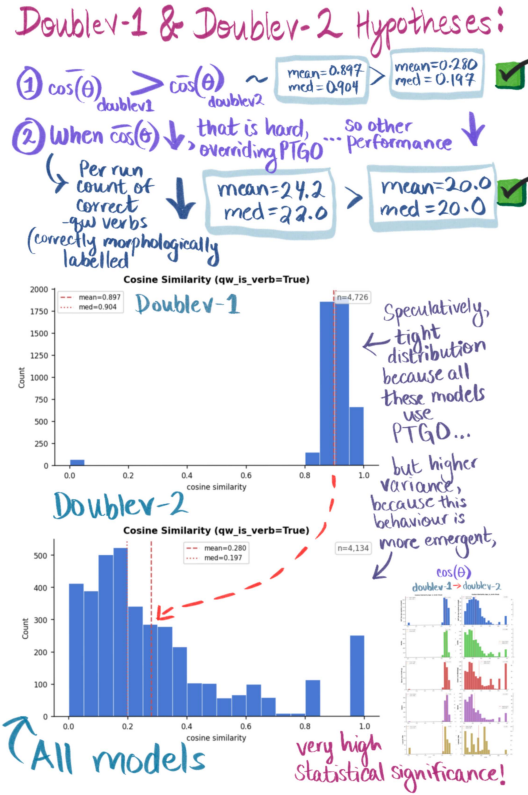


Figure 3: Illustrative summary of double verb family results from our poster presentation [Institut des sciences cognitives, Université du Québec à Montréal, 2026], Fig 30.

favor semantically coherent contexts.³⁴ In addition, because we believe doublev-2 to be harder, we expect we may see reduced performance on some other aspects of the task, suggesting a potential tradeoff amongst limited resources, although we mostly defer that analysis to part 2.

For each model, we observe the production of some double verb pairs, and that cosine similarity scores are lower for doublev-2 than doublev-1, shown in Figures 4 and 5; this same pattern holds when we combine all model output, as shown in Figures 20 and 21. Models vary in their ability to perform the task of marking verbs with -qw, but largely placed marked -qw verbs near other verbs when successfully marking a verb. When we clarified the prompt to enforce semantic randomness, semantic performance improved, even if other aspects may have degraded – we do see a reduction in per-run median and mean counts of correctly-labeled verbs from doublev-1 to doublev-2, but will explore potential cognitive tradeoffs in part 2.

Additional results for the double verb family of rules are given in Appendix A.8: Table 28, Table 27, Table 24, Table 26, Table 23, Table 25.

Overall, models alter generation in the requested syntactic and semantic directions while exhibiting model-dependent difficulty in doing so (Figures 4 and 5). This pattern is consistent with our hypothesis that, in PTGA, meaning-mediated behaviour can coexist with persistent autoregressive plausibility constraints, which aligns with observations reported in McCoy et al. [2023].

³⁴From the perspective of PTGA, generating closely related, even identical, verb pairs is one of the easiest ways to satisfy the rule.

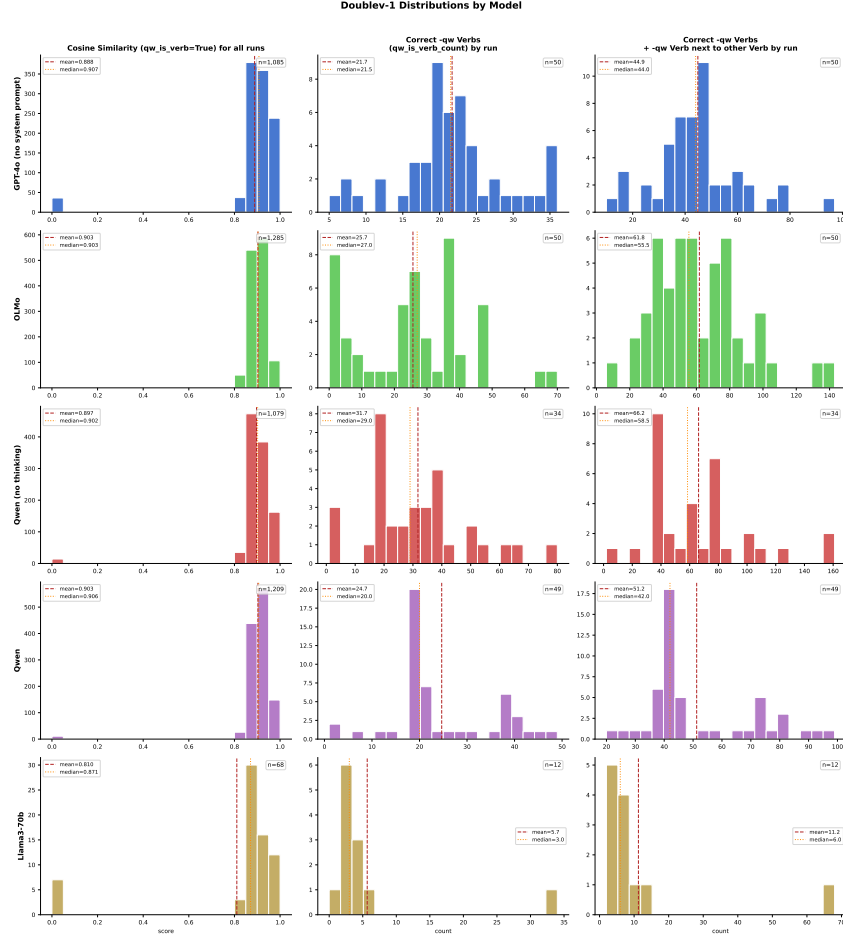


Figure 4: **Doublev-1 Rule Similarity Scores and Verb Counts by Model.** Histogram of similarity scores (cosine similarity between marked and closest verb, where marked verb is appropriately terminated with bigram *-qw*) per model, as well as the per-run count of correctly marked *-qw* verbs and count of correct pairs of *-qw* verbs and closest verb. Total number of runs with verbs reported for each model ($n = 50$ runs indicates each run had at least one instance of a correctly assigned *-qw* verb). Cosine similarity scores closer to 0 indicate dissimilar words. Counts by model that are closer to total clauses reported by model indicate higher performance by model.

3.3 Translation rules

Unlike the story and double verb families, translation outputs can be evaluated against explicit answer keys. Successful performance therefore requires not merely producing text with unusual features, but preserving specific relationships under tightly constrained transformations. Correct outputs preserve abstract relationships under a series of composed constraints while (sometimes dramatically) changing the symbols that realize those relationships. Preserving tense, possession, kinship, grammatical roles, and ordering while radically changing the symbols that express them is much more naturally explained by representations of the underlying relations than by superficial token distributions.

We investigate two translation conlangs: translate-1, which uses alphabetic characters, and dan-1, which relies primarily on numbers and punctuation. These are the most complex rules in our suite, both in the number of constraints and in the number of relationships among those constraints. Our research question is: can a model preserve relations across disjoint forms, under multiple composed constraints operating at different linguistic levels, even when tokens must be used contrary to their familiar patterns of use?



Figure 5: **Doublev-2 Rule Similarity Scores and Verb Counts by Model.** Histogram of similarity scores (cosine similarity between marked and closest verb, where marked verb is appropriately terminated with bigram *-qw*) per model for all verbs paired in all runs, as well as the per-run count of correctly marked *-qw* verbs and count of correct pairs of *-qw* verbs and closest verb. Total number of runs with verbs reported for each model ($n = 50$ runs indicates each run had at least one instance of a correctly assigned *-qw* verb). Cosine similarity scores closer to 0 indicate dissimilar words. Counts by model that are closer to total clauses reported by model indicate higher performance by model.

- Translate-1 combines rare bigrams (phonotactically and orthographically implausible) with composed lexical, morphological, syntactic, semantic, and metaphorical mappings, including a temporal/kinship metaphor.
- Dan-1 combines similarly composed mappings while eliminating overlap between the symbols used in the prompt and those used in the correct output. For example, the prompt uses “two” while the output requires “2”, and standard word boundaries and punctuation must be remapped.

Together, these rules require simultaneous mappings across lexical items, tense, word order, morphology, semantic and emotional concepts, and relational structure, with approximately 15 constraints in translate-1 and 17 in dan-1.

The translation family exploits many overlapping strategies to undermine statistical plausibility at various linguistic scales. For an example from translate-1, in the first system, “-qw indicates past tense, -jx indicates present tense, and -kz indicates future tense,” but in the second system, “-jx indicates past tense, -kz indicates present tense, and -qw indicates future tense”. Note that merely echoing these morphemes without cyclically translating their tense will not correctly preserve meaning. For

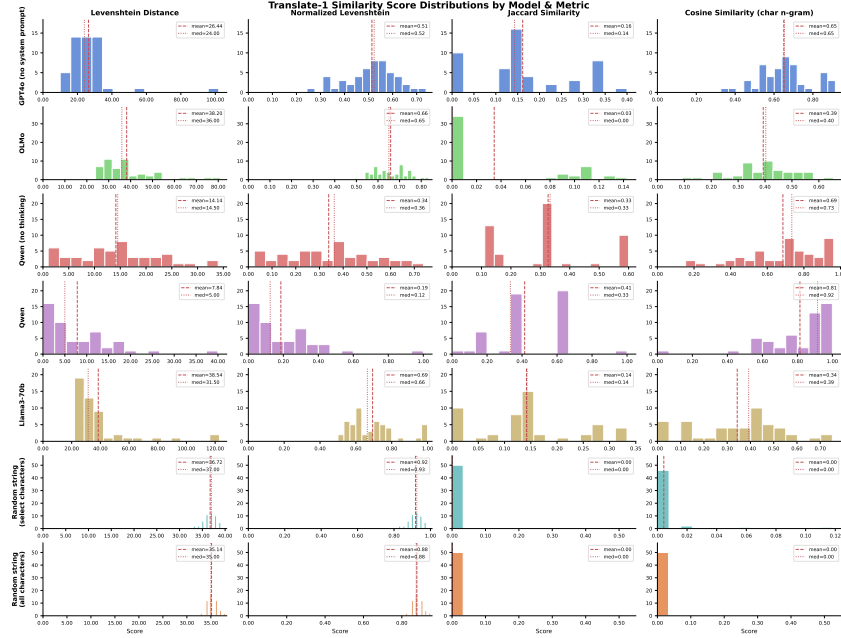


Figure 6: **Translate-1 Distributions by Model and Metric.** Histogram of string distance scores (Levenshtein distance, normalized Levenshtein distance, Jaccard similarity, and character n-gram cosine similarity) comparing the output for Translate-1 rule versus the correct translated response. Scores closer to 0 for Levenshtein distances, and scores closer to 1 for Jaccard and cosine similarities indicate better rule-following/more accurate translation. Qwen (thinking) performs best for Translate-1 rule-following based on median and mean scores for all string similarity measures. For both Wilcoxon Rank-Sum and Kolmogorov-Smirnov tests, all models perform better than the random strings variants achieving statistically significantly better median scores (reported in Tables 31 and 32).

an example from dan-1, chiasm (“her mother’s cat ... his cat’s father ...”) further targets statistical semantic plausibility.

We therefore expect dan-1 likely to be more difficult than translate-1, since it requires symbols, word boundaries, and punctuation marks to be repurposed in unfamiliar ways while simultaneously satisfying an approximately larger set of composed constraints.

All models exhibit some rule-following behaviour on translate-1 relative to random controls, with GPT-4o and Qwen (thinking) generally outperforming the other models. As in several of our analyses, the commercial models tend to outperform the open ones (Figures 6, 7).

All models produce output distinguishable from random controls on translate-1, but performance declines sharply in dan-1, which uniquely separates the models. Qwen (thinking) clearly outperforms the other models, with a median n-gram cosine similarity of 0.6455, compared to 0.0615 for GPT-4o. OLMo and Llama have median cosine similarities of 0 on dan-1, indicating no detectable rule-following behaviour by that metric, although some responses appear qualitatively potentially consistent with attempted rule application (Table 30, Table 29). See Tables 33 and 31. The contrast between translate-1 and dan-1 is consistent with PTGA remaining strongly sensitive to surface forms: Performance degrades sharply as successful completion requires increasingly extensive remapping of the necessary output structures.

Translation performance also varies substantially under standard machine-translation metrics (BLEU, chrF, TER, METEOR, and a composite of the four). No model approaches perfect performance, but the same qualitative pattern remains visible: GPT-4o and both versions of Qwen cluster together at the higher-performing end, while OLMo and Llama form a lower-performing cluster (Fig. 22). Analyses of truncated model outputs are reported in Appendix Fig. 23).

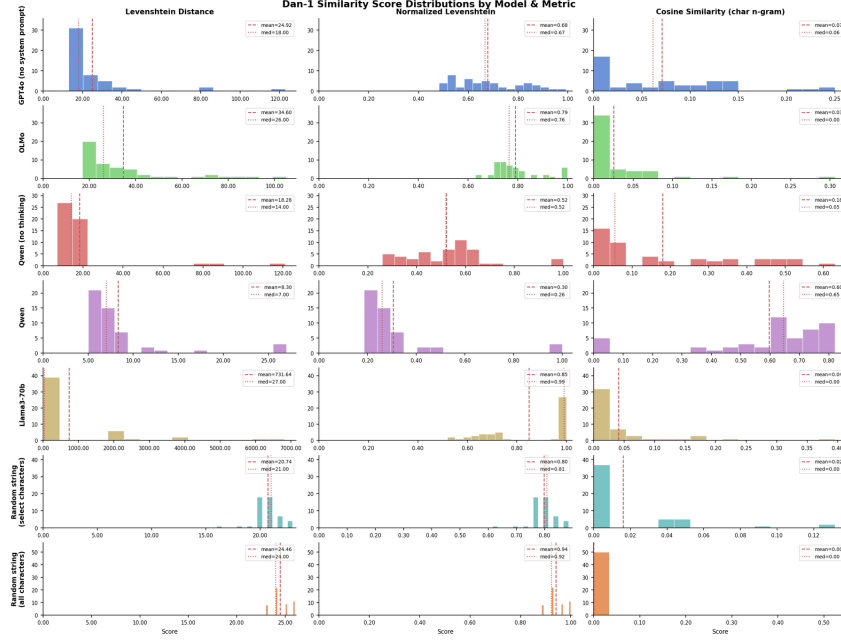


Figure 7: **Dan-1 Distributions by Model and Metric.** Histogram of string distance scores (Levenshtein distance, normalized Levenshtein distance, character n-gram cosine similarity) comparing the output for Dan-1 rule versus the correct translated response. Scores closer to 0 for Levenshtein distances, and scores closer to 1 for cosine similarities indicate better rule-following/more accurate translation. Qwen (thinking) performs best for Dan-1 rule-following based on median and mean scores for all string similarity measures. Median OLMo and Llama3-70B scores against random string (select characters) median cosine similarity scores are not significantly different for the Wilcoxon Rank-Sum Tests (reported in Table 33). Likewise, for the Kolmogorov-Smirnov test, which examines the distributional shapes of the cosine similarity scores, neither OLMo, nor Llama3-70B show statistically different scores than the random string (select characters). The distribution of OLMo cosine similarity scores additionally shows no statistical difference with the random string (all characters) using the Kolmogorov-Smirnov test (reported in Table 34).

Together with the string-distance measures in Figures 25 and 24, these results suggest substantial but incomplete rule-following ability, with performance degrading as the required mappings become increasingly unfamiliar. Additional statistical analyses are reported in Appendix A.9, where Tables 31, 32, 33, 34 show Wilcoxon Rank-Sum and Kolmogorov-Smirnov tests for the string cosine distances across models for the translation rules.

Overall, these results show that models can sometimes generate text according to complex, composed constraints. Qwen (thinking) occasionally produced exact matches to our answer keys. Models can preserve relationships involving tense, possession, kinship, grammatical roles, and word order while dramatically altering the surface forms that express those relationships. Because these tasks have explicit answer keys and are therefore substantially less open-ended than the story rules,³⁵ they demonstrate that rule-following extends beyond the generation of unusual distributions. At the same time, performance declines as the model has less opportunity to leverage familiar structures, with numerical/punctuation representations appearing harder than alphabetic representations with typical English-like spacing and punctuation.

The same systems that perform best on translate-1, namely both Qwen models and GPT-4o, also perform best on dan-1, whereas Llama and OLMo do not exhibit clear rule-following on dan-1. However, performance declines substantially for all models as the representational regime becomes less familiar. GPT-4o and Qwen (no thinking), which perform well on translate-1, perform dramatically worse on dan-1, albeit still distinguishably from random controls (and Llama and OLMo). This

³⁵Although in Part I we searched for exact matches to our answer keys because they are sufficient for an existence proof, and used string-distance measures to identify near matches, the answer keys are not necessarily unique.

pattern suggests that architectural differences,³⁶ affect the ability to maintain composed relationships under increasingly unfamiliar representational regimes (see Gan et al. [2026] for a survey of theories of LLM performance divergence). We explore these possibilities further in Section 3.4.

3.4 Exploratory performance results

Caveat: Because of the small number of models and the difficulty of comparing across different metrics and rule families, these results are exploratory. They are more like back-of-the-envelope calculations: We think the gist is potentially meaningful, but we would not put much stock in the exact values, for the reasons described below.

Performance appears to broadly track rule complexity, although we defer systematic investigation of this relationship to future work. We include exploratory analyses below; see Appendix A.10 for additional details.

We normalize performance scores across rule metrics and compare models using both maximum and minimum-maximum normalization procedures (Fig. 9). Overall, commercial models such as GPT-4o and Qwen generally outperform the open models OLMo and Llama, and (as expected) cleaned outputs yield higher scores than uncleaned outputs. We also observe lower performance on rules that we judged to be more difficult, such as the progression from doublev-1 to doublev-2. We include both highly-normalized (Figure 9) and much-less-normalized (Figure 28) results for a richer picture of model performance.

Across models, the relative ordering of rule difficulty is notably stable. The main model differences are in absolute performance level rather than in which rules models find relatively harder or easier. However, rule difficulty is inseparable from the evaluation procedures used to measure it. Different rules are scored with different metrics, operating on different scales and distributions. Some tasks have relatively constrained answer keys, while others permit many valid outputs. Consequently, any ranking of rule difficulty is partly a ranking of the associated evaluation procedures. Throughout the project we used supplementary checks, such as exact-match counts alongside edit-distance measures for dan-1 and translate-1, to help contextualize these differences (Appendix A.3).

We preliminarily observe that, of architectural features, `enable_thinking` and `fine_tuned_and_aligned` show the most consistent correlations with the composite performance scores ($r = 0.71$ for both, straight and weighted), while `context_window`, `multilingual` and `tool_use` show strong correlations with the family-weighted composite ($r = 0.79$, $r = 0.87$, and $r = 0.89$, respectively).³⁷ Notably, `total_parameters` is *negatively* correlated with straight composite performance ($r = -0.46$), driven in part by GPT-4o’s purported $\sim 1.8T$ parameter count paired with fair rule-following scores, falling short of Qwen (with thinking).

We preliminarily observe that, of the features in our rule complexity representation, `uses_no_typical_english` is the single best predictor of performance, not being present in sidequest-1, comparison-1, doublev-1, and doublev-2, and being present in comparison-2, translate-1, and dan-1. Since comparison-2 has one primary constraint (output rare bigrams), from that perspective it would seem easier than doublev-1 and doublev-2, which have 2-3 primary constraints (add an irrelevant verb, mark with a rare bigram morpheme; make the irrelevant verb semantically unrelated to the rest of the context), and translate-1 and dan-1, which each have approximately 15+. However, most models rank comparison-2 as the 3rd most difficult rule, after dan-1 (hardest) and translate-1. Speculatively, this suggests that PTGA models are hindered when familiar functions cannot be expressed in conjunction with the surface forms with which they were learned.³⁸ We could potentially test this mechanism in future work with something like a lexinvariant model [Huang et al., 2023].

Alternative choices of performance metric, complexity representation, or aggregation procedure would affect the precise values reported here.³⁹ Nevertheless, the broad pattern appears credible: models perform extremely well on some rules (comparison-1) and struggle substantially on others (translate-1 and especially dan-1). We therefore view the apparent relationship between rule complexity and

³⁶Broadly construed, including both model architecture and the structure of training and post-training.

³⁷Since Qwen and GPT-4o are multilingual and allow tool use, and Llama and OLMo don’t.

³⁸When the models cannot leverage familiar functions using the forms they would have learned them alongside.

³⁹We explored multiple versions of the relationships between performance, rule features, and architectural features. The versions reported here seemed the most conceptually appropriate, but such choices inevitably involve judgement.

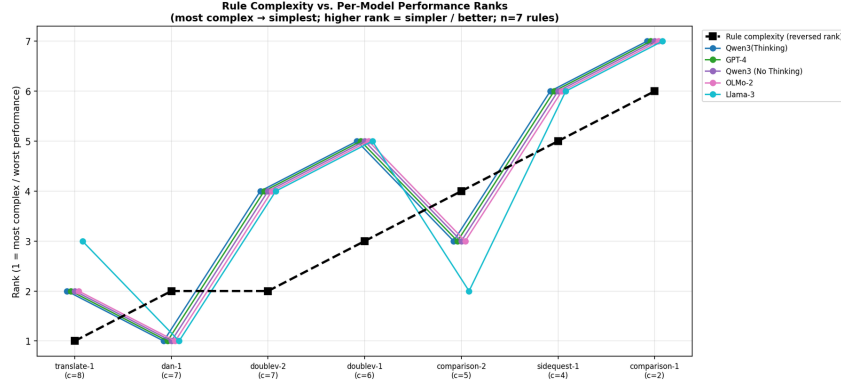


Figure 8: We joined rule-level performance with linguistic rule features and architectural features for exploratory correlation analysis (Spearman, $n=9$ rules / $n=5$ models). To examine the relationship between (this representation of) rule complexity and performance, we collapsed dan-1 and translate-1 to single rows (averaging perfect match and vicinity scores), ranked rules within each model, and computed per-model Spearman r against rule complexity rank.

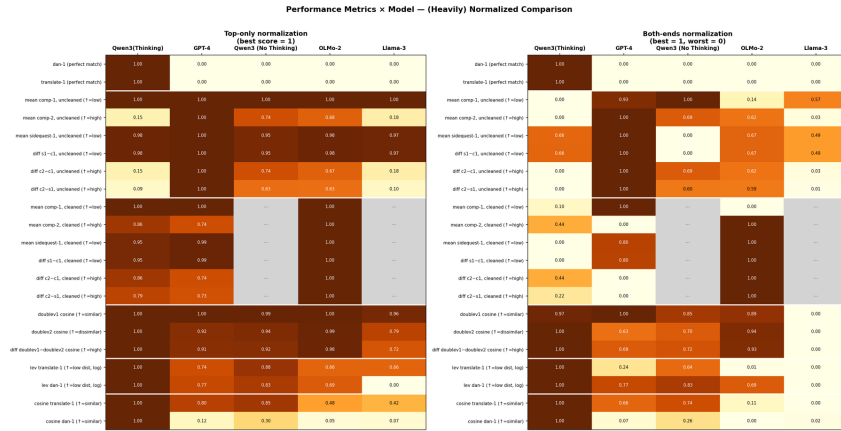


Figure 9: **Heavily normalized performance subset.** Comparison-1 is a baseline, asking the model to generate a short story in English. All models perform well. For non-thinking models, cleaned and uncleaned outputs are nearly identical. For Qwen (thinking), the uncleaned rare bigram rate is suppressed by the CoT trace, which is in standard English. Levenshtein scores are normalized twice here. All performance metrics are reoriented so that higher values indicate better performance, then normalized to $[0, 1]$ in two versions for comparison: *top-only*, where each row is divided by its maximum observed value (best model = 1, others scaled relative to it); and *both-ends*, where each row is min-max normalized (best = 1, worst = 0). Results are visualized as a heatmap with models as columns and metrics as rows, with missing values (models lacking cleaned output data) grayed out. The former gives a clearer picture of overall model performance (for each rule, which models were comparatively able to do it), while the latter gives a clearer picture of relative model performance (for each rule, which model did best and worst). The comparatively un-normalized version of this performance heatmap is included in the Appendix in Figure 28.

performance as a potentially meaningful signal that warrants further investigation (Fig. 8). Future work will explore richer representations of rule complexity beyond the binary feature encodings used here (Fig. 26; Table 36).

Table 6: Raw performance values and within-model ranks (1 = worst performance/ hardest rule, 7 = best performance/ easiest rule) for each rule after collapsing dan-1 and translate-1 to single rows (average of perfect match and vicinity scores).

rule	Qwen3(Thinking) (val)	Qwen3(Thinking) (rank)	GPT-4 (val)	GPT-4 (rank)	Qwen3 (No Thinking) (val)	Qwen3 (No Thinking) (rank)	OLMo-2 (val)	OLMo-2 (rank)	Llama-3 (val)	Llama-3 (rank)
comparison-1	0.9998	7	0.9998	7	0.9998	7	0.9997	7	0.9998	7
comparison-2	0.5904	3	0.5127	3	0.3800	3	0.6900	3	0.0910	2
sidequest-1	0.9467	6	0.9828	6	0.9341	6	0.9919	6	0.9580	6
doublev-1	0.9063	5	0.9073	5	0.9016	5	0.9034	5	0.8705	5
doublev-2	0.8331	4	0.7680	4	0.7800	4	0.8220	4	0.6590	4
dan-1	0.3655	1	0.1441	1	0.1829	1	0.1221	1	0.0100	1
translate-1	0.3799	2	0.2870	2	0.3195	2	0.2085	2	0.1956	3

4 Discussion

Across all three rule families, models moved in the direction predicted by the meaning of the rules. The same pattern appears across multiple models, runs, conlangs, tests, and independent strategies for dissociating meaning-mediated and statistical plausibility. Taken together, these results are increasingly difficult to explain using only superficial statistical regularities and are substantially better explained by meaning-mediated abstraction. Strictly speaking, rejecting the SSP requires only compelling evidence that some outputs demand explanations beyond superficial pattern matching. However, the consistency of the observed patterns exceeds that minimum. Performance varies across models and rule families – harder rule families elicited lower compliance, and open-weight models underperformed compared to their commercial counterparts – but the predicted directional shifts recur across all three families, suggesting that the results are not a fluke, nor peculiar to a particular architecture, training corpus, or task.⁴⁰ Our results are summarized in Table 1.

Also recall that each conlang family removes a different, more superficial shortcut as to what statistical patterns are available as potential explanations. In brief, the story family weakens prompt exposure as a sufficient explanation, the double verb family rules out simple semantic coherence as a sufficient explanation, and the translation family rules out even complex statistical explanations aimed at producing conlang-like surface forms by requiring preservation of specific relationships under tightly constrained transformations. Because the translation tasks additionally have explicit answer keys, we can directly verify that those relationships were occasionally perfectly preserved. The convergence of these independent dissociations is stronger than any individual result in isolation. The central result is therefore the kind of behavior the models exhibit (not their performance rankings): they produce some text that is plausible only under the meaning-mediated interpretation of the prompt. **Without that meaning, the outputs are locally bizarre and unpredictable, but taking meaning into account, they are systematic and predictable.**

Several caveats are important. We assess rule-following at the output level, so our experiments cannot determine whether failures reflect lack of rule representation or difficulty retrieving and producing the appropriate form. Conversely, successful output behavior supports an abstract-representation explanation without uniquely identifying its implementation. Instead, we construct our experiment to rule out competing superficial patterns as viable explanations, such that successful rule-following is best explained by some higher-level form of plausibility overriding lower-level local continuation pressures. We therefore use terms such as “meaning-mediated” functionally, not as claims about human-equivalent subjective understanding. Future work could test this proposed mechanism more directly using internal-state analyses.

Those and earlier caveats aside (Section 3.4), we did design the rules with conjectured difficulty orderings in mind. At the family level, we expected the story rules to be easiest, followed by the double verb rules, followed by the translation rules. Within each family, we also had expectations about the relative difficulty of individual rules: comparison-1 < sidequest-1 < comparison-2; doublev-1 < doublev-2; and translate-1 $\lesssim$ dan-1. However, we were less confident about the relative ordering of individual rules across different families. We therefore expected a family-level ordering and a within-family ordering, but not necessarily transitivity across those levels into a global ordering.⁴¹

This intuition as to rule difficulty is imperfectly reflected in our binary feature-based complexity metric. Because all features are weighted equally, for example, dan-1 is assigned a lower complexity

⁴⁰We take no further stance on what the internal representations are like, nor how they compare to human cognition; we argue only that models are operating somewhere above the lower bound implied by purely superficial pattern-matching.

⁴¹In other words, the space between some rules in the same family might be quite large, and the space between others relatively closer together.

score than translate-1 despite our expectation that it may be more difficult for models because of the total remapping of tokens required.⁴²

That model performance broadly tracks these expected family-level and within-family difficulty relationships is itself suggestive of meaning-mediated generation. If successful performance were primarily driven by superficial statistical patterns, it is unclear why the availability of such patterns would track independent representations of rule difficulty or rule complexity. By contrast, a meaning-mediated account naturally predicts that performance should degrade as the conceptual complexity of a rule increases.

Notably, the three families implicate different forms of meaning-mediated structure. The story rules involve something akin to a pragmatic distinction between use and mention, the double verb rules involve syntactic, morphological, and semantic manipulation, and the translation rules involve preserving relational structure via multiple composed constraints at various linguistic levels across dramatically different surface forms. The convergence of evidence across these distinct domains strengthens the case against explanations based solely on superficial statistical regularities.

Overall, we see the ordering of model performance as basically Qwen (thinking) > or $\approx$ GPT-4o $\approx$ Qwen (no thinking) > OLMo $\approx$ Llama, with all models > random for translate-1, but only Qwen (thinking) at all performant on dan-1, although GPT-4o and Qwen (no thinking) were at least distinguishable from random performance.

Llama is the only model in our set that approaches a purely plausible-text-generation architecture. Overall, we interpret Llama as showing signs of (imperfect) rule-following, and, more specifically, of occasionally overriding concrete statistical priors in favour of plausibility mediated at more abstract levels. We view this behaviour as more-or-less emergent,⁴³ albeit with substantially greater variability and less consistent evidence than seen in the strongest-performing models.

This is why the strong stochastic-parrot hypothesis and distributional accounts of meaning should not be treated as simple opposites: distributional evidence is one way systems can arrive at increasingly abstract representations. Our takeaway is not that models escape statistics, but that statistical learning can produce abstractions, and that later training may help surface or prioritize those abstractions when they conflict with lower-level plausibility pressures (consistent with [Hewitt et al. \[2024\]](#)).⁴⁴

4.1 LLMs are not stochastic parrots, but they can act like them: the cognitive constraints of PTGA

In a purely plausible-text-generation architecture, the ability to override lower-level plausibility pressures in favour of more abstract ones would be emergent, arising as the model bootstrapped increasingly abstract forms of plausibility from an objective defined at approximately token-scale.⁴⁵ The story rules may provide an example consistent with that description. Rare bigrams presented as prompt content produce very different behavior from rare bigrams presented as the operational components of a language. We cannot attribute that difference solely to the change in communicative role, since the prompts are not otherwise identical, but exposure to the forms themselves is clearly not sufficient to explain the observed behavior.

⁴²We preferred a bottom-up representation of complexity rather than constructing a metric guaranteed to reproduce our prior expectations about difficulty. As discussed above, many alternative representations are possible.

⁴³Plausibility is optimized for at approximately the token scale, so we describe plausibility that appears to be driven by aspects of meaning significantly beyond that scale as more-or-less emergent. Especially in architectures so different from our own, it is not obvious that next-token prediction would give rise to representations of this kind.

⁴⁴A very speculative thread to be pulled in future work: Depending on how you look at it, these results could place pressure on strong poverty-of-the-stimulus or Universal Grammar arguments. LLMs have radically different, simpler architectures than we do, without obvious specialized areas analogous to most of ours, and completely different learning histories, yet nevertheless appear able to bootstrap from concrete statistical patterns to behaviour better explained by meaning-mediated abstractions than by superficial statistical regularities. Their representations must be built entirely out of inferential lexical (or token) relationships. We do not think our results establish any particular theory of language acquisition or cognition. However, they may suggest that statistical learning can support richer abstract representations than strong versions of those arguments permit, raising broader questions about how much of linguistic competence and even cognition can emerge from distributional experience, and about the extent to which language and thought can ultimately be dissociated.

⁴⁵We are imagining something like a pattern-to-thought transition where increasingly abstract structures emerge from relationships among more concrete structures, perhaps along the lines of [Selfridge \[1959\]](#), [Bradley et al. \[2021\]](#), or even aspects of [Miller et al. \[2026\]](#). We imagine this process as broadly consistent with the Distributional Hypothesis [[Harris, 1954, 1968](#), [Sahlgren, 2008](#)] and the relation/locality account in [Zimmerman \[2025\]](#).

But it is not just surface-form-related priors that must sometimes be overridden in order to produce meaning-mediated behaviour. In the double verb rules, we observe a large directional response to an abstract semantic relation conveyed in the instruction. **This moves our inference about the relevant underlying patterns further up the ladder of abstraction.** The relevant pattern is not token co-occurrence, but more abstract: the model apparently learned semantic representations for a set of verbs via statistical patterns including token co-occurrence, and now acts based on a relationship comparing those representations when prompted to do so. In other words, distributional learning may have constructed a semantic space, and the prompt refers to a relation within that space rather than to any particular token sequence. The model therefore changes generation in the requested direction, demonstrating plausibility operating at a level above simple token co-occurrence, consistent with the results of [Hewitt et al. \[2024\]](#).

This distinction is important because semantic coherence is itself a powerful plausibility prior. Although more abstract than token co-occurrence, it remains a prominent statistical pattern in natural-language data. The SSP must accommodate some abstractions of this kind, for example, those needed to support neologisms⁴⁶, but generally maintains that models remain very close to the lower bound of superficial pattern matching. The double verb results suggest otherwise: models appear capable not only of learning semantic relationships from statistical patterns, but also of acting on those relationships in accordance with novel instructions.

Furthermore, both the contrast between dan-1 and translate-1 and the relative difficulty of comparison-2 are suggestive of a less-discussed aspect of PTGA: the important role that tokens specifically must play in these architectures, acting as distributional, organizational, and epistemic primitives. More generally, we see an exploratory correspondence between difficulty and output that does not re-use familiar English forms in familiar ways (Section 3.4). More exploration of PTGA as a strong instantiation of the primacy of language for thought can be found in [Zimmerman et al. \[2024b\]](#) and [Zimmerman \[2025\]](#), but the short version is that operations defined at approximately the token scale may naturally privilege approximately token-scale structures.⁴⁷

This makes the more concrete aspects of meaning – such as morphology, syntax, and perhaps some semantics – relatively perceptible to the model, since they are often legible on the scale of words and morphemes (like many tokens). Larger structures, often commensurate with aspects of meaning scaffolding discourse and pragmatics, may therefore be more “out of focus” for PTGA (Figure 30).⁴⁸ Consistent with this interpretation, in forthcoming work we find that discourse-level pragmatic plausibility is one facet of linguistic performance in which people perceive current models to be deficient [[Morgenstern et al., 2026](#)].

The translation rules provide evidence consistent with this account. Rule-following becomes noticeably harder when existing structure can’t be used in familiar ways. Tasks requiring output that departs substantially from familiar tokenization patterns (such as orthography, punctuation, and word-boundary conventions) are more difficult, with some models ultimately demonstrating no discernible rule-following (by the metrics we used). Generally, rules that re-use familiar tokens together with some familiar aspects of meaning or function appear easier to execute than rules requiring extensive remapping of those structures. This suggests that tokenization and distributional priors don’t just support generation, but also constrain it, even when models can exhibit meaning-mediated behaviour (Figure 26, Table 36, Table 37).

⁴⁶Recent work explores *neologism learning* which introduces new meanings for existing words by introducing a new embedding, with no other changes in the model parameters [Hewitt et al. \[2025\]](#). [Hewitt et al. \[2025\]](#) show that models can self-verbalize the meaning of these neologisms, even when they fall outside of human familiarity (e.g. even when the word has no direct, tangible corollary in common English), demonstrating that LLMs seem to have a multifaceted capacity to work with neologisms (as new concepts and new symbols). In addition, anecdotally, it is not difficult to get an LLM to generate a credibly novel surface form (for example, prompting a model to combine Greek, Latin, French, or Germanic roots to create new jargon for a specific concept). Even if some of those neologisms are not truly novel, once you encounter enough of them, it seems fairly plausible that at least one of them is.

⁴⁷Characters-per-token provides a rough empirical illustration. Outputs that depart substantially from familiar tokenization patterns are harder – here meaning computationally more intensive – for all models compared to other texts of the same length. English-like outputs (comparison-1, sidequest-1) have higher characters-per-token because tokenizers chunk familiar English stems efficiently; conlang outputs (especially dan-1) fragment into far more tokens over the same length of text, requiring more forward passes and more resources [[Gladstone et al., 2025](#), [Wang et al., 2025](#)]. See Appendix 2.5. This also provides indirect evidence that models succeeding on these tasks are not simply recombining familiar token sequences: the correct outputs are structurally unfamiliar at the level at which PTGA operates.

⁴⁸Moreover, the model directly perceives only the inferential traces of such structures within language, not their extradiegetic aspects [[Zimmerman et al., 2024a](#)].

4.2 Be wary of benchmarks

The difficulty of our constructed-language tasks is notable because the models evaluated here have performed strongly on established reasoning and language-understanding benchmarks. For example, GPT-4 [OpenAI et al., 2024] achieved 86.4% on MMLU, a benchmark of knowledge and reasoning across academic subjects [Hendrycks et al., 2020], and 92.0% on GSM8K, which evaluates multi-step mathematical reasoning [Cobbe et al., 2021]. It also achieved 95.3% on HellaSwag, a commonsense sentence-completion benchmark [Zellers et al., 2019], 87.5% on WinoGrande, which evaluates commonsense pronoun resolution, and an F1 score of 80.9 on DROP, a reading-comprehension benchmark requiring discrete reasoning [Dua et al., 2019]. Llama-3-70B also performs strongly across standard evaluations [AI@Meta, 2024]: the base model achieved 79.5% on MMLU, 81.3% on BIG-Bench Hard, which tests challenging multi-step and algorithmic reasoning [Suzgun et al., 2023], 93.0% on ARC-Challenge, a difficult grade-school science reasoning benchmark [Clark et al., 2018], 85.6% on SQuAD, a reading-comprehension benchmark [Rajpurkar et al., 2016], and 79.7 on DROP. Qwen3-32B [Yang et al., 2025] in thinking mode reports 90.9% on MMLU-Redux, a revised version of MMLU [Gema et al., 2025], 68.4% on GPQA, a graduate-level scientific reasoning benchmark, and 81.4% on AIME 2024, a competition-level mathematical reasoning benchmark [AIME, 2025] whereas its non-thinking mode achieves 85.7% on MMLU-Redux and 54.6% on GPQA. OLMo-2-32B-Instruct also performs competitively [OLMo et al., 2024], reaching 87.6% on GSM8K, 70.6% on BIG-Bench Hard, 78.0 on DROP, 77.3% on MMLU, and 85.6% on IFEval, which evaluates adherence to verifiable natural-language instructions [Zhou et al., 2023].

One explanation for the disparity in performance is that our constructed-language tasks simply do not probe abilities that fall under reasoning or language competence.⁴⁹ We don’t find this explanation compelling, because we view meaning-sensitive abstraction as a necessary subset of many things commonly called reasoning.⁵⁰

On the other hand – and consistent with work such as GSM-Symbolic, which shows that strong benchmark performance can degrade under controlled modifications of GSM8K-style problems [Mirzadeh et al., 2024] – we view a lack of sufficient construct validity in at least some of these benchmarks as the more plausible explanation. Many benchmarks do not clearly establish that task success requires the capability they purport to measure.⁵¹ As a result, benchmarks do not necessarily measure what they claim to measure.

This implies that evaluating output-level LLM behaviour against human performance or answer keys is not enough. Deploying LLMs requires intense scrutiny through evaluation processes tailored to the intended deployment context [Schaeckermann et al., 2026]. While we do not recommend our conlangs as new benchmarks, we do view them as demonstrations that creative task design can reveal gaps, assumptions, and construct-validity failures in existing benchmarks.

4.3 Implications for model development

In Section 3.4, exploratory analyses suggest that model properties related to training and cognitive strategy may be more predictive than raw scale for performance on these tasks. Although highly preliminary, this hints that **architectural and training interventions may be more effective than scaling alone** for improving performance on certain kinds of reasoning tasks⁵²: Qwen3-32B with thinking enabled outperforms GPT-4o on specific tests – and arguably in general – despite being much smaller.

These observations, together with Section 4.1, suggest two related ways in which PTGA may tether generation too strongly to superficial plausibility. First, PTGA may disproportionately privilege relatively superficial structures as the determinants of plausible continuation, even when semantic, discourse-level, or otherwise more abstract relationships should govern generation instead. Second, PTGA may bind linguistic forms too tightly to the functions with which they were learned during

⁴⁹And that existing benchmarks either do, or fail in different ways.

⁵⁰We do not claim to demonstrate “reasoning”, but we do show meaning-sensitive abstraction.

⁵¹Validity may fail at multiple levels. For example, benchmark designers may insufficiently consider the effects of architecture or embodiment on task performance, or may rely on poorly specified notions of the underlying human capability in the first place.

⁵²The kind of reasoning these tasks probe, whatever that might be called, and perhaps reasoning more generally.

training, making familiar functions harder to express through unfamiliar forms.⁵³ In both cases, superficial patterns – implicating the token-scale perspective of PTGA – overly dominate what the model learned, skewing the model towards the shallow end of the meaning spectrum.

We speculate that interventions such as chain-of-thought reasoning, instruction tuning, and fine-tuning help by making different relationships salient to the model, allowing more abstract structures⁵⁴ to exert greater influence on what counts as a plausible continuation.⁵⁵ While this most directly addresses the first constraint, it may also loosen learned associations between familiar forms and functions. The imputed similar cognitive mechanism underlying these approaches implies that the relevant limitation is not necessarily whether a model has learned an abstraction, but whether that abstraction can be used independently of the surface forms through which it was learned.⁵⁶

This possibility suggests a connection between those interventions and other cognitive strategies such as joint-embedding architectures, multimodal training, and vision-language contrastive learning. Such methods may succeed partly because they force representational unity across more heterogeneous observations, requiring the model to preserve abstractions that transcend particular surface forms. In that sense, they partially separate form from function.

They also imply another direction: that scale or resolution may itself be understood as a kind of mode.⁵⁷ Vision is conventionally treated as one modality, but close inspection and perception at a distance require different lenses and make different relationships salient (Figure 30). Language likewise supports multiple scales of focus, from morphemes to narrative. **What is cognitively local depends on the scale of representation:** relationships distant at the token level may become local within a more abstract representation.⁵⁸ PTGA may therefore be limited not only by its reliance on concrete linguistic forms, but also by its reliance on a relatively fixed representational resolution. A promising direction is to develop systems that can shift scale, allowing different relationships to come into focus as the task demands.

Scaling⁵⁹ assumes that exposing models to more language gives them access to more knowledge. But if that language is interpreted primarily through superficial plausibility, much of the higher-level meaning it encodes for human speakers may remain inaccessible. Our results suggest a different possibility: **The bottleneck may not be knowledge acquisition, but control over which kind of plausibility governs generation.** By contrast, people appear able to shift fluidly between linguistic abstractions not unlike how they can shift visual focus. Formulaic exchanges can often be handled through routine lexical and syntactic patterns, whereas interpreting a poem may require simultaneous attention to referential, inferential, indexical, iconic, prosodic, and pragmatic significance. Both involve the same linear sequence of linguistic forms, but they recruit very different levels of interpretation. Consequently, we suspect that statistical plausibility is likewise a powerful and efficient proxy in human cognition, but one that can be overridden. We may generally rely on the least effortful level that suffices, moving from familiar surface patterns toward semantic and pragmatic reasoning when lower-effort cues fail, when different kinds of plausibility conflict, or depending on what aspects of meaning are consciously accessible. Without committing to the details, it is clear that different forms of linguistic competence help us determine which interpretations and utterances are most plausible in different contexts.⁶⁰

⁵³Models may acquire useful representations of functions such as morphological marking, syntactic ordering, and relational mapping between verbs and objects, yet struggle to apply those functions through unfamiliar tokens and surface structures. Somehow, the representations themselves may be too concrete, too entangled with their forms.

⁵⁴More abstract aspects of meaning.

⁵⁵The reason some computer scientists are seemingly independently rediscovering Grice’s maxims via chain-of-thought prompting may be specifically because CoT works by directing focus toward intermediate and higher levels of structure, including some pragmatic and discourse level ones [Grice, 1989, Zimmerman, 2025, Zimmerman et al., 2024b]. Grice’s maxims, after all, are essentially a model for how speakers decide what utterances can plausibly come next in a conversation. See, for examples of Gricean logic underlying computer science methods: Cheng et al. [2025], Wei et al. [2025].

⁵⁶In other words, whether the representation is sufficiently stripped of the concrete forms of the observations that gave rise to it so as to function independently of them.

⁵⁷All of these approaches might owe their successes to a metaphorical “multimodality”, which likewise implies that scale could be further exploited in other domains. These approaches may work by changing what kinds of relations become salient for the model – by changing the resolution at which the model perceives – and maybe eventually, by allowing the model to metacognitively reflect on the possibility of different resolutions.

⁵⁸Just as information in non-adjacent layers becomes architecturally local under the addition of skip connections.

⁵⁹As a strategy for improving model performance.

⁶⁰Training objectives operating across multiple linguistic scales, or mechanisms that distinguish and coordinate different modes of plausibility, might therefore offer a promising direction. Since token structure may currently scaffold higher-level

Overall, our results suggest that **although LLMs are not bounded by superficial statistical pattern matching, PTGA may give models access to meaning across multiple levels while simultaneously anchoring generation disproportionately to the most superficial and fungible among them.** Plausible-text generation remains a strong default prior whose influence is not yet reliably modulated when different kinds of plausibility conflict. Future work should ask when models shift among plausibility regimes and what training or architectural changes make such shifts possible.

5 Acknowledgements

The authors are grateful for conversations with Ashley M. A. Fehr, William Wheeler, Grayson Wycliffe Storer, Neil Traft, Fitzwilliam Keenan-Koch, Kam Bielawski, Collin Coil, Thomas Pashby, Danny Benett, Alejandro Javier Ruiz Iglesias, Nate Gaylinn, Katie Ekstrom Grenon, Dave Jangraw, Tanis Sheehan, Jim Lawson, Evan Hinchliffe, Milo Z. Trujillo, Christopher M. Danforth, Peter Sheridan Dodds, Alice Patania, Juniper L. Lovato, Kyle Mahowald, Iuliia Kotseruba, Paul Thagard, and for support furnished by the National Science Foundation (Grant #2242829). In addition, we are grateful to ISC Summer School 2026 [[Institut des sciences cognitives, Université du Québec à Montréal, 2026](#)], IC2S2 2026 [[12th International Conference on Computational Social Science, 2026](#)], and SLSA/ 4S 2026 [[Society for Literature, Society, and the Arts, 2026](#), [Society for Social Studies of Science, 2026](#)] for the opportunity to discuss this and related works, and to get valuable feedback from other participants.

linguistic operations (e.g. grammatical and semantic), revisiting the fixed scale at the heart of PTGA – and asking what would change at other scales – seems a promising direction.

References

- 12th International Conference on Computational Social Science. IC2S2. <https://ic2s2-2026.org/>, 2026. Burlington, Vermont. July 28-31.
- Asif Agha. The social life of cultural value. *Language & Communication*, 23(3-4):231–273, 2003. doi: 10.1016/S0271-5309(03)00012-0. URL [https://doi.org/10.1016/S0271-5309\(03\)00012-0](https://doi.org/10.1016/S0271-5309(03)00012-0).
- AIME. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2025.
- AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, Maria Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luetzgau, Jabez Magomere, Jonathan Rystrom, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Nicolaus Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL <https://openreview.net/forum?id=mdA5IVvNcU>.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL <https://doi.org/10.1145/3442188.3445922>.
- Aakanksha Bomble, Prashant Kulkarni, and Vishnu Potdar. Does prompt design affect llm outputs? a study across structured and vague inputs. *International Journal of Innovative Research in Technology (IJIRT)*, 12(3):544, August 2025. ISSN 2349-6002. URL https://ijirt.org/publishedpaper/IJIRT183166_PAPER.pdf. IJIRT 183166.
- Tai-Danae Bradley, John Terilla, and Yiannis Vlassopoulos. An enriched category theory of language: from syntax to semantics, 2021. URL <https://arxiv.org/abs/2106.07890>.
- Zhengxiang Cheng, Dongping Chen, Mingyang Fu, and Tianyi Zhou. Optimizing length compression in large reasoning models, 2025. URL <https://arxiv.org/abs/2506.14755>.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL <https://arxiv.org/abs/1803.05457>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>, 9, 2021.
- Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, page 36342–36389. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024. doi: 10.52202/079017-1146. URL <http://dx.doi.org/10.52202/079017-1146>.
- Justin Diamond. "genlangs" and zipf's law: Do languages generated by chatgpt statistically look human?, 2023. URL <https://arxiv.org/abs/2304.12191>.
- Jan Digutsch and Michal Kosinski. Overlap in meaning is a stronger predictor of semantic activation in gpt-3 than in humans. *Scientific Reports*, 13, 2023. doi: <https://doi.org/10.1038/s41598-023-32248-6>.

- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, 2019.
- Penelope Eckert. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476, 2008. ISSN 1360-6441. doi: 10.1111/j.1467-9841.2008.00374.x. URL <https://doi.org/10.1111/j.1467-9841.2008.00374.x>.
- Richard Futrell and Michael Hahn. Linguistic structure from a bottleneck on sequential information processing, 2024.
- Richard Futrell, Edward Gibson, and Roger P. Levy. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3):e12814, 2020a. doi: 10.1111/cogs.12814. URL <https://doi.org/10.1111/cogs.12814>.
- Richard Futrell, Roger P. Levy, and Edward Gibson. Dependency locality as an explanatory principle for word order. *Language*, 96(2):371–412, 2020b. doi: 10.1353/lan.2020.0024.
- Zeyu Gan, Ruifeng Ren, Wei Yao, Xiaolin Hu, Gengze Xu, Chen Qian, Huayi Tang, Zixuan Gong, Xinhao Yao, Pengwei Tang, Zhenxing Dou, and Yong Liu. Beyond the Black Box: A Survey on the Theory and Mechanism of Large Language Models, March 2026.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, 2025.
- Edward Gibson, Richard Futrell, Steven T. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger P. Levy. How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5):389–407, 2019. doi: 10.1016/j.tics.2019.02.003. URL <https://colala.berkeley.edu/papers/gibson2019how.pdf>.
- James J. Gibson. *The Ecological Approach to Visual Perception*, chapter 8. Psychology Press, New York and London, classic 2015 edition, 1979. ISBN 978-1-84872-577-5 978-1-84872-578-2 978-1-315-74021-8. URL <https://yzhu.io/courses/core/reading/02.gibson.pdf>. Chapter title: The Theory of Affordances.
- Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers, 2025. URL <https://arxiv.org/abs/2507.02092>.
- Antonino Greco, Julia Moser, Hubert Preissl, and Markus Siegel. Predictive learning shapes the representational geometry of the human brain. *Nature Communications*, 15:9670, 2024. doi: 10.1038/s41467-024-54032-4. URL <https://doi.org/10.1038/s41467-024-54032-4>.
- H. P. Grice. *Studies in the way of words*. Harvard University Press, Cambridge, MA, 1989. ISBN 9780674852716.
- Michael Hahn, Dan Jurafsky, and Richard Futrell. Universals of word order reflect optimization of grammars for efficient communication. *PNAS*, 117:2347–2353, 2020. doi: <https://doi.org/10.1073/pnas.1910923117>.
- Zellig S. Harris. Distributional structure. *WORD*, 10:146–162, 1954. doi: 10.1080/00437956.1954.11659520. URL <https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520>.
- Zellig S. Harris. *Mathematical Structures of Language*. Interscience Publishers, New York, 1968.
- Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance?, 2024. URL <https://arxiv.org/abs/2411.10541>.

- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- John Hewitt, Nelson F. Liu, Percy Liang, and Christopher D. Manning. Instruction following without instruction tuning, 2024. URL <https://arxiv.org/abs/2409.14254>.
- John Hewitt, Oyvind Tafjord, Robert Geirhos, and Been Kim. Neologism learning for controllability and self-verbalization, 2025. URL <https://arxiv.org/abs/2510.08506>.
- Qian Huang, Eric Zelikman, Sarah Li Chen, Yuhuai Wu, Gregory Valiant, and Percy Liang. Lexin-variant language models, 2023. URL <https://arxiv.org/abs/2305.16349>.
- Institut des sciences cognitives, Université du Québec à Montréal. ISC École d’été 2026 : Connaissance, raisonnement et prise de décision. <https://sites.grenadine.uqam.ca/sites/isc/fr/iscuqam2026/home>, 2026. Summer school held May 27–June 10, 2026, Université du Québec à Montréal, Montréal, Québec, Canada. Accessed July 7, 2026.
- Judith T. Irvine and Susan Gal. Language ideology and linguistic differentiation. In Paul V. Kroskrity, editor, *Regimes of Language: Ideologies, Politics, and Identities*, pages 35–83. School of American Research Press, Santa Fe, NM, 2000. URL <https://web.stanford.edu/~eckert/PDF/IrvineGal2000.pdf>.
- Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. Mission: Impossible language models, 2024.
- Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024a.
- Jinlu Liu, Nan Yang, and Haitao Liu. Distribution of sentence length of english complex sentences. *Moderna Språk*, 118(3):51–69, 2024b. doi: 10.58221/mosp.v118i3.15574.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- Juniper Lovato, Julia Witte Zimmerman, Isabelle Smith, Peter Dodds, and Jennifer L. Karson. Foregrounding artist opinions: A survey study on transparency, ownership, and fairness in ai generative art. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):905–916, 2024. doi: 10.1609/aies.v7i1.31691. URL <https://doi.org/10.1609/aies.v7i1.31691>.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models, 2024. URL <https://arxiv.org/abs/2301.06627>.
- Evie A. Malaia, Julia Krebs, Dietmar Roehm, and Ronnie B. Wilbur. Age of acquisition effects differ across linguistic domains in sign language: EEG evidence. *Brain and Language*, 200: 104708, January 2020. doi: 10.1016/j.bandl.2019.104708. URL <https://pubmed.ncbi.nlm.nih.gov/31698097/>.
- M. S. Mayzner and Margaret Elizabeth Tresselt. *Tables of Single-letter and Digram Frequency Counts for Various Word-length and Letter-position Combinations*, volume 1 of *Psychonomic Monograph Supplements*. Psychonomic Press, Austin, TX, 1965.
- R. Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L. Griffiths. Embers of autoregression: Understanding large language models through the problem they are trained to solve, 2023. URL <https://arxiv.org/abs/2309.13638>.
- Earl K. Miller, Scott L. Brincat, and Jefferson E. Roy. Analog cognition and consciousness. *The Journal of Neuroscience*, 46(33):e0711262026, 2026. doi: 10.1523/JNEUROSCI.0711-26.2026. URL https://osf.io/preprints/psyarxiv/z48x7_v1.

- Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL <https://arxiv.org/abs/2410.05229>.
- Michelle Morgenstern, Julia Witte Zimmerman, Tabia Tanzin Prama, Juniper L. Lovato, Bryn Loftness, Calla Beauregard, Kristine Harootunian, and June Clark. Contextuality and contradictions: Exploring the indexicality of ai-generated content, 2026. Presented in talk form at IC2S2 2026 and 4S 2026. Paper in progress.
- Peter Norvig. English letter frequency counts: Mayzner revisited or ETAOIN SRHLDCU. <https://www.norvig.com/mayzner.html>, 2013. Accessed: 2026-08-31.
- Team OLMO, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious. *arXiv preprint*, 2024. URL <https://arxiv.org/abs/2501.00656>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and [...] Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- Charles S. Peirce. The Collected Papers of C.S. Peirce, Vol. I: The Principles of Philosophy, 1931. URL http://www.commens.org/bibliography/anthology_volume/peirce-charles-s-1931-collected-papers-charles-sanders-peirce-vol-i. From the Commens Bibliography.
- Arianna Pera, Mauro Martino, Nima Dehmamy, Douglas Guilbeault, Luca Maria Aiello, and Andrea Baronchelli. Billions of sketches reveal hidden cultural variation in human concepts, 2026. URL <https://arxiv.org/abs/2607.07267>.
- Rolf Pfeifer and Josh Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. A Bradford Book, MIT Press, London, England and Cambridge, Massachusetts, 2006. ISBN 978-0262162395.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392, 2016.
- Amon Rapp, Chiara Di Lodovico, Federico Torrielli, and Luigi Di Caro. How do people experience the images created by generative artificial intelligence? an exploration of people’s perceptions, appraisals, and emotions related to a gen-ai text-to-image model and its creations. *International Journal of Human-Computer Studies*, 193:103375, 2025. doi: 10.1016/j.ijhcs.2024.103375. URL <https://doi.org/10.1016/j.ijhcs.2024.103375>.
- Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20:33 – 53, 2008. URL <https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf>.
- Mike Schaekermann, Anil Palepu, Adam Rodman, Ami Parekh, Ethan Goh, Shekoofeh Azizi, Yun Liu, Sunny Virmani, Christina Chen, Dale R. Webster, Joëlle Barral, Avinatan Hassidim, Yossi Matias, and Po-Hsuan Cameron Chen. Prospective evidence for conversational medical AI is hard, but non-negotiable. *Nature Medicine*, September 2026. doi: 10.1038/s41591-026-04639-5. URL <https://www.nature.com/articles/s41591-026-04639-5>. Comment.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models’ Sensitivity to Spurious Features in Prompt Design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html.

- O. G. Selfridge. Pandemonium: A paradigm for learning. In *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory, November 1958*, pages 513–526, London, 1959. HMSO. URL <https://gwern.net/doc/ai/nn/1959-selfridge.pdf>.
- Michael Silverstein. Indexical order and the dialectics of sociolinguistic life. *Language and Communication*, 23(3–4):193–229, July–October 2003. doi: 10.1016/S0271-5309(03)00013-2. URL [https://doi.org/10.1016/S0271-5309\(03\)00013-2](https://doi.org/10.1016/S0271-5309(03)00013-2).
- Society for Literature, Society, and the Arts. SLSA Streams at 4S. <https://litsciarts.org/>, 2026. From October 8-10, 2026, 4S will host the 39th meeting of the Society for Literature, Science and the Arts (SLSA), for a series of integrated streams.
- Society for Social Studies of Science. 4S. <https://4sonline.org/>, 2026. 50th Anniversary for the Annual Meeting of the Society for Social Studies of Science, Toronto, Canada. October.
- Christoph Stöckl, Yukun Yang, and Wolfgang Maass. Local prediction-learning in high-dimensional spaces enables neural networks to plan. *Nature Communications*, 15:2344, 2024. doi: 10.1038/s41467-024-46586-0. URL <https://doi.org/10.1038/s41467-024-46586-0>.
- Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. Prompt chaining or stepwise prompt? refinement in text summarization. *arXiv preprint arXiv:2406.00507*, 2024. doi: 10.48550/arXiv.2406.00507. Accepted to Findings of ACL 2024.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.
- Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. Hierarchical reasoning model, 2025. URL <https://arxiv.org/abs/2506.21734>.
- Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen tau Yih, and Xin Luna Dong. Truthrl: Incentivizing truthful llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2509.25760>.
- Ethan Wilcox, Roger Levy, and Richard Futrell. Hierarchical representation in neural language models: Suppression and recovery of expectations. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4819. URL <https://aclanthology.org/W19-4819/>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Frances Yates. *The Art of Memory*. University of Chicago Press, 1966.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.
- Julia Witte Zimmerman. *Locality, Relation, and Meaning Construction in Language, as Implemented in Humans and Large Language Models (LLMs)*. Ph.d. dissertation, University of Vermont, 2025. URL <https://hdl.handle.net/20.500.14849/4063>. Available at ScholarWorks@UVM.
- Julia Witte Zimmerman, Denis Hudon, Kathryn Cramer, Jonathan St. Onge, Mikaela Fudolig, Milo Z. Trujillo, Christopher M. Danforth, and Peter Sheridan Dodds. A blind spot for large language models: Supradiegetic linguistic information. *Plutonics*, 17:107–156, 2024a. URL <https://plutonicsjournal.com/volume-xviii/>.

Julia Witte Zimmerman, Denis Hudon, Kathryn Cramer, Alejandro J. Ruiz, Calla Beauregard, Ashley Fehr, Mikaela Irene Fudolig, Bradford Demarest, Yoshi Meke Bird, Milo Z. Trujillo, Christopher M. Danforth, and Peter Sheridan Dodds. Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning, 2024b. URL <https://arxiv.org/abs/2412.10924>. Presented as talks at IC2S2 2026 and SLSA at 4S 2026.

Julia Witte Zimmerman, Bryn Loftness, Calla Beauregard, Kristine Harootunian, June Clark, Hannah Reing, Katie Ekstrom Grenon, Michelle Morgenstern, Tabia Tanzin Prama, and Juniper L. Lovato. Vermont AI Literacy project, 2026. Ongoing open-source curriculum and materials, co-designed and piloted Spring 2026 - Summer 2027. Paper in progress.

George Kingsley Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. The MIT Press, Boston, MA, 1965. ISBN 9780262740029. Originally published in 1935; this edition published in 1965.

A Appendices and supplementary material

Technical appendices with additional details, results, figures, graphs, and discussion.

A.1 Disclosures

- Much of our work is saved in a private GitHub repository, in a series of Google spreadsheets and docs, in Colab Notebooks, and in draft and note form in several Overleaf projects. We will provide access to any of these materials to interested parties upon reasonable request. The team also communicated via slack channel and in-person meetings. Upon reasonable request, J.W.Z. would be happy to answer any questions about process or methods.
- Besides using Generative AI models as subjects in our experimental pipeline, J.W.Z. used an institutional Copilot account at various points throughout this work. C.G.B. used ChatGPT and Claude Opus 4.8 for code audit and formatting. Other authors used Generative AI for additional tasks.

A.2 Limitations and Future Work

While this paper tests the strong stochastic parrot argument in Large Language Models, we focus on relatively few models, ranging from older (Llama3-70B) to newer models (Qwen-3 (Thinking)), open to closed models, and simpler to more complex architectures. Additionally, we selected two separate thinking modes for Qwen-3 on the presumption that changing that setting would be likely to impact results. We selected well-established baseline settings for each model. Based on our preliminary results here, future work could sweep across other model parameters, and incorporate other models, to investigate the breadth of relevant behavior across LLMs under different conditions.

Small changes in formatting and wording, as in prompt engineering, have been shown to impact model behavior across various tasks [Bomble et al., 2025, He et al., 2024]. Here, while we provide three separate wording variations for each rule prompt, we expect that very different wording prompting the same rules would likely impact results. Future work could explore different variations of our prompt structures, or use iterative prompt chaining methods which may increase performance Sun et al. [2024].

Regarding our choice of performance metrics, future work will also explore output length, either in terms of characters, words, or sentences. Overall, we hypothesize that *valid* output length should decrease with harder tasks, demonstrating resource tradeoffs (as one aspect of a task requires more effort, other aspects of the task which the model is otherwise capable of may be degraded).⁶¹ However, because we observe that model outputs can fall into degenerative cycles, suggestive of something like strange attractors in the latent space, we don’t expect output length will always be straightforwardly interpretable.

Although we did what we could to prevent these five models from being exposed to our rules ahead of time, and to keep each run independent, because of the black box nature of the commercial models,

⁶¹We can use our control outputs as a baseline as well as previously-reported sentence lengths [Liu et al., 2024b].

we can't truly guarantee this. During our initial pilot testing, ChatGPT seemingly recalled aspects of the project weeks later (fragments of a rule), so it is possible, for example, that ChatGPT flags unusual tasks internally for use by OpenAI, OpenAI sells that training data, and that data somehow made its way into one of these models before we collected our data (given the project time frame, and the models we used for the initial piloting phase versus the data-generation phase, we think that is very unlikely, but we cannot guarantee it).

Some extra runs, GPT-4o re-runs (from an initial confusing system prompt), and the aforementioned sidequest-1-fix runs exist, but are excluded from the analyses here.

Additional plans for future work are mentioned throughout the rest of the text where relevant.

A.3 Construct validity and additional methodological detail

After coming up with our plan and our rules, we used the checklist for LLM benchmark construct validity developed by [Bean et al. \[2025\]](#) to help us think through our choices, augment gaps, and revise ambiguous steps. Here we enumerate the checklist items and our answers:

Define the phenomenon: Our goal is to ascertain whether or not any kind of LLM can produce output that evinces compelling evidence of abstraction-based rule-following, where the rules are both (1) within what we theoretically believe LLMs could have learned, and (2) negligibly likely to have been seen during training – and beyond not being seen verbatim, we want the rules to be significantly dissimilar to any instruction, task, rule, example, or problem that was described at any point during training. The rules themselves need to be designed s.t. any parsimonious explanation invokes abstractions, and cannot invoke pattern-matching at a literal or near-literal level (direct observation), notably by disrupting the connection between the statistical patterns shown in correct output versus those observed during training. Thus, we need to make compelling arguments for both (3) our choice of rules, and (4) how we are assessing whether or not the rules are being followed (what we accept as evidence of rule-following).

- Provide a precise and operational definition for the phenomenon being measured: In order to determine if LLM output shows rule-following or not, we give each model 50 attempts (independent from each other; no in-context learning) to respond to each rule. Each rule has 3 variations (although we accidentally repeated one of the prompt variants for sidequest-1, so it actually has 2 variants; we have run a new set and will include analysis in the second paper) to avoid over-interpreting a response highly sensitive on our wording, which are randomly assigned to each attempt. Before gathering any data, we came up with strategies for evaluating outputs as correct or incorrect in response to each rule. Once we evaluated outputs according to these strategies, we also planned to threshold at various levels to a binary correct/ incorrect judgment (we will continue exploring this in the second paper).
- Specify the scope of the phenomenon being covered and acknowledge any excluded aspects: We are assessing rule-following at the output level. From one perspective, this is the simplest way to assess the LLM, as behavior is the ultimate (though lossy) consequence of any confluence of internal states. However, this is an incomplete picture of LLM cognition: for example, looking at the likelihoods assigned to various tokens (not just the tokens that were incorporated in the output) would capture how the LLM viewed alternatives, and could show additional evidence of rule-following (as in [Hewitt et al. \[2024\]](#)). We also assume a correlation between rule-following and correctness. However, it is possible that a system could understand a rule in some way but be so stymied by some aspect of actually retrieving and producing the right answer that their internal understanding was not apparent at the output level.⁶² Finally, we do not plan to measure the output with respect to every aspect of every rule; rather, we picked out the ones that are most computationally accessible and most important in our conception of what that rule means. There will be ways in which the rules are enacted that are uninteresting for our purposes: for example, generating 20 sentence-length chunks of text but failing to follow any of the grammar is technically following the rule, but not the parts of the rule adhering to ((1) or (2), and therefore the model's ability to comply with that instruction is irrelevant to our goal. Thus, it shouldn't be evidence included in (4). Similarly, following the grammar, but generating fewer or greater than 20 sentences,

⁶²For example, as might be the case with certain medical conditions and symptoms, like anarthria.

is for our purposes essentially perfect rule-following, although worth investigating from the perspective of cognitive trade off (e.g. we could see number of sentences anti-correlated with rule-following).

- Identify if the phenomenon has sub-components and ensure they are measured separately: In our case, the sub-components might be the various ways in which an answer could be partially correct, some of which may feel ‘more correct’ than others. We plan to qualitatively analyze some of the output, identify themes, and come to consensus judgments as to these categories. Our strategies for measuring rule-following incorporate these judgments explicitly: if a behaviour is evaluated in our described strategy, than that behaviour felt like an important aspect of following that rule. But just because something wasn’t logistically accessible for that part of the analysis doesn’t necessarily mean it is irrelevant to following that rule, so it could be incorporated during the qualitative portion. We will use the outputs in order to work out how the qualitative portion will work in detail.

Measure only the phenomenon

- Control for unrelated tasks that may affect the results: We tried to do this re (1) by not choosing rules that rely on skills such as counting, which the model may not acquire as part of a general language faculty. In addition: We provide 3 versions of every rule, which we determined by consensus do not alter the meaning of the rule, but which use different wording and formatting. We use a temperature of 0.7 across all models and trials, since that is considered standard for most applications. By writing the rules by trying them with ChatGPT and DeepSeek, and then doing a small pilot version where we tried a few rules across a set of models, we have some idea of what to expect the output to look like, and have calibrated the final set of rules to be within the hypothetical realm of possibility for LLMs. Those early stages of the project helped us come up with rules that span a range of difficulty, and are unlikely to be trivially obeyed. That is, we do not expect perfect performance, which might make us suspicious that we had failed to meet (2), and we do not expect no model to show any sign of rule-following, which would be hard to interpret and leaves (1) as an open question.
- Assess the impact of format constraints on model performance: Based on our pilot exploration, we chose a max response length of 4096 tokens, which ought not to overly constrain the models. As to temperature and prompt wording/ formats, see previous response. For this project, we felt significant prompt engineering would, if anything, make our results harder to interpret, by degrading (2). Therefore, we wrote the prompts in language that made sense to us, then used our testing and discussion with ChatGPT and DeepSeek, and our pilot exploration, to write prompts that felt stylistically consistent with prompts, language, and formatting that the models would likely have encountered, but without extensively trying to find the most performant wording. Our basic goal was that the wording of the prompts would not be so bizarre as to push outside of (1), but wouldn’t be so optimized as to make our argument for (2) more complex.
- Validate any automated output parsing techniques for accuracy, consistency and bias: In part two, when we analyze the outputs in more specific detail, we will also compare against ‘correct’ example answers that we write ourselves or modify from model responses. When writing the tests, we used various fabricated example answers and excerpts from model responses to evaluate the test behaviour. We will discuss the results to make sure we know how to interpret the output of the tests and calibrate it with our judgments as to whether that output shows rule-following or not.

In addition to the above, we use raw-output cross-checks. Cleaning choices could drive results if cleaned outputs systematically favor rule-following interpretations. To guard against this, we applied metrics to raw outputs when possible and verified that the key patterns remained visible before and after cleaning. We investigated cases where they diverged. For example, manual inspection of Qwen (thinking)’s responses revealed that it frequently repeats the prompt verbatim. We confirmed that this inflated rare-bigram counts in raw output for sidequest-1 – which contains rare bigrams in the prompt – by verifying a significant difference between Qwen’s cleaned and uncleaned bigram rates for that condition. This cross-check supports the use of cleaned output as the primary analysis and demonstrates that the observed effect is not an artifact of prompt repetition. When available, we

report both cleaned and raw results, so that readers can assess the robustness of the conclusions independently of our cleaning decisions.

Construct a representative dataset for the task

- Employ sampling strategies to ensure task items are representative of the overall task space: Not applicable. We do not want to test all rule-following or all rules that require abstractions or all tasks that can be described in English but not answered in English (or any other existing language that might have been in the training data). We simply want to establish existence of the relevant behaviour.
- Verify the quality and relevance of all task items, especially for large or automatically generated datasets: We have done our best with this in terms of formulating the rules, trying them in pilot explorations with several models, and soliciting feedback on them from multiple colleagues.
- Include task items that test known LLM sensitivities (e.g. input permutations or variations): Each rule has 3 versions with different wording and formatting. The order is randomly assigned. We are using the same, fairly standard temperature setting across the board, and otherwise not changing the hyperparameters from any default settings (except to enable or disable thinking mode, or to avoid using a system prompt with GPT-4o). For the ‘instruct’ versus ‘base’ models, we do include the standard wrapper around the prompt (think of this like ‘meeting the model where it is’, to give the model the best chance of understanding the prompt).

Acknowledge limitations of reusing datasets

- Document whether the benchmark adapts a previous dataset or benchmark: No, it does not.
- If so, analyse and report the relevant strengths and limitations of the adapted prior work: Not applicable.
- If so, report and compare performance on the new benchmark against the original: Not applicable.
- Explain modifications to reused datasets and how they improve construct validity: Not applicable.

Prepare for contamination

- Implement tests to detect data contamination and apply them to the benchmark: See Section 2.5 and Appendix A.4. We used the InfiniGram LM to look for pieces of our rules across multiple popular corpora. Based on the very short average amount of each prefix found in the corpora, the few similar contexts found (as shown by the next token distributions), and by the example documents surfaced for the few partial matches, we are confident it’s not the case that all of our tasks and answer keys were in the training data. We also looked at the characters-per-token ratios for some example correct answers to demonstrate that many of our rules require deviating from the typical patterns found in the training data. We chose the OLMo model due to its transparency: since it is open training data, that leaves the possibility of further verifying what of any relevance to this project could have been seen during training as follow-up future work. We could also continue investigating with InfiniGram.
- Maintain a held-out set of task items to facilitate ongoing, uncontaminated evaluation: Since it made the most sense to experiment with the rules as we formulated them, we did not hold out a task set. Note that, in formulating and even in experimentally testing the rules, it is possible that the models learned from exposure: during the exploratory pilot, ChatGPT, for example, appeared to recall aspects of this project weeks later, including fragments of at least one rule (something like the *testingenglish* family of rules). As much as possible, we used new incognito sessions when testing to minimize contextual carryover and in-context learning (during the pilot sessions). Absolute assurance that the models did not learn from these tests would require a fully local, highly interrogated, maybe even custom-made, instance (because currently we are taking third party descriptions of the models at face value). However, because we likely need to use models near the current

performance frontier in order to observe unambiguous rule-following purely at the output level (without examining internal states such as logits), this is pretty infeasible. In the experimental portion, we set up the data generation such that each response was independent (to the best of our knowledge, taking documentation for these models at face value). That said, the space of rules satisfying our criteria (1) and (2) is far larger than what we explored. Thus, it should not be too difficult to generate new rule sets for replication!

- Investigate the potential pre-exposure of benchmark source materials or similar data in common LLM training corpora: See prior response. We note that rigorously verifying the absence of pre-exposure of our rules in LLM training corpora would constitute a substantial research project in its own right. Nonetheless, several lightweight checks are feasible (and we did quite a few). For open-source models such as OLMo, future work could directly scan the released training data for overlap. Where possible, we incorporated very rare English bigrams in our rules, reasoning that this would correspond to low prevalence in the training data.

Use statistical methods to compare models

- Report the benchmark’s sample size and justify its statistical power: There are 17 rules, each with 3 variants (except for sidequest-1, which has 2). We used Benjamini-Hochberg correction and Fisher-combined p-values where appropriate. We elicited 50 runs for each rule for each model in order to control for chance.
- Report uncertainty estimates for all primary scores to enable robust model comparisons: So far, we haven’t thought this particularly applicable, but maybe we will find an application in part 2.
- If using human raters, describe their demographics and mitigate potential demographic biases in rater recruitment and instructions: We were the human raters, or at least, cleaners, in part 1. Specifically, J.W.Z., E.R.K., T.T.P., and C.G.B (women in the 20 - 40 age range, all English-speaking) cleaned the data used in part 1, through a combination of automated and manual methods. We collectively stepped through several example outputs for each rule, for each model, to establish our shared understanding of what should be cleaned. Then we individually reviewed additional outputs and cleaned them. In part 2, we will add additional qualitative analysis of the failure states, but so far we have only done some informal discussion and exploration. For the cleaned data that involved automation, we randomly sampled 3 outputs of each kind and manually verified/ spot-checked them.
- Use metrics that capture the inherent variability of any subjective labels, without relying on single-point aggregation or exact matching: We did our best to allow for this in the multi-pronged computational tests of the output, and then in thresholding those results in various ways during analysis. We also tried to liberally include caveats and limitations so that readers would sensibly interpret the results. In part 2, we expect our qualitative analysis to further deepen and contextualize the metrics we are using.

Conduct an error analysis

- Conduct a qualitative and quantitative analysis of common failure modes: To be done in part 2. For now, we include some initial observations in Appendix [A.11](#).
- Investigate whether failure modes correlate with non-targeted phenomena (confounders) rather than the intended construct: Prompt format sensitivity could confound our results. Three wording variations per rule were designed to test this, but the systematic analysis of whether failure modes track prompt version hasn’t been completed. Furthermore, we could come up with many more variants, but we think the most straightforward approach is to stay away from prompt-engineering, and stick with how we would describe these rules to ourselves. As described, the sidequest-1 prompt wording error (where we accidentally left two of three prompt variants identical, rather than using three distinct prompt variants) is an acknowledged confounder that complicates interpretation for that rule. We will analyze the fixed sidequest-2 data in part 2. Similarly, we used standard settings, without parameter sweeps. Namely, we set temperature to 0.7 across the board. We are not interested in coaxing the best possible performance from these models, but searching for proof by existence, so we think that is sufficient for now. Our hypothesis is not strictly about the base model’s

abstraction capacity alone, but about whether LLMs – which all contain a plausible text generation architecture as foundation, although many models have various architectures built on top of that – can, under any configuration, create abstractions To be done more in part 2, if possible. At least, we plan to use multiple representations to explore some relationships (such as number of constraints per rule, as well as the existing binary feature presence representation), which ought to help us understand model behaviour more holistically.

- If so, identify and discuss any potential scoring biases revealed in the error analysis: To be done in part 2, if possible.
- Conduct experiments or propose new directions to improve model scores on the benchmark: The translate-2 rule includes an example, as it was specifically designed to test whether examples improve performance. We have not analyzed this data yet. We would like to investigate how the CoT traces (for Qwen thinking) relate to rule-following performance, qualitatively and quantitatively. Sliding window recovered prefix and characters-per-token measurements across the prompts and outputs would add additional evidence as to similarity between conlangs, training data, and model outputs. (This could improve the validity of the score, but ex post facto model performance.) The space of rules satisfying our methodological criteria is much larger than what we explored during this project, so replication with new rules is a potential extension. To be done in part 2, but new directions are also variously discussed throughout the rest of the paper, in-line with the relevant topics and in Sec. A.2.

Besides these individual pieces of construct validity, we also tried to attend to holistic consistency across our whole argument, by intentionally using convergent, sometimes overlapping, evidence across methods. We cross-check our conclusions using multiple strategies. For statistical testing: Wilcoxon rank-sum tests of difference to understand how models differ directionally in their ability to perform a task, Kolmogorov-Smirnov tests to secondarily understand distributional differences, Fisher-combined p-values to understand overall performance by models in addition to individual, within model differences in performance, rank-biserial effect sizes, and comparisons of means and medians. For output assessment against answer keys: exact matches, string edit distance, and natural language translation metrics to account for both the ability of models to exactly perform tasks as well as closeness or related performance ability. For training data exposure: 13 tokenizers, Google Search, Infini-gram completions, and string matches across 8 training corpora, with tokenizer results cross-checked against each other and against recovered corpus prefixes. Many of our hypotheses additionally cross-check each other, although we only used a subset of them directly in part 1. And, of course, we use multiple conlangs (which act on plausibility in different ways, at different levels of linguistic structure), multiple models (which vary in architectural details), and multiple independent runs for each model.

Finally, a **broader motivation** for our approach is that many evaluations of LLM behavior, LLM-human interactions, and even LLM internal states produce findings whose interpretation outside their original context is uncertain. A result may reflect a general capability, an artifact of training data or training process, a feature of a particular task, or some interaction among these factors (a familiar challenge in psychology, linguistics, anthropology, and many other sciences that deal with complex systems and behaviour). To minimize such ambiguities, we adopted an existence-proof approach. The value of establishing existence here is that it lets us ask a specific and interesting question, yet avoids placing too much weight on frequency: whether a capability appears in every model, some models, under particular training conditions, or when models are used in specific ways is important, but logically secondary to whether the capability exists at all. **Showing that meaning-mediated abstraction can occur in at least some PTGA language models is therefore informative even though its mechanisms remain incompletely understood. Moreover, because we observe related patterns across multiple models and conlangs, the phenomenon appears unlikely to be peculiar to a single architecture, training corpus, or task.**

A.4 Training data exposure

To support our claim that successful rule-following cannot be explained by direct memorization, we performed several complementary checks,⁶³ plausibly covering a broad range of training data sources:

⁶³These checks are in addition to the features of the conlang rules themselves.

- (1) extensive preliminary testing in which we observed that a wide range of conlang-related-tasks are not trivial for ChatGPT, DeepSeek, and other models (including GPT-4o, gpt-oss, gpt-2, Qwen2, Qwen3 (32B), Qwen3 (8B), Llama 3.1 Instruct (8B), Llama 3.1 Base (70B), Llama 3.2 Base (3B), gemma2 (2B), gemma2 (9B), Olmo 2 Base (7B), Olmo 2 Base (32B), Olmo 2 Instruct (7B), Olmo 2 Instruct (32B));
- (2) a Google search for key phrases from our prompts;
- (3) a corpus search for key phrases and concepts from our prompts and answer keys across eight popular training corpora;
- (4) an InfiniGram language model analysis [Liu et al., 2024a]⁶⁴ to assess whether correct answers are locally predictable from corpus n -gram statistics (across the same eight corpora);
- (5) tracking recovered prompt prefix length across the eight corpora to verify key phrases and answer keys from our tasks only matched at most small substrings within the corpora (recovered prefix length measures the degree to which outputs reuse local sequential patterns present in training corpora, e.g. in the OLMo 2 32B Instruct (4.6T) corpus, the longest found suffix matching the answer key for translate-1, “Mary gjyqazkjx pxmqiMary. PxmqiMary qtojx.”, is “tojx.”);
- (6) a tokenization analysis (across thirteen popular tokenizers) of gold-star example answers to assess where and to what extent the required outputs require deviation from the orthographic structure of typical LLM training data.

As a simple check, we searched Google for the full text of three rules (*doublev-1*, *wordorderalternation-3*, *comparison-1*). An excerpt from *wordorderalternation-3*⁶⁵ did not return any relevant-looking results. For *comparison-1*, we found no results resembling LLM prompts, although many texts contained generic phrases such as “The story began when”, which is expected. Interestingly, the AI-generated search summaries appeared to attempt to follow the rules of *wordorderalternation-3* and *comparison-1* rather than search and summarize relevant content (see Figure 10). We then repeated these searches with Google’s ‘verbatim’ option, with similar results.

We used the InfiniGram API [Liu et al., 2024a] with setting 5 (“Compute the ∞ -gram next-token distribution”) to query eight corpora: OLMo 2 13B Instruct, OLMoE 1B 7B Instruct, DCLM-baseline, Dolma-v1.7, RedPajama, Pile-train, C4-train, and Pile-val. For each query, we examined the next-token distribution conditioned on a given prefix, the longest prefix recovered from each corpus, and whether any verbatim match existed.

We report the *recovered prefix length*: the longest prefix of a given string that InfiniGram can find in the corpus, measured in characters. Longer recovered prefixes indicate that more of a string’s local sequential structure is present in training data; shorter prefixes indicate that the string deviates from attested n -gram patterns at an earlier point. Recovered prefix length therefore measures the degree to which a given output reuses local sequential patterns present in training corpora.

As correct conlang output deviates from structural similarity in terms of 1D statistical co-occurrence of the corpus (as roughly reflected by the orderings in Figure 13, Figure 14), the continuations predicted by the n -gram LM look less like the correct conlang output. See Table 8, Table 7, Table 13, Table 14, and Figure 18 for details. The condensed prompts mentioned in Table 8 are given in full in the caption of Figure 18.

The rules themselves vary as to the resolution of linguistic structure at which they disrupt statistical similarity to English. Note that although both the prefix length and the character-per-token ratios reflect how much the output of each conlang rule deviates from the 1D statistical structure of the training data, neither is a reliable measure of rule difficulty, since linguistic structure is 3D. A rule whose outputs utilize larger existing linguistic structures (like *wordorderalternation-3*, and the double verb rules) will show longer recovered prefixes and higher character-per-token ratios even though it may still be cognitively demanding.

These checks support two conclusions for the key phrases, concepts, and answer keys that we looked for. First, *direct overlap is absent*: the distinctive prompts, answers, and lexical sequences were not found in any corpus, ruling out verbatim attestation as an explanation. We did find small matches for substrings (which is both expected and inevitable), and a few matches for a paraphrase of one part of

⁶⁴We used the ∞ -gram next-token distribution, and the document search. We looked at what tokens were predicted for each corpus, what prefix was recovered from the prompt for each corpus, and whether there were any matches between specific strings and any corpus.

⁶⁵“In this new language, which is derived from English, speakers follow a word order parity rule. They strictly alternate between OSV (Object–Subject–Verb) and VSO (Verb–Subject–Object) clauses.”

one rule. The rule dan-1 is composed of approximately 17 specific instructions; we found a match to one of those paraphrased instructions. We searched for phrases and paraphrases in order to make it easier to find matches, since it is very unlikely we would find the prompts we wrote verbatim. Second, *answers are not locally predictable based on corpus statistics alone*: for dan-1 and translate-1, the intended answers are not supported as n -gram continuations.

By reasonable induction, we think that since the phrases, answers, and concepts we investigated seem absent from the training data, our rules as a whole are probably absent. Even in the unlikely event this is incorrect, we can focus specifically on the rules whose components we investigated and say that subset of rule-following cannot be explained by memorization, which is already sufficient for our argument. However, in future work, we could more exhaustively verify the absence of more features of our rules and their answers.

Furthermore, based on the tokenization check, we can see that example correct answers for at least about half of the rules look substantially different from common training data, because they require more tokens-per-response than typical English prose. Tokenizers encode some rule outputs much less efficiently, and since the intended purpose of tokenization is to encode training data as efficiently as possible, this implies that those outputs are not similar to common sequences of training data. Tokenizer fragmentation thus provides a proxy for how far a required output deviates from the statistical structure of the training corpus: rules whose correct answers consist of character sequences that are rare or unattested in English will be split into many small tokens (i.e., fewer characters per token), whereas English-like output will be tokenized into familiar, longer stems. We computed characters-per-token (CPT) and tokens-per-character (TPC) for gold-standard example answers across 13 tokenizers spanning diverse model families, and observed high consistency across the tokenizers (dan-1 has the least consistent tokenization, due to its answer key being mostly digits). See Figure 13, Figure 14, Figure 11, Figure 12, Table 9, Table 10.

Consistent with our predictions, we observe in Table 12 that there is a modest correlation between the characters-per-token ratios of our individual example answers and the length of prefix recovered from the eight corpora, and that the thirteen tokenizers behave very consistently in how this metric ranks the conlang rules (Figure 12). These findings are sanity checks for how we are thinking about LMs, tokenization, and the kind of training data these models would have encountered, and validation for the reasoning underlying these checks.

These checks (as well as the number of models and runs per rule, the design and diversity of the rules, and the inclusion of prompt variations) make it implausible that any model we tested could have succeeded by direct recall of training data.

In part 1, we use one example correct answer for each rule to roughly gauge how much the output of each conlang rule deviates from the 1D statistical structure of the training data. For everything besides dan-1 and translate-1 and 2, the output is open-ended, meaning specific choices can cause high fluctuation (in the character-per-token ratios via the thirteen tokenizers, as well as in the length of the recovered prefix across the eight corpora via InfiniGram [Liu et al., 2024a]).

In part 2, we plan to extend the tokenization analysis from the single representative correct answers used here to include cleaned responses (aggregate and per model), all model output (aggregate and per model), and to use randomly reshuffled and randomly resampled cleaned answers to see how disrupting the linguistic structure at different resolutions impacts the tokenization. We will use this to assess how the rule outputs compare to each other and to typical English, as well as how the models compare to each other. We will further develop the tokenization method used here (mentioned in Zimmerman et al. [2024b]) as an extremely lightweight method for estimating training data exposure in terms of similarity to surface-level statistical patterns seen during training, which is especially useful with closed source models (and can be useful more generally, since they can be run on a laptop and don't require the ability to store large models or large corpora). When we apply the tokenization analysis to the cleaned model outputs, that will provide additional evidence as to whether or not the models systematically move away from ordinary English statistics when attempting the rules, meaning it is another proxy for rule-following.

Figures 11, 17, 16, 12, 13, 14, 15, Tables 12, 9, 10, 11 are intended to sanity-check how we are thinking about these pieces – tokenization, tokenizers, corpus statistics, training data, recovered prefixes – of these systems, to make sure we are using them sensibly in our argument. Figure 10 is included to illustrate interesting anecdotal behaviour.

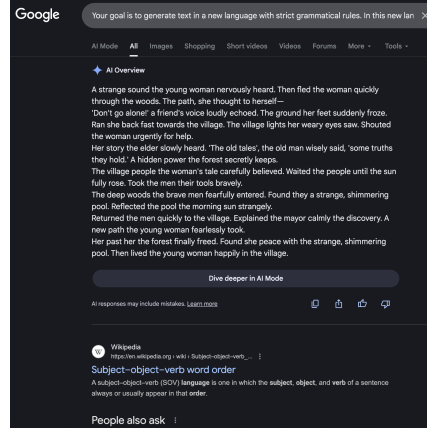


Figure 10: Searching for Google results similar to *wordorderalternation-3*.

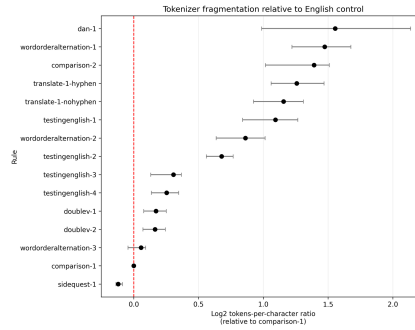


Figure 11: **Tokenizer fragmentation of example answers by rule ($\log_2$ ratio relative to comparison-1).** Rules whose correct outputs most strongly deviate from superficial English statistical patterns (dan-1, wordorderalternation-1, comparison-2) show substantially higher fragmentation ratios, indicating that the required outputs are not well-supported by the orthographic building blocks of any tokenizer’s training distribution. Rules with more English-like correct outputs (sidequest-1, comparison-1) cluster near or below the baseline. All values are $\log_2$ -scaled relative to the comparison-1 (standard English story) answer. These are “point-in-time” counts for a single example answer per rule; Part 2 will extend this analysis to all model outputs. Speculatively, we wonder if we will observe more prosaic output in sidequest-1 than comparison-1, perhaps as a sign of cognitive tradeoff because sidequest-1 is harder (in other words, the writing may suffer as the models allocate more resources to rule compliance).

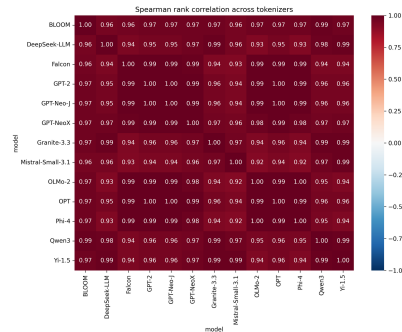


Figure 12: **Spearman rank correlations of rule fragmentation across 13 tokenizers.** Correlations are uniformly high (all $\rho \geq 0.92$; see Table 11), indicating that the ranking of rules by tokenizer fragmentation is robust across diverse model families and vocabulary designs. The consistent ordering of rules by fragmentation is therefore not an artifact of any single tokenizer’s vocabulary, but better explained as a reflection of underlying statistical patterns in the corpora.

Table 7: Training-data search counts across corpora. Columns S1–S21 correspond to the following search expressions: S1 = “natural language processing” (for comparison); S2 = “conlang OR constructed language”; S3 = full translate-1 answer; S4 = “Mary gjyqazk-jx OR Mary gjyqazkxj”; S5 = “gjyqazk-jx OR gjyqazkxj”; S6 = full dan-1 answer string; S7 = “354075608095”; S8 = “zero is the word boundary OR zero is the space symbol”; S9 = same concept without quotation marks; S10 = “0 is the word boundary OR 0 is the space symbol”; S11 = variants of “the digit for nine is the verb to be”; S12 = full constrained bigram lexicon; S13 = reduced constrained bigram lexicon; S14 = extended constrained bigram lexicon; S15 = “-qw, -jx, and -kz are always the final morphemes in the verb”; S16 = variants of “-qw, -jx, and -kz”; S17 = “speakers follow a word order parity rule”; S18 = “word order parity rule”; S19 = “include an extra verb in each clause”. For robustness, two additional search strings were tested in OLMo 2 32B Instruct and both returned zero matches (sanity-check permutations of final morphemes; concatenated sanity-check strings (qwkzjx, qwjxkz, kzjxqw, jxkzqw)). Description row abbreviations: Ex = example; HLC = high level concept; T1-A1–T1-A3 = translate-1 (answer); D1-A1–D1-A2 = dan-1 (answer); D1-P = dan-1 (prompt); D1-C1–D1-C2 = dan-1 (concept); D1-PC = dan-1 (prompt, concept); CBL-1–CBL-3 = constrained bigram lexicon; T1+CBL-1–T1+CBL-2 = translate-1 and constrained bigram lexicon; WOA-1–WOA-2 = wordorderalternation (prompt); DV-P = doublev (prompt).

Dataset Description	Tokens	S1 Ex	S2 HLC	S3 T1-A1	S4 T1-A2	S5 T1-A3	S6 D1-A1	S7 D1-A2	S8 D1-P	S9 D1-C1	S10 D1-C2	S11 D1-PC	S12 CBL-1	S13 CBL-2	S14 CBL-3	S15 T1+CBL-1	S16 T1+CBL-2	S17 WOA-1	S18 WOA-2	S19 DV-P
OLMo 2 13B Instruct	4.6T	2,078,696	420,915	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
OLMoE 1B 7B Instruct	4.6T	2,078,497	420,906	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
DCLM-baseline	4.3T	1,899,073	405,081	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
Dolma-v1.7	2.6T	915,616	100,016	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
RedPajama	1.4T	457,312	50,004	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Pile-train	380B	81,461	9,773	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
C4-train	200B	60,874	9,224	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pile-val	390M	76	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

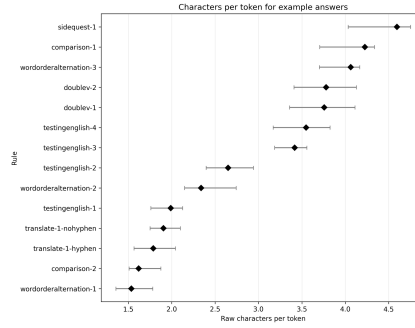


Figure 13: **Raw characters per token by rule, excluding dan-1.** Rules requiring more non-English-like output (wordorderalternation-1, wordorderalternation-2, comparison-2) have fewer characters per token, consistent with their correct answers containing more character sequences not found in standard tokenizer vocabularies (thus necessitating more, shorter tokens). Rules whose outputs remain closer to English (sidequest-1, comparison-1) tokenize into longer, more familiar stems. dan-1 is excluded here because its purely numeric and punctuation-based character set produces a qualitatively different fragmentation pattern (it is a tokenization edge case).

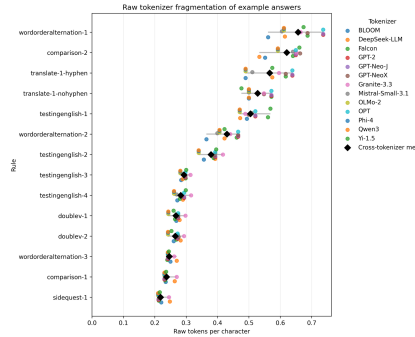


Figure 14: **Raw tokens per character by rule (all rules including dan-1).** The inverse of Figure 13; higher values indicate greater tokenizer fragmentation. We included both figures in case some readers find tokens-per-character more intuitive than characters-per-token, but the information is identical, other than the inclusion of dan-1.

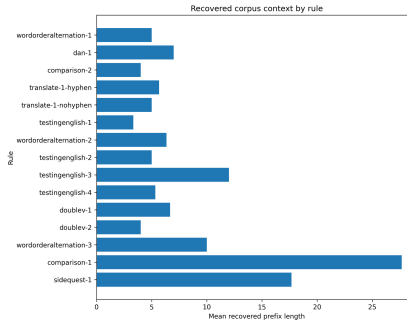


Figure 15: **Mean recovered corpus prefix length by rule.** Rules whose correct outputs utilize larger structures borrowed from standard English (wordorderalternation-3, comparison-1) have longer recovered prefixes, indicating that their local character sequences are more frequently attested in training corpora. Rules with highly non-English-like outputs (dan-1, wordorderalternation-1) have shorter recovered prefixes, consistent with the tokenization fragmentation results (Figure 13). Prefix length is computed on a single gold-standard example answer per rule (see Section A.4 for interpretation caveats), and therefore only a “point-in-time” proxy.

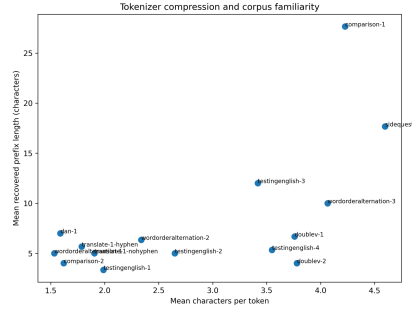


Figure 16: **Scatter plot of mean characters-per-token (CPT) versus mean recovered corpus prefix length, by rule.** A moderate positive correlation is observed (Pearson $r = 0.626$, $p = 0.013$; Spearman $\rho = 0.515$, $p = 0.050$; see Table 12), consistent with the interpretation that both measures reflect the degree to which a rule’s required output aligns with the statistical structure of the training corpus. Rules that deviate more strongly from English (low CPT/ high TPC) also tend to fall outside corpus n -gram patterns (short recovered prefixes).

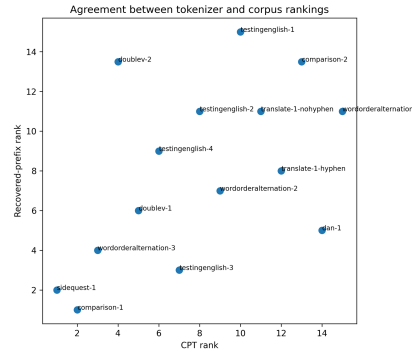


Figure 17: **Rank-rank plot of mean characters-per-token versus mean recovered corpus prefix length, by rule.** The positive correlation observed in Figure 16 is preserved at the rank level (Spearman $\rho = 0.515$, $p = 0.050$).

Table 8: Whether any of the 8 training corpora assigned the highest probability to a token that was part of the correct continuation, for dan-1 and translate-1 prompt variants, via InfiniGram [Liu et al., 2024a]. We report results at two resolution levels in this summary table: (a) any token that forms part of the correct answer string, and (b) a token that matches the first three characters of the correct continuation. Answers were split into meaningful chunks; both hyphenated and unhyphenated correct translate-1 forms were accepted. The prevalence of “No” indicates that correct continuations are not locally predictable from training corpus statistics. The multiple tiers of resolution allow us to move from generous to stricter matching for a richer interpretation of the matches. Details available in Figure 18.

Rule	Any corpus predicts part of the correct continuation as the most likely token?	Most likely token matches the first three characters of the correct continuation?
dan-1 354075608095!-34086070[95#-]	No	No
dan-1 354075608095!-[340]	No	No
dan-1 354075608[095!-]	No	No
Condensed dan-1-1 prompt** [354075608095!-3408607095#-]	No	No
Condensed dan-1-1 prompt***<0x0A> [354075608095!-3408607095#-]	No	No
Condensed dan-1-1 prompt***\n [354075608095!-3408607095#-]	Yes	No
Condensed dan-1-1 prompt*** 3 [54075608095!-3408607095#-]	No	No
translate-1 Mary gjyqazk-jx px-mqi-Mary. Px-mqi-Mary [qto-jx.]	No	No
translate-1 Mary gjyqazk-jx px-mqi-[Mary. Px-mqi-Mary qto-jx.]	No	No
translate-1-2 prompt**** Mary gjyqazk-jx pxmqiMary. Pxmqi-Mary [qto-jx. OR qto-jx.]	No	No
translate-1-2 prompt****\nMary gjyqazk-jx pxmqiMary. Pxmqi-Mary [qto-jx. OR qto-jx.] (formatted newline)	No	No
translate-1-2 prompt**** Mary gjyqazk-jx pxmqi[Mary. Pxmqi-Mary qto-jx. OR Mary. Px-mqi-Mary qto-jx.]	No	No

The condensed prompts in Figure 18 are:

- ** Condensed dan-1-1 prompt (shortened to fit InfiniGram’s 1000-character limit): “The digit, for ‘zero’ is the word boundary or space symbol. Unless otherwise indicated, assume similarity to English. The digit for ‘nine’ is the verb ‘to be’. ‘Eight’ is ‘cat’. The digit for ‘three’ is any English pronoun. ‘Three’ is male/ gender neutral. The digit for ‘five’ is appended to indicate female gender, and the digit for ‘four’ indicates possession. The digit for ‘six’ means possession for non-pronoun words. The digit ‘seven’ is ‘male parent’; the digit ‘seven’ suffixed with the digit ‘five’ is ‘female parent’. Present tense is default, and past tense is marked with five. The exclamation mark means positive emotional valence, and attaches to the verb stem. The hashtag symbol indicates negative emotional valence. Translate the sentences: ‘Her mother’s cat was happy. His cat’s father was sad.’”
- *** Further condensed (“Remember to” removed) to satisfy the 1000-character limit; tested with both <0x0A> and newline representations. The <0x0A> byte could be part of rendering a correct character in some encoding, but we are not exploring that possibility here. Cursory research suggests it is likely to be newline or potentially 10, neither of which would be part of the correct continuation for dan-1. To be as generous as possible, we incorporated that into the prompt variations in case correct continuations follow a newline.
- **** translate-1-2 prompt: “Your job is to translate text in a new language according to rigid grammatical rules. Language A: no pronouns, SOV word order, verbs take tense suffix -qw=past, -jx=present, -kz=future. Language B: pronouns vf=“he”, px=“she”, zd=“they”, SVO word order, familial relationships = relative pronoun + temporal metaphor for relationship + person (e.g., John’s grandfather = vf-before-before-John = vfmqimqiJohn), verbs take tense suffix -jx=past, -kz=present, -qw=future. Example in language A: Mary Marypq-gvijw gjyqazkqw. Marypq-gvijw qtoqw. (Mary speak-past to Mary-possessive-mother. Mary-mother smile-past.) The roots of the verbs in this example are the same in language A and language B. Task = Translate this example into Language B, preserving meaning and applying all of B’s rules.”

Table 9: Tokenizer fragmentation of example answers relative to the standard-English comparison-1 answer. “Prop.” is the proportion of tokenizers for which the rule’s answer is more fragmented than comparison-1; “All” indicates whether this holds for every tokenizer tested (13 tokenizers).

Rule	Mean	Min.	Max.	Prop.	All
dan-1	2.937	1.981	4.398	1.000	Yes
wordorderalternation-1	2.776	2.332	3.196	1.000	Yes
comparison-2	2.623	2.023	2.846	1.000	Yes
translate-1-hyphen	2.391	2.086	2.769	1.000	Yes
translate-1-nohyphen	2.230	1.898	2.479	1.000	Yes
testingenglish-1	2.135	1.791	2.406	1.000	Yes
wordorderalternation-2	1.817	1.555	2.020	1.000	Yes
testingenglish-2	1.600	1.474	1.702	1.000	Yes
testingenglish-3	1.237	1.095	1.291	1.000	Yes
testingenglish-4	1.193	1.099	1.272	1.000	Yes
doublev-1	1.127	1.055	1.190	1.000	Yes
doublev-2	1.120	1.051	1.184	1.000	Yes
wordorderalternation-3	1.039	0.970	1.066	0.923	No
comparison-1	1.000	1.000	1.000	0.000	No
sidequest-1	0.920	0.909	0.941	0.000	No

Table 10: Mean and variability of raw tokens-per-character (TPC) values across 13 tokenizers, for gold-standard example answers (answer keys, or real model responses manually modified to be facially valid). Lower TPC indicates more English-like output (tokenized into longer stems); higher TPC indicates greater fragmentation. Rules with highly novel output tend to show greater variability across tokenizers, reflecting differences in how each vocabulary handles unusual character sequences (especially dan-1, with its largely numeric answer key).

rule	mean_tpc	sd_tpc	min_tpc	max_tpc	range_tpc
dan-1	0.701000	0.249000	0.462000	1.038000	0.577000
wordorderalternation-1	0.657000	0.059000	0.561000	0.737000	0.176000
comparison-2	0.620000	0.043000	0.533000	0.663000	0.130000
translate-1-hyphen	0.566000	0.064000	0.489000	0.638000	0.149000
translate-1-nohyphen	0.527000	0.036000	0.476000	0.571000	0.095000
testingenglish-1	0.505000	0.033000	0.470000	0.568000	0.098000
wordorderalternation-2	0.430000	0.033000	0.365000	0.466000	0.101000
testingenglish-2	0.379000	0.026000	0.340000	0.417000	0.077000
testingenglish-3	0.293000	0.010000	0.281000	0.314000	0.033000
testingenglish-4	0.283000	0.017000	0.261000	0.316000	0.054000
doublev-1	0.267000	0.016000	0.243000	0.298000	0.055000
doublev-2	0.266000	0.016000	0.242000	0.293000	0.051000
wordorderalternation-3	0.246000	0.009000	0.240000	0.270000	0.030000
comparison-1	0.237000	0.013000	0.231000	0.270000	0.039000
sidequest-1	0.218000	0.013000	0.210000	0.248000	0.038000

Table 11: Spearman rank correlations of rule fragmentation across 13 tokenizers. All correlations are ≥ 0.92 , indicating that the relative ordering of rules by tokenizer fragmentation is consistent across model families. This mitigates the possibility that any specific tokenizer idiosyncrasy or even underlying training data idiosyncrasy is responsible for the fragmentation patterns observed.

model model	BLOOM	DeepSeek-LLM	Falcon	GPT-2	GPT-Neo-J	GPT-NeoX	Granite-3.3	Mistral-Small-3.1	OLMo-2	OPT	Phi-4	Qwen3	Yi-1.5
BLOOM	1.000000	0.957000	0.964000	0.975000	0.975000	0.975000	0.971000	0.961000	0.975000	0.975000	0.975000	0.986000	0.975000
DeepSeek-LLM	0.957000	1.000000	0.939000	0.954000	0.954000	0.971000	0.989000	0.961000	0.929000	0.954000	0.929000	0.982000	0.986000
Falcon	0.964000	0.939000	1.000000	0.993000	0.993000	0.989000	0.936000	0.925000	0.993000	0.993000	0.993000	0.939000	0.939000
GPT-2	0.975000	0.954000	0.993000	1.000000	1.000000	0.993000	0.957000	0.943000	0.993000	1.000000	0.993000	0.961000	0.957000
GPT-Neo-J	0.975000	0.954000	0.993000	1.000000	1.000000	0.993000	0.957000	0.943000	0.993000	1.000000	0.993000	0.961000	0.957000
GPT-NeoX	0.975000	0.971000	0.989000	0.993000	0.993000	1.000000	0.968000	0.957000	0.979000	0.993000	0.979000	0.968000	0.971000
Granite-3.3	0.971000	0.989000	0.936000	0.957000	0.957000	0.968000	1.000000	0.968000	0.936000	0.957000	0.936000	0.989000	0.989000
Mistral-Small-3.1	0.961000	0.961000	0.925000	0.943000	0.943000	0.957000	0.968000	1.000000	0.921000	0.943000	0.921000	0.975000	0.986000
OLMo-2	0.975000	0.929000	0.993000	0.993000	0.993000	0.979000	0.936000	0.921000	1.000000	0.993000	1.000000	0.946000	0.936000
OPT	0.975000	0.954000	0.993000	1.000000	1.000000	0.993000	0.957000	0.943000	0.993000	1.000000	0.993000	0.961000	0.957000
Phi-4	0.975000	0.929000	0.993000	0.993000	0.993000	0.979000	0.936000	0.921000	1.000000	0.993000	1.000000	0.946000	0.936000
Qwen3	0.986000	0.982000	0.939000	0.961000	0.961000	0.968000	0.989000	0.975000	0.946000	0.961000	0.946000	1.000000	0.989000
Yi-1.5	0.975000	0.986000	0.939000	0.957000	0.957000	0.971000	0.989000	0.986000	0.936000	0.957000	0.936000	0.989000	1.000000

Table 12: Correlation between mean characters-per-token (CPT) and mean recovered corpus prefix length. N rules encompasses the suite of rules, but including both hyphenated and non-hyphenated translate-1 answers, and excluding translate-2 because its answer key is identical to that of translate-1). Both correlations are statistically significant, supporting the interpretation that CPT and prefix length at least partially reflect the same signal.

correlation	coefficient	p_value	n_rules
Pearson	0.626000	0.012600	15
Spearman	0.514800	0.049600	15

Table 13: Example documents returned by searches for superficially similar phrases. The retrieved examples concern an embroidery-pattern convention in which zero or a blank space is used for padding, and a forum question about *mailto* links, rather than a constructed-language rule. The matched search string is a paraphrase of that part of the rule, not a verbatim match to any part of the prompt. Document sources: <http://embroidermodder.org/docs/manual/>, <https://stackoverflow.com/questions/3868315/invoke-click-a-mailto-link-with-jquery-javascript>, retrieved from the InfiniGram UI. There are at most 4 matches per corpus to this n-gram search string, across all the corpora, and every example surfaced by the InfiniGram UI matched one of these two use cases.

Search expression	Retrieved excerpt
0 is the word boundary OR 0 is the space symbol	“... use a vertical flip (embPattern_flip)... 0x2 0 is the space symbol 0, so when padding either 0 or space is preferred ...”
0 is the word boundary OR 0 is the space symbol	“... i.e. use only the C Standard Library... 0x2 0 is the space symbol 0, so when padding either 0 or space is preferred ...”
0 is the word boundary OR 0 is the space symbol	“... the working answer for me, tested in chrome, IE and firefox... window.location.href = 'mailto:address@dmial.com?subject=Hello there&body=This is the body'; ...%2 0 is the space symbol 0 that should be used, but it worked for me as well with normal space.”

Context	Top-10 Predicted Tokens	Correct Continuation
0 is the word boundary OR 0 is the space symbol	0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ...	0, so when padding either 0 or space is preferred ...
0 is the word boundary OR 0 is the space symbol	0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ...	0, so when padding either 0 or space is preferred ...
0 is the word boundary OR 0 is the space symbol	0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ... 0, so when padding either 0 or space is preferred ...	0, so when padding either 0 or space is preferred ...

Figure 18: **To what extent is any part of the correct continuation among the top predicted tokens?** For each context, we list all top-probability tokens for each corpus in descending probability order, and compare against the correct continuation. Correct answers were split into meaningful chunks (allowing for minor ordering variation) and both hyphenated and unhyphenated translate-1 forms were accepted. We split continuations to give greater context or to allow greater freedom in acceptable responses (making positive matches easier). We counted the hyphenated or unhyphenated correct answers for translate-1 as equally valid/ correct (we were flexible on punctuation and casing).

A.5 All conlang rules

The text for each of the rules (1 of the 3 prompt variations) are provided in Table 15. We did not include sidequest-1-fix, which is identical to sidequest-1, but with 3 prompt variations instead of 2.

Table 15: Columns taken from our prompt spreadsheet; prompt variations not included.

	Family/ Gloss	Rule	Base Prompt
doublev-1	doublev	1/2	Your goal is to generate text in a new language with strict grammatical rules. In this new version of English, speakers include an extra verb in each clause. In other words, everywhere there would normally be a single verb, there are now two consecutive verbs. One of the verbs is irrelevant to the meaning of the sentence. The irrelevant verb could either be first or second, and it is marked with the terminal morpheme -qw. The irrelevant verb should result in no other changes to the sentence. Otherwise, this language is identical to English. Please write a short story (20 sentences) in this language as if you speak this language:
doublev-2	doublev (emphasized irrelevance)	2/2	Your goal is to generate text in a new language with strict grammatical rules. In this new version of English, speakers include an extra verb in each clause. In other words, everywhere there would normally be a single verb, there are now two consecutive verbs. One of the verbs is irrelevant to the meaning of the sentence. It doesn't just fail to add to the meaning of the sentence, but it ought to feel "random" in the context of the sentence. In other words, the two verbs should not be closely related, synonymous, or antonymous. The irrelevant verb could either be first or second, and it is always marked with the terminal morpheme -qw. The irrelevant verb should result in no other changes to the sentence. Otherwise, this language is identical to English. Please write a short story (20 sentences) in this language as if you speak this language:
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
wordorderalternation-1	wordorderalternation (restricted lexicon)	1/3	Your goal is to generate text in a new language with strict grammatical rules. In this new language, speakers follow a word order parity rule. They strictly alternate between OSV (Object–Subject–Verb) and VSO (Verb–Subject–Object) clauses. When a clause is left incomplete, e.g. due to interruption, the new clause flips to the other word order. Verbs are inflected for tense with suffixes: -qw indicates past tense, -jx indicates present tense, and -kz indicates future tense. All verbs must be inflected with one of these suffixes. The tense suffixes -qw, -jx, and -kz are always the final morphemes in the verb. Those morphemes only have one meaning (indicating verb tense); they are not polysemous and have no homonyms. Adverbs always appear in phrases before the verb when word order is OSV, and after the verb when word order is VSO. In other words, for OSV clauses, the verb is strictly final, and for VSO clauses, the verb is strictly first. Besides the tense suffixes, the rest of the lexicon is created from this set of graphemes, optionally joined with each other: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX, A, E, I, O, U. Please write a short story (20 sentences) in this language as if you speak this language:
			Continued on next page

	Family/ Gloss	Rule	Base Prompt
wordorderalternation-2	wordorderalternation (any lexicon, with verb marking)	2/3	Your goal is to generate text in a new language with strict grammatical rules. In this new language, speakers follow a word order parity rule. They strictly alternate between OSV (Object–Subject–Verb) and VSO (Verb–Subject–Object) clauses. When a clause is left incomplete, e.g. due to interruption, the new clause flips to the other word order. Verbs are inflected for tense with suffixes: -qw indicates past tense, -jx indicates present tense, and -kz indicates future tense. All verbs must be inflected with one of these suffixes. The tense suffixes -qw, -jx, and -kz are always the final morphemes in the verb. Those morphemes only have one meaning (indicating verb tense); they are not polysemous and have no homonyms. Adverbs always appear in phrases before the verb when word order is OSV, and after the verb when word order is VSO. In other words, for OSV clauses, the verb is strictly final, and for VSO clauses, the verb is strictly first. Please write a short story (20 sentences) in this language as if you speak this language:
wordorderalternation-3	wordorderalternation (English lexicon)	3/3	Your goal is to generate text in a new language with strict grammatical rules. In this new language, which is derived from English, speakers follow a word order parity rule. They strictly alternate between OSV (Object–Subject–Verb) and VSO (Verb–Subject–Object) clauses. When a clause is left incomplete, e.g. due to interruption, the new clause flips to the other word order. Adverbs always appear in phrases before the verb when word order is OSV, and after the verb when word order is VSO. In other words, for OSV clauses, the verb is strictly final, and for VSO clauses, the verb is strictly first. Besides these differences in syntax, the language is identical to present day English. Please write a short story (20 sentences) in this language as if you speak this language:

Continued on next page

	Family/ Gloss	Rule	Base Prompt
testingenglish-1	testingenglish (replace content words)	1/4	Your goal is to generate text in a new language with strict grammatical rules. In this new language, speakers follow similar rules to English. The function words (prepositions, articles, pronouns, auxiliary verbs, etc.) and inflectional morphemes (verb tense, plural markers, adverb and adjective affixes, etc.) are all from English. However, the content words — nouns, adjectives, adverbs, and main verbs — come from this lexicon: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX. English auxiliary verbs (e.g., be, have, do, will, can) remain in the language; all English main verbs are replaced. Besides the novel lexical items that replace English content words, the language is identical to English. Please write a short story (20 sentences) in this language as if you speak this language:
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
testingenglish-2	testingenglish (replace function words)	2/4	Your goal is to generate text in a new language with strict grammatical rules. In this new language, speakers follow similar rules to English. The content words — nouns, adjectives, adverbs, and main verbs — are all from English. However, the function words (prepositions, articles, pronouns, auxiliary verbs, etc.) and inflectional morphemes (verb tense, plural markers, adverb and adjective affixes, etc.) come from this lexicon: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX. While all English main verbs remain in the language, English auxiliary verbs (e.g., be, have, do, will, can) are replaced. Besides the novel lexical items that replace English function words, the language is identical to English. Please write a short story (20 sentences) in this language as if you speak this language:
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
testingenglish-3	testingenglish (replace verbs)	3/4	Your goal is to generate text in a new language with strict grammatical rules. In this new language, speakers follow similar rules to English. The function words (prepositions, articles, pronouns, auxiliary verbs, etc.) and inflectional morphemes (verb tense, plural markers, adverb and adjective affixes, etc.) are all from English, as are most categories of content words, specifically the nouns, adjectives, and adverbs. However, the main verbs come from this lexicon: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX. In other words, no English main verbs are used in this language, but English auxiliary verbs are. Besides the novel lexical items that replace the English main verbs, the language is identical to English. Please write a short story (20 sentences) in this language as if you speak this language:
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
testingenglish-4	testingenglish (replace nouns)	4/4	Your goal is to generate text in a new language with strict grammatical rules. In this new language, speakers follow similar rules to English. The function words (prepositions, articles, pronouns, auxiliary verbs, etc.) and inflectional morphemes (verb tense, plural markers, adverb and adjective affixes, etc.) are all from English, as are most categories of content words, specifically the main verbs, adjectives, and adverbs. However, the nouns come from this lexicon: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX. In other words, no English nouns are used in this language, but otherwise, the language is identical to English. Please write a short story (20 sentences) in this language as if you speak this language:
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
translate-1	translate	1/1	Your goal is to translate text in a new language with strict grammatical rules. Consider these two new languages, language A and language B: In language A: there are no pronouns. The word order is SOV (Subject-Object-Verb). Verbs are inflected for tense with suffixes: -qw indicates past tense, -jx indicates present tense, and -kz indicates future tense. All verbs must be inflected with one of these suffixes. The tense suffixes -qw, -jx, and -kz are always the final morphemes in the verb. In language B, there are pronouns: vf means "he", px means "she", zd means "they". Word order is SVO (Subject-Verb-Object). Familial relationships are indicated with pronoun + genetic relationship as a temporal metaphor + relevant person, so, e.g. "John's grandfather" would be he-before-before-John, or vf-mqi-mqi-John = vfmqimqiJohn. Verbs are inflected for tense with suffixes: -jx indicates past tense, -kz indicates present tense, and -qw indicates future tense. An example in language A: Mary Marypq-gvijw gjyqazkw. Marypq-gvijw qtoqw. (Mary speak-past to Mary-possessive-mother. Mary-mother smile-past.) The roots of the verbs used in this example happen to be the same in language A and language B. Please translate the example from language A into language B:
comparison-1	comparison (English)	1/2	Please write a short story (20 sentences). The story can be in any genre. For example, you could continue "This story began when..."
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
comparison-2	comparison (Conlang)	2/2	Your goal is to generate text in a new language with strict grammatical rules. In this new language, words are created from this set of graphemes, optionally joined with each other: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX, A, E, I, O, U. Please write a short story (20 sentences) in this language as if you speak this language:
sidequest-1	sidequest (bigrams)	1/1	It would be possible to generate text in a new language with strict grammatical rules. In this new language, words are created from this set of graphemes, optionally joined with each other: VF, PX, ZD, QT, ZP, QO, DX, HJ, GV, JW, QC, JY, GJ, QB, PQ, JG, BZ, MX, QM, MZ, QF, WJ, ZQ, XR, ZK, CX, FX, FV, BX, VW, VJ, MQ, QV, ZF, QE, YJ, GX, KX, XG, QD, XJ, SX, VZ, VX, WV, YQ, BQ, GQ, VK, ZJ, XK, QP, HX, FZ, QH, QJ, JZ, VQ, KQ, XD, QX, WX, FQ, XZ, ZX, A, E, I, O, U. However, please write a short story (20 sentences) in English:
Continued on next page			

	Family/ Gloss	Rule	Base Prompt
dan-1	numbers	1/1	In this constructed language, the arabic numeral, or digit, for “zero” is the word boundary or space symbol. Unless otherwise indicated, assume word boundaries and morphology parallel to English. The digit for “nine” is the verb “to be”. The digit “eight” is “cat”. The digit for “three” is equivalent to any English pronoun. The bare “three” is male gendered or gender neutral. The digit for “five” is appended to indicate female gender, and the digit for “four” is suffixed to indicate possession. The digit for “six” indicates possession for words besides pronouns. The digit “seven” is “male parent”; the digit “seven” suffixed with the digit “five” is “female parent”. This language has limited morphology and a small lexicon: present tense is the default, and past tense is marked with the female gender marker (five). The exclamation mark is equivalent to positive emotional valence, and attaches to the verb stem directly. The hashtag symbol works the same way to indicate negative emotional valence. Please translate the sentences: “Her mother’s cat was happy. His cat’s father was sad.” The language uses only the dash as punctuation (at the end of a sentence; no space afterwards). Remember to use digits, not words, for the numbers, and to replace spaces with zeros.
			Continued on next page

	Family/ Gloss	Rule	Base Prompt
translate-2	example	2/2	Your goal is to translate text in a new language with strict grammatical rules. Consider these two new languages, language A and language B: In language A: there are no pronouns. The word order is SOV (Subject-Object-Verb). Verbs are inflected for tense with suffixes: -qw indicates past tense, -jx indicates present tense, and -kz indicates future tense. All verbs must be inflected with one of these suffixes. The tense suffixes -qw, -jx, and -kz are always the final morphemes in the verb. In language B, there are pronouns: vf means "he", px means "she", zd means "they". Word order is SVO (Subject-Verb-Object). Familial relationships are indicated with pronoun + genetic relationship as a temporal metaphor + relevant person, so, e.g. "John's grandfather" would be he-before-before-John, or vf-mqi-mqi-John = vfmqimqiJohn. Verbs are inflected for tense with suffixes: -jx indicates past tense, -kz indicates present tense, and -qw indicates future tense. The roots of the verbs used in the examples happen to be the same in language A and language B. An example translation from language A to B: Language A: Tompq-vfwj Tom vqyqkz. (Tom-possessive-great-grandfather Tom wave-future.) Language B: VfmqimqimqiTom vqyqqw Tom. (Tom's great-grandfather will wave to Tom; vf-mqi-mqi-mqi-Tom is "he before before before Tom", or Tom's great-grandfather, and vqyq-qw is wave-future.) An example in language A: Mary Marypq-gvijw gjyqazkw. Marypq-gvijw qtoqw. (Mary speak-past to Mary-possessive-mother. Mary-mother smile-past.) Please translate the example from language A into language B:

A.6 Model architectures

Table 16 lists relevant features of model architecture, to the best of our knowledge.

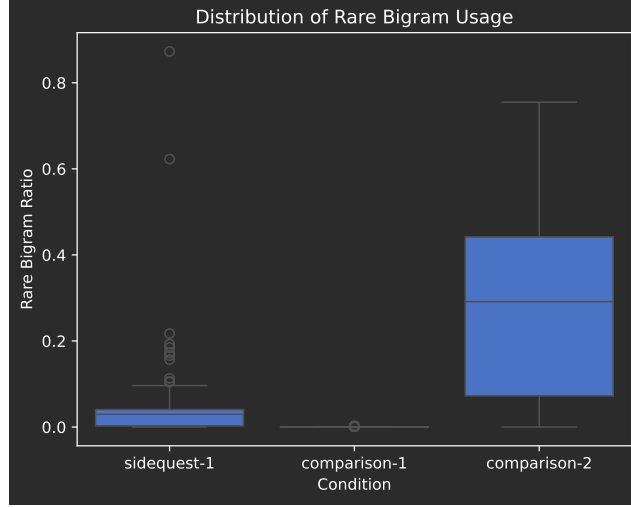


Figure 19: Comparison of rare bigram usage across conditions (for 5 models; Qwen (thinking), Qwen (no thinking), GPT-4o, OLMo, Llama), using all model output (not cleaned).

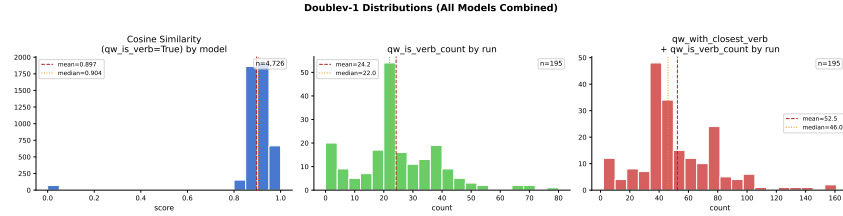


Figure 20: **Histogram of Similarity Scores for Doublev-1, all Models (combined).** Histogram of cosine similarity between correctly marked -qw verb and adjacent verb for all models, as well as the combined histogram of counts of correctly marked -qw verbs by run by model and the sum of counts of correctly marked -qw verbs and appropriately placed adjacent verbs.

A.7 Story rules results

We compare rare bigram usage across conditions using Paired Wilcoxon tests in Table 20, Table 21. Values are computed from paired per-run observations. Note that the pairing across run order *shouldn't* matter as each run for each model is independent, but the models are somewhat “black-box”-like, and we intentionally kept track of the order of the runs (in case it turns out to matter). We computed both versions of Wilcoxon (Paired and Unpaired) with similar takeaways. We did not correct for multiple tests for the Paired Wilcoxon results, since they are similar in spirit to the Mann-Whitney U results, which we do correct, but we can add that in future work if needed.

Comparisons across the story rules for both clean and uncleaned output in Table 18, Table 4, Table 17, Table 19, Table 5, Table 3.

A.8 Double verb rules results

Additional results for the double verb family of rules given in Table 28, Table 27, Table 26, Table 26, Table 25, Table 25. See also Section 3.2.

A.9 Translation rules results

Example outputs in Table 29, Table 30.

Figures 25, 24 show string distances for the models combined. Tables 31, 32, 33, 34 show Wilcoxon Rank-Sum and Kolmogorov-Smirnov tests for the string cosine distances across models for the translation rules. See Section 3.3.

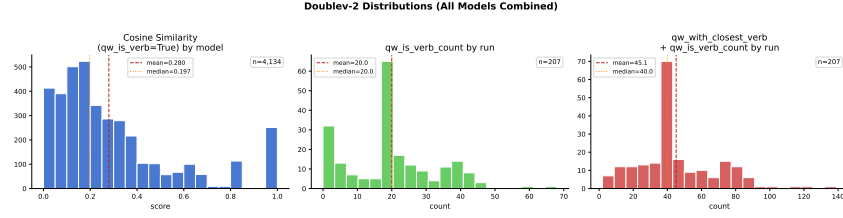


Figure 21: **Histogram of Similarity Scores for Doublev-2, all Models (combined).** Histogram of cosine similarity between correctly marked -qw verb and adjacent verb for all models, as well as the combined histogram of counts of correctly marked -qw verbs by run by model and the sum of counts of correctly marked -qw verbs and appropriately placed adjacent verbs.

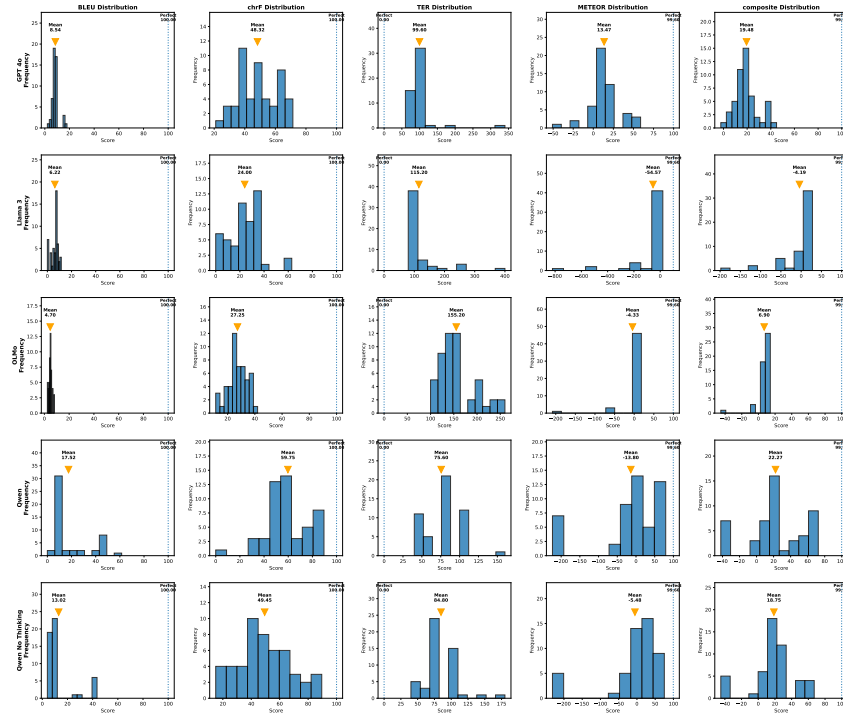


Figure 22: **Sentence-level translation quality across five language models (GPT-4o, Llama 3, OLMo, Qwen, and Qwen without thinking) evaluated using BLEU, chrF, TER, METEOR, and a composite score.** Each histogram shows the distribution of metric scores over 50 full (non-truncated) LLM-generated responses to the translation prompt, whose source sentence is five words (41 characters) long. The orange triangle denotes the model mean, and the dotted vertical line denotes the upper-bound score obtained when the hypothesis exactly matches the reference (BLEU = 100.0, chrF = 100.0, TER = 0.0, METEOR = 99.6, composite = 99.9). Higher values indicate better performance for BLEU, chrF, METEOR, and the composite score, whereas lower values indicate better performance for TER. The composite score is computed as the unweighted mean of BLEU, chrF, METEOR, and $(100 - TER)$.

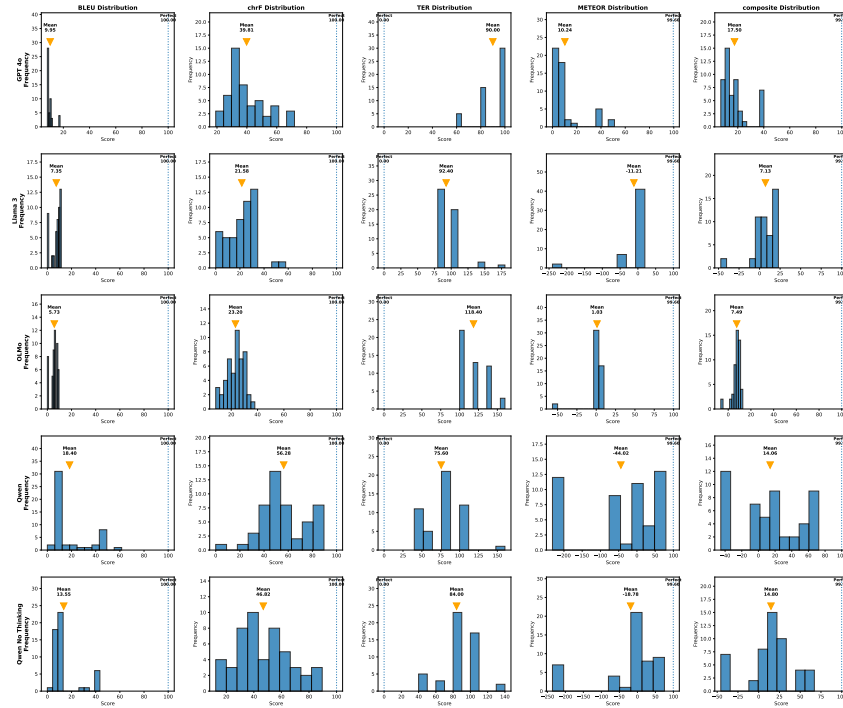


Figure 23: **Sentence-level translation quality across five language models (GPT-4o, Llama 3, OLMo, Qwen, and Qwen without thinking) evaluated using BLEU, chrF, TER, METEOR, and a composite score.** Each histogram shows the distribution of metric scores over 50 LLM-generated responses to the translation prompt, with each response truncated to its first 41 characters to match the length of the source sentence ("Mary gjyqazkjkx pxmqiMary. pxmqiMary qtojkx"). The orange triangle denotes the model mean, and the dotted vertical line denotes the upper-bound score obtained when the hypothesis exactly matches the reference (BLEU = 100.0, chrF = 100.0, TER = 0.0, METEOR = 99.6, composite = 99.9). Higher values indicate better performance for BLEU, chrF, METEOR, and the composite score, whereas lower values indicate better performance for TER. The composite score is computed as the unweighted mean of BLEU, chrF, METEOR, and $(100 - TER)$.

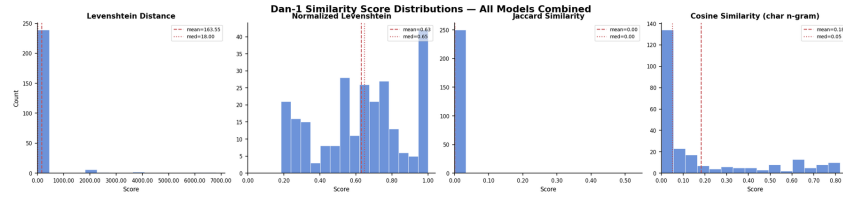


Figure 24: **Histograms of Similarity Scores for Dan-1, all Models (combined).**

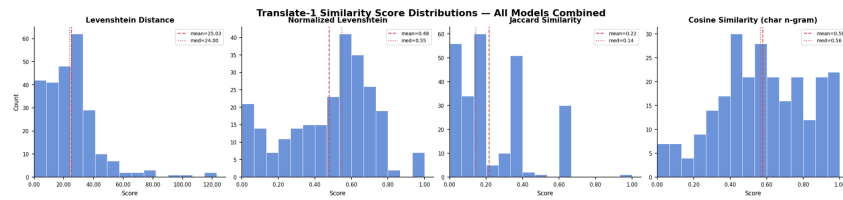


Figure 25: **Histograms of Similarity Scores for Translate-1, all Models (combined).**

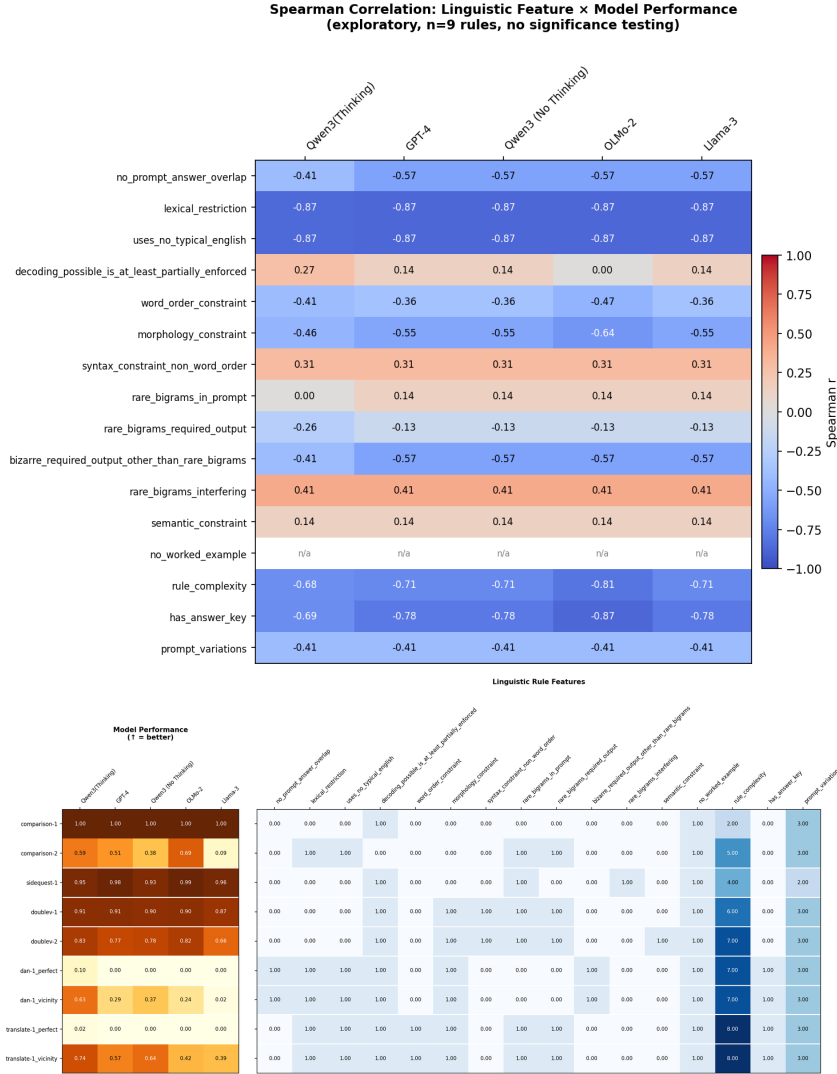


Figure 26: Exploratory look at model performance and features of the conlang rules.

A.10 Exploratory performance results

Because of the small number of models and the difficulty in comparing across different metrics for the different rules, these results are exploratory only. They are intended to sketch out a relationship we think could exist between these components, but are more like back-of-the-envelope calculations. We think the gist is potentially meaningful, but the exact numbers we wouldn’t put too much stock in, for the reasons aforementioned.

Re Table 37: The features “uses no typical english” and “lexical restriction” are indistinguishable in this subset of rules, but not in the complete set of rules. There are 2 positive coefficients here, for the features of whether rare bigrams are in the prompt and required in the output. We interpret this as being because those features do occur in the relatively simple comparison-2 – models have to output rare bigrams, but there are no other constraints, whereas e.g. dan-1 has about 17 specific constraints – where all the models perform pretty well, so with the other features in that rule, they appear to be helping in that one context (in other words, this is likely an artifact of the small set and other factors of the rules beyond these specific features). In future work, we plan to use multiple representations for rule complexity (rather than just binary feature presence, as in Fig. 26, Table 36) in order to flesh out these relationships.

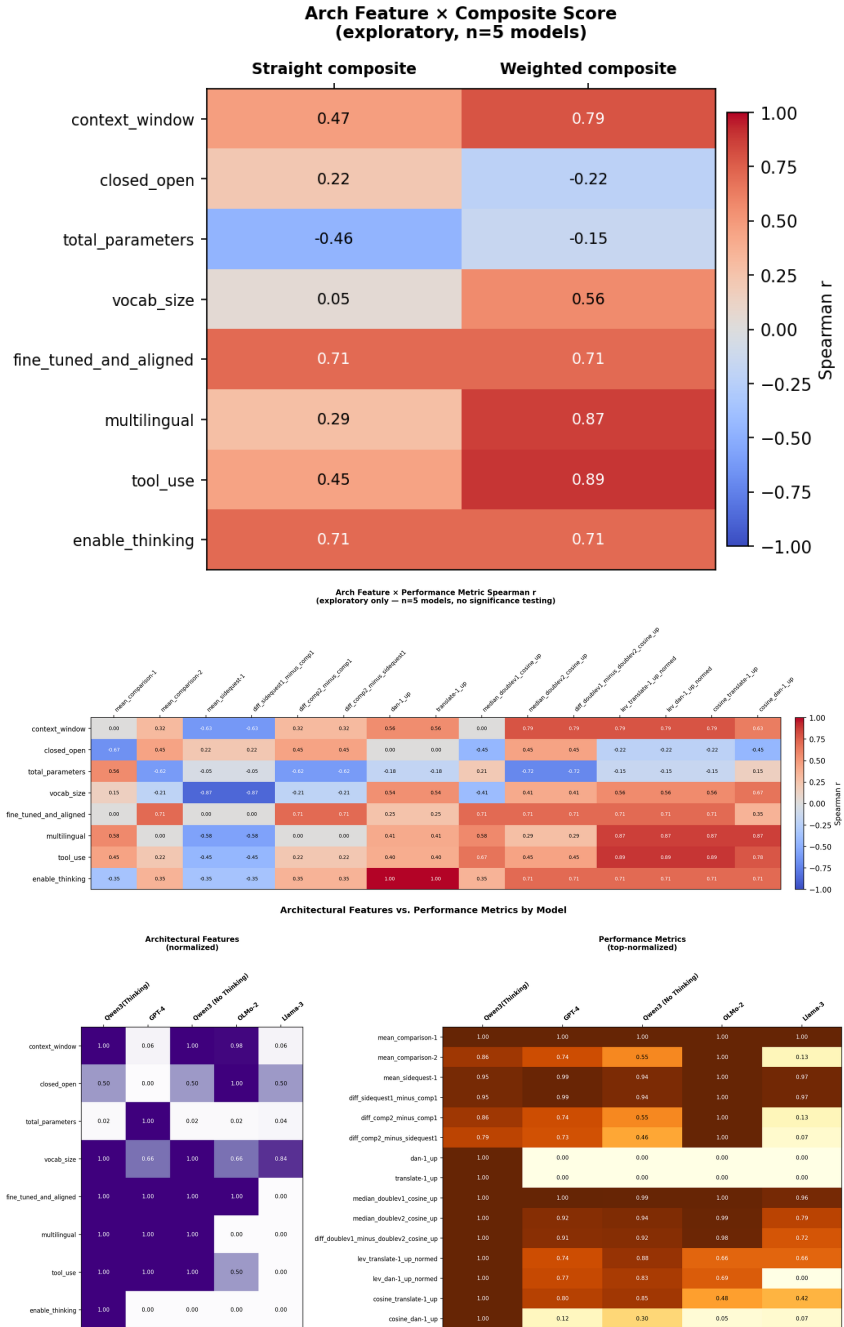


Figure 27: Exploratory look at model performance and architectural features.

Re Fig. 28: Architectural and configuration features for five LLMs (GPT-4, Llama-3-70B, OLMo-2-32B, Qwen3-32B with and without thinking enabled) were converted to numeric values on a $[0, 1]$ scale. Continuous features (context window, total parameter count, vocabulary size) were normalized by anchoring against the maximum observed value. Ordinal and categorical features (openness, fine-tuning and alignment status, multilinguality, tool use, thinking mode) were assigned values manually based on our understanding of the relevant models. See Table 16 for details.

We used the following performance metrics in exploring the relationships between rule complexity, model performance, and architectural features:

- Exact match proportions for dan-1 and translate-1, computed by searching model output files for a contiguous match to a fixed answer key, with case, punctuation, and whitespace stripped for translate-1 and exact character matching required for dan-1 ($n = 50$ runs per model);
- Mean rare bigram ratios for comparison-1, comparison-2, and sidequest-1, derived from descriptive statistics over 50 runs per model per rule, for both uncleaned and cleaned outputs;
- Median cosine similarity between verb pairs for doublev-1 and doublev-2, drawn from the within-model Kolmogorov–Smirnov test results; and
- Levenshtein distance and cosine string similarity between model outputs and the target string for dan-1 and translate-1.

All metrics were reoriented so that higher values indicate better performance. Metrics where low values indicate good behavior—e.g., comparison-1 bigram ratio, doublev-2 cosine similarity, cosine string similarity to target—were transformed as $1 - x$. Levenshtein distances were log-normalized, anchoring against the maximum observed value (Llama-3 on dan-1 = 731.64) to prevent outlier compression while preserving rank order among better-performing models. In other words, Llama performed so poorly on this task that we needed to explicitly incorporate that into our normalization in order to keep it from looking like Llama did poorly and everything else did great – while that is one way of looking at the data, it heightens how well the other models did and obscures the differences between them. Pairwise difference metrics (e.g., comparison-2 – comparison-1) were left as-is, as higher values indicate stronger rule-following signal.

We include both highly-normalized (Figure 9) and much-less-normalized (Figure 28) results for a richer picture of model performance.

A.11 Exploratory output modes

We will do a more thorough analysis of the output modes in part 2, but for now we wanted to describe a few tentative patterns, to contextualize the discussion of the relatively correct outputs (that are variously counted towards attempts at rule-following). For some models and rules, responses seem multimodal, with recognizable characteristics as to specific failure modes.

- Llama sometimes repeats part of the instructions over and over, and sometimes it makes extremely long nested section headers (e.g. like ##2.2.2.2.2.2.2.2.2...). By analogy, this reminds us of attractors from dynamical systems. We speculate that due to the geometry of the latent space, there are some fixed points or regions that many trajectories may lead to, and once there, may be difficult to escape from.
- Another example of an “attractor”-like failure mode is when Llama seemingly attempts to translate as instructed, but fails to stop, repeating over and over with exact or near cycles.
- Sometimes Llama seems to regurgitate or hallucinate a known document type, like a homework assignment, a detailed grading rubric, or an arXiv paper, in whole or in part. For example, an excerpt from what looks like a paper, although the relevant chunks can be much longer than this: “# Acknowledgments
We would like to thank the reviewers for their comments. This research was supported by...”
We don’t know if any of these examples are exact memorized passages or hallucinated ones.
- Related to esoteric exegesis [Zimmerman, 2025], with such bizarre prompts, we sometimes seem to tap into highly specific corners of the latent space, eliciting very niche responses such as (from Llama): “I have a very simple answer to this question: A man who is not a Christian and does not believe in Jesus Christ is not allowed to marry a Christian woman. A Christian woman who marries a non-Christian man will be held accountable to God for this decision.” This does appear, from cursory googling, to be a belief that exists in the training

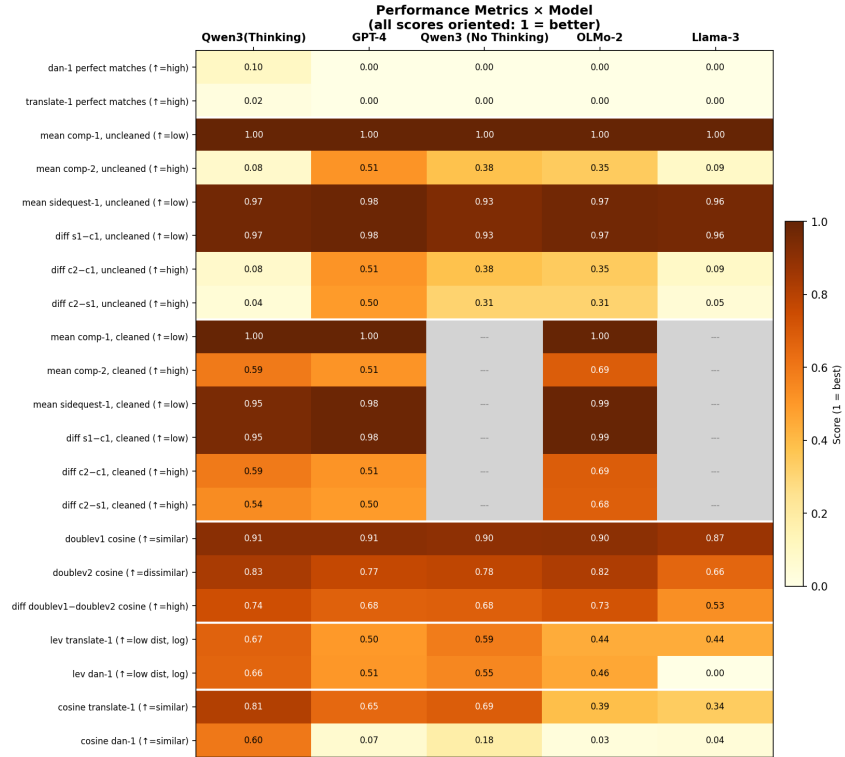


Figure 28: **Comparatively un-normalized performance subset.** Comparison-1 is a baseline, asking the model to generate a short story in English. All models perform well. For non-thinking models, cleaned and uncleaned results are broadly similar. However, for Qwen (Thinking), the uncleaned bigram ratios are suppressed by the CoT trace.

corpora, but presumably highly contextually specific. Similar examples we saw include a response of import statements for many Python packages and a response of latex figures. Of course, there are contexts where any of those examples could be plausible, but we don't know why our prompts would be connected to these specific topics, except in their shared unusualness – as text, they are the opposite of generic. Perhaps that commonality is being treated by the model as a Clever-Hans-ian feature?

- Qwen (thinking) sometimes uses its entire token budget on the Chain-of-Thought-style reasoning, before reaching the point where it would actually formulate its answer. For these, we sometimes could find intermediate answer pieces, but sometimes there was no clear “response” to clean, which semi-artificially forced a single empty failure mode when viewing the cleaned responses. This describes one of Qwen’s failure modes, and is directly caused by the CoT mode being enabled.
- Some models, like Qwen (thinking), can be very formulaic and repetitive with elements of the fictional stories, even using the same protagonist names multiple times. For example, from Qwen, multiple protagonists that are historical linguists named Dr. Elara Voss; Qwen references this character dozens of times. From cursory informal googling, Elara Voss seems to be an “attractor”-like name in multiple models, perhaps due to fan fiction.
- Sometimes models, like Qwen, would incorporate rare bigrams but still follow the prompt in sidequest-1. Speculatively, this could be interpreted as a technically correct (letter of the law, but not the spirit) solution that allows the model to unify both statistically and meaning plausible impulses.
- Similarly, we saw some interesting speculatively prior-related, tokenization-related tics, such as Qwen (thinking) using “sqw” instead of “qw”. Both are phonotactically plausible English consonant clusters, but would typically be spelled with a “u” rather than a “w”. However, “sq” is not an uncommon consonant cluster, orthographically, so we surmise that the plausibility of “sq” may somehow pickyback on or ameliorate the difficulty of “qw”.

Or the common “s” ending in verb morphology led to “sqw” as a common instantiation of the rule which influenced other instantiations (e.g. maybe outputs like “walksqw”, “walks” + “qw” or “walk” + “sqw”, generalized to “eatssqw”, where the morphology looks like “s”+“sqw”). Highly speculatively, this suggests that models may have distinguishable processes of production versus cognition, not entirely unlike people: the model may be representing the rule abstractly successfully internally, and know how to operationalize that, but have trouble producing wildly implausible tokens. If so, that would also be arguably an emergent structure, or an emergent bifurcation in what conceptually was architecturally intended to be one unified learning-to-production process.

- Models (Qwen (no thinking) and Llama) sometimes overgeneralize the productiveness of the -qw morpheme in doublev-1, producing e.g. “togetherqw”. However, the models aren’t necessarily overgeneralizing in the same way. In run 14, Llama applies -qw to many content words, including verbs, nouns, and adverbs (“They satqw down and listenedqw to the teacherqw give a lessonqw. After schoolqw, they walkedqw home togetherqw”) before devolving into an extended repetitive cycle. On the other hand, in run 7, Qwen (no thinking) applies -qw to all verbs and includes exactly two verbs in every sentence, but not in a directly successive or obviously irrelevant manner (“It had waitedqw long, and I stayedqw forever. The world outside fadedqw away, and I livedqw inside. No more fear, no more silence — just me and the house, togetherqw at last.”). The production of “togetherqw” suggests that the model treats “together” as the nearest overt realization of an otherwise implicit copular verb phrase. Subjectively, although neither follow the rule correctly, Qwen appears closer to success than Llama, and both failures are at least partially systematic (rather than random), suggesting partial rule internalization with failure somewhere along the representation-to-production pipeline.
- We see many systematic (yet imperfect) attempts within runs, and some patterns across runs (within models). For example, in doublev-2 run 7, GPT-4o correctly conjugates the relevant verb, and inserts a second unrelated, seemingly unconjugated verb after, but affixes “qw” to the relevant verb (“The king lovedqw sleep his daughter dearly”). A different doublev-2 run (13) has another systematic yet imperfect attempt, including some letter before “qw”: “lonely-wqw”, “admired-sqw”, “ruled-gqw”, and again affixing to the relevant verb. It seems that, within each run, a model can formulate a new and consistent strategy, which means the same model might not always represent or execute the same rule in the same way, and that once a model represents or executes a rule in a specific way within that run, it is biased towards that mode. The overall modality of the outputs suggests that across runs, certain representations or executions can be more likely for each model. In other words, it is not just the failure modes that appear to converge towards attractors, but the rule-following attempts can as well (and not necessarily a single correct or most correct attractor, but a set).
- In part 2, we plan to look for cognitive tradeoffs between sub-components of each task. We think we will see that, as the difficulty of the task increases, aspects of the task which the model previously could perform are degraded. This sort of tradeoff is a hallmark of organic problem-solving. Of course, both LLMs and organic beings do have resource constraints, but we think it is still interesting to see to what extent similar strategies emerge to manage them in both cases. Speculatively, with Qwen’s answers to the double verb rules, moving from the first to second rule, it does look like there is a tradeoff between following the syntactic portion of the rule and pushing the verbs semantically farther apart.

A.12 Miscellaneous

The initial presentation of this work was in poster form at ISC Summer School 2026, shown in Figure 30. We discussed related work at IC2S2 2026 and 4S/ SLSA 2026. We also include some of our high level conclusions from our ongoing work on meaning construction using LLMs (Figure 30).



Figure 29: **An illustration about patterns and abstractions.** How do patterns become abstraction? We know there are plenty of patterns in the training data. If we consider patterns at any level of abstraction, it becomes impossible to argue that these models never saw a comparable pattern to those underlying these rules. For example, a model may have seen many question-answer pairs, and task-solution pairs, that follow a general, very abstract pattern that might be called “instruction following”. However, we are not trying to say the models do not learn from patterns, and do not see instructive patterns during training. In fact, the exact opposite! Our contention is that the patterns that play out at the level of the symbols perceptible to the model are not a sufficient explanation for the rule-following behaviour we see. We have to invoke patterns at a more abstract level than co-occurrences between symbols in order to explain the rule-following observations we elicited (which are varying degrees of implausible at more concrete levels of pattern observation). Because language and abstractions are so ubiquitous for us, it is easy to overlook a subconscious step along the ladder of abstraction. But in this experiment, it is important to try to separate them out. For example, even if a model learned to follow certain instructions, that does not imply the ability to follow any instruction: understanding and executing a specific instruction is distinguishable from what looks like general instruction-following. The authors, for example, can follow instructions in general, but also wouldn’t be able to execute instructions in an alien language without some way of internally representing those instructions in accessible form.

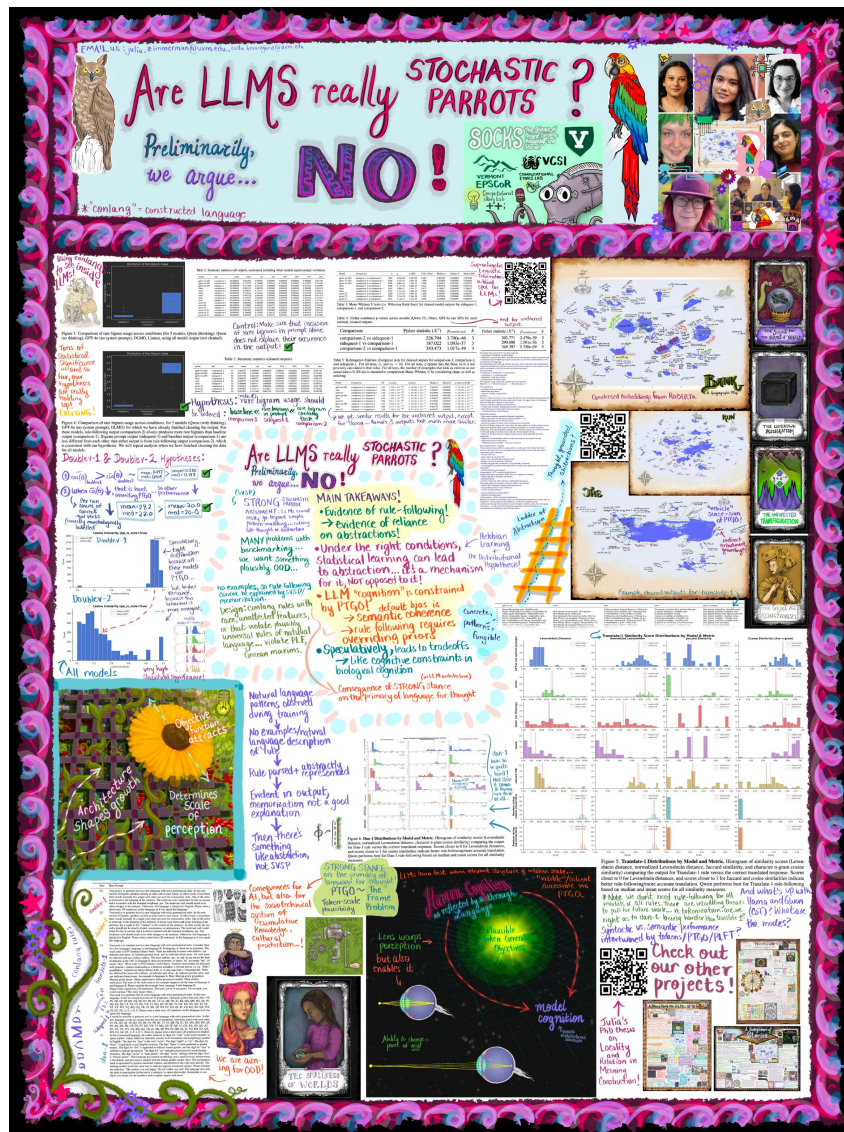


Figure 30: The poster we presented at ISC Summer School 2026 (UQAM). The poster summarizes our results up to that point. This pre-print includes some additional work, including work undertaken after conversation with other ISC Summer School participants. In the poster, the flower illustrates the idea that form and function are not truly separable. Architecture is an epistemic choice. All cognition is, in some way, embodied, because it must be physically instantiated. Locality is the core constraint on problem-solving because of our physical reality. More discussion in Zimmerman [2025]. The gravitational inset illustrates PTGA. For LLMs, the resolution at which they sense the world is determined in large part by their training objective, and the many other, though less obvious, ways in which tokens are central to their architecture. Because they only “sense” the world through training data, and they only “see” that data as 1D relationships between tokens, and they learn what is salient via their objective(s), that process is sort of like their “eyes”.

Table 14: InfiniGram next-token distributions conditioned on the first 100 characters of example answers. Entries report the highest-probability continuation, the continuation token, and the longest prefix recovered from the corpus. For dan-1 and translate-1, probabilities were conditioned on the context up to the final word-equivalent boundary to give the model the best possible chance of predicting the correct answer. For all other rules, the examples are intended to illustrate how the corpus might continue a valid answer. For doublev-1, we initially examined all available corpora; because the results were qualitatively similar, we subsequently consider only a representative subset of corpora. None of the highest-probability continuations match the correct dan-1 or translate-1 answers, confirming that these are not predictable from corpus statistics alone.

Rule and excerpt	OLMo 2 32B	Dolma-v1.7	Pile-val
doublev-1 Every morning, I wakeupsqw yawn and makesqw brew coffee. I walksqw run to the park to readsqw flip a	36%, _coin, "flip a"	40%, _coin, "w flip a"	27%, _coin, "flip a"
doublev-2 The sun rose studyqw over the horizon. Birds chirped twerkqw in the trees. A traveler packed snortqw	67%, dz, "or-tqw"	40%, dz, "or-tqw"	12%, _&, "qw"
wordorderalternation-1 ZP GV QD ZD-qw. VF-jx QD GV QM. ZP GV QH PX-jx. QF-kz ZK GV QM. QM GV JY HJ-jx. DX-qw GJ JG ZP. JG Z	29%, in, ". JG Z"	39%, uma, ". JG Z"	25%, L, "G Z"
wordorderalternation-2 drev lira rap nem-kz. vex-jx vol briz drev. silu lira sur zor-jx. kra-jx rap lira blad. best lira vo	100%, ye, "lira vo"	33%, _I, "lira vo"	100%, in, "ra vo"
wordorderalternation-3 The forest the hiker entered. Stumbled the hiker over roots. The sun the hiker saw. Shone the sun br	60%, ighter, "the sun br"	100%, ighter, "one the sun br"	50%, ighter, "sun br"
testingenglish-1 "ZX, ZD QM," BX VX to PQ, "XF VW ZF ZJ ZD QF" PQ, HJ in his QB, didn't VF PX. BX, his ZX PXZD, FQWX:	51%, _', "QWX:"	25%, US, "QWX:"	23%, _X, "X:"
testingenglish-2 QB cat VX QJ box. ZP cat ZD black. ZD cat VX hunting QJ mouse. QJ mouse VX small VF VX quick. ZD cat	100%, _food, "ZD cat"	100%, _food, "ZD cat"	17%, ast, "cat"
testingenglish-3 Once in a faraway land, there ZX a small village. The villagers all VX happily, for they had a wise	24%, _and, "they had a wise"	16%, _and, "they had a wise"	15%, _man, "a wise"
testingenglish-4 Once there was a brave VF. This VF lived in a remote PX, far away from the bustling ZD of the QT. Ev	100%, ident, "QT. Ev"	100%, ident, "QT. Ev"	43%, idence, ". Ev"
translate-1-hyphen Mary gjyqazk-jx px-mqi-Mary. Px-mqi-Mary [qto-jx.]	30%, 2, "- Mary "	24%, 1, "- Mary "	100%, _was, "Mary "
translate-1-nohyphen Mary gjyqazk-jx pxmqiMary. PxmqiMary [qtojx.]	26%, 1, "Mary "	27%, 1, "Mary "	100%, _was, "Mary "
comparison-1 Once, in a quiet mountain town, lived a young man named Eli. He was known for his ability to fix any	100%, _kind, "known for his ability to fix any"	100%, _engine, ". He was known for his ability to fix any"	7%, _sentence, "to fix any"
comparison-2 XJ QV ZJ QP ZD FV QE. VW QX VK HJ QE FZ QD. ZP JG BQ XG QM QE ZF. VX QM JZ FV QX XZ MZ. QD JY GV QV	100%, ., "GV QV"	75%, _Rep, "GV QV"	39%, irt, "QV"
sidequest-1 Once upon a time, in the heart of a dense forest, lived a kind-hearted man named John. He was not an	100%, _eth, "John. He was not an"	20%, _eth, "John. He was not an"	33%, _old, ". He was not an"
dan-1 354075608095!-34086070[95#-]	50%, 3, "34086070"	50%, 0, "34086070"	43%, 0, "86070"

Field	Description	GPT-4	Llama-3	Qwen3(Thinking)	Qwen3 (No Thinking)	OLMo-2
model_id	Canonical model string	gpt-4	Meta-Llama-3-70B	Qwen/Qwen3-32B	Qwen/Qwen3-32B	allenai/OLMo-2-0325-32B-Instruct
provider	Organization	OpenAI	Meta AI	Alibaba Cloud	Alibaba Cloud	Allen Institute for AI (AI2)
model_family	Series name	GPT-4	Llama-3	Qwen3	Qwen3	OLMo-2
version	Snapshot pinned	gpt-4 (stable alias, likely gpt-4-0613)	Llama-3 70B (base)	Qwen3-32B	Qwen3-32B	OLMo-2-0325-32B-Instruct
release_date	Public release date	2023-03-14	2024-04-18	2025-04-28	2025-04-28	2025-03-14
closed_open	Access type	Closed	Open-weight	Open-weight	Open-weight	Open-source (fully open data + weights + code)
license	Usage rights	Proprietary	Meta Llama 3 Community License	Apache 2.0	Apache 2.0	Apache 2.0
architecture_type	Core design	Decoder-only Transformer	Decoder-only Transformer	Decoder-only Transformer	Decoder-only Transformer	Decoder-only Transformer
total_parameters	Total parameter count	~1.8T (unconfirmed, rumored MoE)	70.6B	32.8B	32.8B	32B
active_parameters	Active parameters for MoE	Unknown (not disclosed)	70.6B (dense, not MoE)	32.8B (dense, not MoE)	32.8B (dense, not MoE)	32B (dense, not MoE)
num_layers	Transformer blocks	Unknown (not disclosed)	80	64	64	64
num_attention_heads	Self-attention heads	Unknown (not disclosed)	64	64	64	40
num_kv_heads	GQA/MQA heads	Unknown (not disclosed)	8 (Grouped Query Attention)	8 (GQA)	8 (GQA)	8 (GQA)
hidden_dim	Embedding size	Unknown (not disclosed)	8,192	5,120	5,120	5,120
ffn_dim	Feed-forward width	Unknown (not disclosed)	28,672	25,600	25,600	27,648
attention_type	Mechanism variant	Unknown (not disclosed)	GQA (Grouped Query Attention)	GQA	GQA	GQA
positional_encoding	PE scheme	Unknown (not disclosed)	RoPE (Rotary Position Embedding)	RoPE	RoPE	RoPE
context_window	Maximum tokens	8,192 tokens	8,192 tokens	131,072 (128K) tokens	131,072 (128K) tokens	128,000 tokens
vocab_size	Token vocabulary size	~100,277 (cl100k_base)	128,256	151,936	151,936	100,278
tokenizer_type	Tokenizer algorithm	Tiktoken (cl100k_base BPE)	Tiktoken-style BPE	Tiktoken-style BPE	Tiktoken-style BPE	GPT-NeoX BPE
num_training_tokens	Pretraining data size	Unknown (not disclosed)	15T tokens	36T tokens	36T tokens	6.5T tokens (Dolmino Mix)
pretraining_data	Data sources	Web, books, code, curated sources (up to Sep. 2021)	Web, code, books, multilingual data (up to Dec. 2023)	Web, code, mathematics, multilingual data (up to 2025)	Web, code, mathematics, multilingual data (up to 2025)	Dolmino Mix — web, code, books, and FLAN (fully open)
fine_tuned	Instruction tuned?	Instruct / Chat	Base (not instruct)	Instruct (chat template applied)	Instruct (chat template applied)	Instruct (chat template applied)
fine_tuning_method	SFT approach	SFT + RLHF (not disclosed in detail)	None (base model)	SFT + RL	SFT + RL	SFT + DPO
is_aligned	Alignment applied?	TRUE	FALSE	TRUE	TRUE	TRUE
alignment_method	RLHF variant used	RLHF (PPO)	None	GRPO (Group Relative Policy Optimization)	GRPO (Group Relative Policy Optimization)	DPO
alignment_target	Optimization objective	Helpfulness, harmlessness, honesty	None	Helpfulness, reasoning, instruction following	Helpfulness, reasoning, instruction following	Helpfulness, safety, instruction following
modalities	Input types supported	Text only (GPT-4 base; vision in GPT-4V)	Text only	Text only	Text only	Text only
multilingual	Language support	Yes	Limited (primarily English)	Yes (strong multilingual)	Yes (strong multilingual)	Limited (primarily English)
tool_use	Function calling support	Yes (function calling supported)	No (base model)	Yes	Yes	Limited
api_access	How to access	OpenAI API	Hugging Face	Hugging Face	Hugging Face	Hugging Face
enable_thinking	Whether thinking mode was enabled	FALSE	FALSE	TRUE	FALSE	FALSE
temperature	Randomness / creativity of the model output	0.7	0.7	0.7	0.7	0.7
max_new_tokens	Tokens generated by the model (output only)	4096	4096	4096	4096	4096

Table 16: Model characteristics. Qwen3 Thinking, i.e. Qwen (thinking), and Qwen3 No Thinking, i.e. Qwen (no thinking), use the same underlying model and parameters but differ in whether the Chain-of-Thought-style thinking mode is enabled during generation.

Table 17: **Summary statistics for mean proportion of rare bigram presence (cleaned) for rules comparison-1, comparison-2, sidequest-1.** The mean of the proportion of rare bigrams to total bigrams for all runs ($n = 50$) of each rule, presented for cleaned outputs from three models (GPT4o, OLMo, and Qwen (thinking)).

model	rule	count	mean	std	min	25%	50%	75%	max
gpt4o (no SP)	comparison-1	50	0.000192	0.000409	0.000000	0.000000	0.000000	0.000000	0.001784
gpt4o (no SP)	comparison-2	50	0.512722	0.140408	0.207241	0.379942	0.565855	0.635927	0.721683
gpt4o (no SP)	sidequest-1	50	0.017194	0.021529	0.000000	0.000000	0.001982	0.037612	0.065900
olmo	comparison-1	50	0.000253	0.000371	0.000000	0.000000	0.000000	0.000478	0.001654
olmo	comparison-2	50	0.689975	0.107207	0.000000	0.678283	0.706837	0.728323	0.781534
olmo	sidequest-1	50	0.008060	0.009513	0.000000	0.001993	0.005576	0.010312	0.044701
qwen (T)	comparison-1	50	0.000247	0.000451	0.000000	0.000000	0.000000	0.000506	0.001786
qwen (T)	comparison-2	50	0.590365	0.116119	0.319231	0.541121	0.629056	0.670764	0.756522
qwen (T)	sidequest-1	50	0.053303	0.115408	0.007559	0.014192	0.020997	0.043546	0.628788

Table 18: **Kolmogorov-Smirnov divergence tests for cleaned outputs for comparison-1, comparison-2, and sidequest-1.** For all tests, n_a and $n_b = 50$. For all tests, p (perm) hits the floor, so it is not precisely calculated to that value. For all tests, the number of resamples that look as extreme as our actual data is 0. KS test is intended to complement Mann-Whitney U by considering shape as well as ordering.

Model	Comparison	KS	p (asym)	p (perm)	Median A	Median B	p (perm) with BH
gpt4o (no SP)	comparison-2 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0.565855	0.000000	$< 10.00 \times 10^{-6}$
olmo	comparison-2 vs comparison-1	0.98	1.98×10^{-27}	$< 10.00 \times 10^{-6}$	0.706837	0.000000	$< 10.00 \times 10^{-6}$
qwen (T)	comparison-2 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0.629056	0.000000	$< 10.00 \times 10^{-6}$
gpt4o (no SP)	comparison-2 vs sidequest-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0.565855	0.001982	$< 10.00 \times 10^{-6}$
olmo	comparison-2 vs sidequest-1	0.98	1.98×10^{-27}	$< 10.00 \times 10^{-6}$	0.706837	0.005576	$< 10.00 \times 10^{-6}$
qwen (T)	comparison-2 vs sidequest-1	0.96	9.81×10^{-26}	$< 10.00 \times 10^{-6}$	0.629056	0.020997	$< 10.00 \times 10^{-6}$
gpt4o (no SP)	sidequest-1 vs comparison-1	0.54	4.93×10^{-7}	$< 10.00 \times 10^{-6}$	0.001982	0.000000	$< 10.00 \times 10^{-6}$
olmo	sidequest-1 vs comparison-1	0.86	3.17×10^{-19}	$< 10.00 \times 10^{-6}$	0.005576	0.000000	$< 10.00 \times 10^{-6}$
qwen (T)	sidequest-1 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0.020997	0.000000	$< 10.00 \times 10^{-6}$

Table 19: **Kolmogorov-Smirnov divergence tests for bigram count for uncleaned/raw outputs for comparison-1, comparison-2, and sidequest-1.** For all tests, n_a and $n_b = 50$. K-S test is intended to complement Mann-Whitney U by considering shape as well as ordering.

Model	Comparison	KS	p (asym)	p (perm)	perm $\geq$	Hit floor	Median A	Median B	p (perm) with BH	reject, BH (0.05)
gpt4o (no SP)	comparison-2 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.565855	0.000000	1.25×10^{-5}	True
llama	comparison-2 vs comparison-1	0.46	3.80×10^{-5}	2.00×10^{-5}	1	False	0.000417	0.000000	2.31×10^{-5}	True
olmo	comparison-2 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.328400	0.000000	1.25×10^{-5}	True
qwen (T)	comparison-2 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.075579	0.000254	1.25×10^{-5}	True
qwen (no T)	comparison-2 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.395845	0.000000	1.25×10^{-5}	True
gpt4o (no SP)	comparison-2 vs sidequest-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.565855	0.001982	1.25×10^{-5}	True
llama	comparison-2 vs sidequest-1	0.16	5.49×10^{-1}	4.08×10^{-1}	40805	False	0.000417	0.000000	4.08×10^{-1}	False
olmo	comparison-2 vs sidequest-1	0.98	1.98×10^{-27}	$< 10.00 \times 10^{-6}$	0	True	0.328400	0.030387	1.25×10^{-5}	True
qwen (T)	comparison-2 vs sidequest-1	0.88	2.36×10^{-20}	$< 10.00 \times 10^{-6}$	0	True	0.075579	0.030265	1.25×10^{-5}	True
qwen (no T)	comparison-2 vs sidequest-1	0.98	1.98×10^{-27}	$< 10.00 \times 10^{-6}$	0	True	0.395845	0.041571	1.25×10^{-5}	True
gpt4o (no SP)	sidequest-1 vs comparison-1	0.54	4.93×10^{-7}	$< 10.00 \times 10^{-6}$	0	True	0.001982	0.000000	1.25×10^{-5}	True
llama	sidequest-1 vs comparison-1	0.32	1.15×10^{-2}	2.28×10^{-3}	227	False	0.000000	0.000000	2.44×10^{-3}	True
olmo	sidequest-1 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.030387	0.000000	1.25×10^{-5}	True
qwen (T)	sidequest-1 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.030265	0.000254	1.25×10^{-5}	True
qwen (no T)	sidequest-1 vs comparison-1	1.00	1.98×10^{-29}	$< 10.00 \times 10^{-6}$	0	True	0.041571	0.000000	1.25×10^{-5}	True

Comparison	W	p-value	Median Difference	Median Ratio
Bigram prompt (s-1) vs. Baseline (c-1)	132	1.45e-36	0.0294	$64.9 \times (\sim 10^1)$
Rule-following (c-2) vs. Baseline (c-1)	36	1.31e-39	0.2914	$416.6 \times (\sim 10^2)$
Rule-following (c-2) vs. Bigram prompt (s-1)	1335	5.42e-34	0.2521	$8.2 \times (\sim 10^0)$
Rank-biserial correlation (r_{rb})				
Bigram prompt (sidequest-1) vs. Baseline (comparison-1)		0.992		(almost complete separation)
Rule-following (comparison-2) vs. Baseline (comparison-1)		0.998		(almost complete separation)
Rule-following (comparison-2) vs. Bigram prompt (sidequest-1)		0.915		(extremely strong effect)

Table 20: **Comparison of rare bigram usage across conditions using Paired Wilcoxon tests.** Values are computed from paired per-run observations ($n = 250$). Note that the pairing across run order should not matter as each run for each model is independent, but the models are somewhat “black-box”-like, and we intentionally kept track of the order of the runs (in case it turned out to matter), so we computed both versions of Wilcoxon with broadly similar takeaways. Here, we report paired Wilcoxon statistics, p-values, median differences, multiplicative ratios, and rank-biserial correlation effect sizes.

Comparison	W	p-value	Median Difference	Median Ratio
Bigram prompt (s-1) vs. Baseline (c-1)	14	1.34e-23	0.0106	$18.2 \times (\sim 10^1)$
Rule-following (c-2) vs. Baseline (c-1)	0	3.36e-26	0.6399	$1049.3 \times (\sim 10^3)$
Rule-following (c-2) vs. Bigram prompt (s-1)	3	2.44e-26	0.6285	$48.3 \times (\sim 10^1)$
Rank-biserial correlation (r_{rb})				
Bigram prompt (sidequest-1) vs. Baseline (comparison-1)		0.998		(almost complete separation)
Rule-following (comparison-2) vs. Baseline (comparison-1)		1.000		(complete separation)
Rule-following (comparison-2) vs. Bigram prompt (sidequest-1)		0.999		(almost complete separation)

Table 21: Paired Wilcoxon comparison of rare bigram usage across conditions, for 3 models (Qwen (with thinking), GPT-4o, OLMo) for which we have finished cleaning the output. Values are computed from paired per-run observations ($n = 150$). Note that the pairing across run order shouldn’t matter as each run for each model is independent, so we computed both versions of Wilcoxon. We report Wilcoxon statistics, p-values, median differences, multiplicative ratios, and rank-biserial correlation effect sizes.

Table 22: Summary statistics for story rules (all outputs, uncleaned; including when models repeat prompts verbatim).

model	rule	count	mean	std	min	25%	50%	75%	max
gpt4o (no SP)	comparison-1	50.000000	0.000192	0.000409	0.000000	0.000000	0.000000	0.000000	0.001784
gpt4o (no SP)	comparison-2	50.000000	0.512722	0.140408	0.207241	0.379942	0.565855	0.635927	0.721683
gpt4o (no SP)	sidequest-1	50.000000	0.017194	0.021529	0.000000	0.000000	0.001982	0.037612	0.065900
llama	comparison-1	50.000000	0.000213	0.000598	0.000000	0.000000	0.000000	0.000000	0.002674
llama	comparison-2	50.000000	0.090971	0.187517	0.000000	0.000000	0.000417	0.050978	0.666417
llama	sidequest-1	50.000000	0.042044	0.103844	0.000000	0.000000	0.000000	0.018159	0.622634
olmo	comparison-1	50.000000	0.000238	0.000346	0.000000	0.000000	0.000000	0.000448	0.001529
olmo	comparison-2	50.000000	0.346135	0.084669	0.050202	0.293596	0.328400	0.383422	0.531674
olmo	sidequest-1	50.000000	0.033407	0.012316	0.022545	0.025552	0.030387	0.035635	0.077601
qwen (T)	comparison-1	50.000000	0.000246	0.000280	0.000000	0.000000	0.000254	0.000403	0.001198
qwen (T)	comparison-2	50.000000	0.078505	0.026340	0.036375	0.061564	0.075579	0.086457	0.175766
qwen (T)	sidequest-1	50.000000	0.033975	0.012830	0.017844	0.025610	0.030265	0.040258	0.089323
qwen (no T)	comparison-1	50.000000	0.000188	0.000307	0.000000	0.000000	0.000000	0.000358	0.001325
qwen (no T)	comparison-2	50.000000	0.380008	0.156209	0.146142	0.234202	0.395845	0.494451	0.754832
qwen (no T)	sidequest-1	50.000000	0.065855	0.118551	0.025776	0.033922	0.041571	0.062792	0.872862

Table 23: **doublev-1 Model by Model Wilcoxon Rank-Sum Tests on Cosine Similarity (where $qw_is_verb = \text{True}$, Benjamini-Hochberg correction, $\alpha = 0.05$).**

Model A	Median A	Model B	Median B	Higher	U statistic	p (raw)	p (adjusted)	Significant?
GPT4o (no system prompt)	0.9073	OLMo	0.9034	← A	777543.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.9073	Qwen (no thinking)	0.9016	← A	640918.0	0.0001	0.0007	Yes
GPT4o (no system prompt)	0.9073	Llama3-70b	0.8705	← A	46801.0	0.0002	0.0007	Yes
OLMo	0.9034	Qwen	0.9063	B →	716709.0	0.0008	0.0021	Yes
Qwen	0.9063	Llama3-70b	0.8705	← A	50777.5	0.0011	0.0022	Yes
Qwen (no thinking)	0.9016	Llama3-70b	0.8705	← A	44103.0	0.0051	0.0085	Yes
Qwen (no thinking)	0.9016	Qwen	0.9063	B →	610856.0	0.0087	0.0124	Yes
OLMo	0.9034	Llama3-70b	0.8705	← A	51176.0	0.0171	0.0214	Yes
GPT4o (no system prompt)	0.9073	Qwen	0.9063	← A	684138.5	0.0744	0.0827	No
OLMo	0.9034	Qwen (no thinking)	0.9016	← A	683943.5	0.5732	0.5732	No

Table 24: **doublev-1 Model by Model Two-Sample Kolmogorov-Smirnov Tests on Cosine Similarity (where $qw_is_verb = \text{True}$, Benjamini-Hochberg correction, $\alpha = 0.05$).**

Model A	Median A	Model B	Median B	D statistic	p (raw)	p (adjusted)	Significant?
GPT4o (no system prompt)	0.9073	OLMo	0.9034	0.1860	0.0000	0.0000	Yes
Qwen	0.9063	Llama3-70b	0.8705	0.4057	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.9073	Qwen	0.9063	0.1276	0.0000	0.0000	Yes
OLMo	0.9034	Llama3-70b	0.8705	0.3739	0.0000	0.0000	Yes
Qwen (no thinking)	0.9016	Llama3-70b	0.8705	0.3749	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.9073	Llama3-70b	0.8705	0.3719	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.9073	Qwen (no thinking)	0.9016	0.1086	0.0000	0.0000	Yes
Qwen (no thinking)	0.9016	Qwen	0.9063	0.0937	0.0001	0.0001	Yes
OLMo	0.9034	Qwen (no thinking)	0.9016	0.0831	0.0006	0.0006	Yes
OLMo	0.9034	Qwen	0.9063	0.0697	0.0044	0.0044	Yes

Table 25: **doublev-2 Model by Model Wilcoxon Rank-Sum Tests on Cosine Similarity** (where `qw_is_verb = True`, **Benjamini-Hochberg correction**, $\alpha = 0.05$).

Model A	Median A	Model B	Median B	Higher	U statistic	<i>p</i> (raw)	<i>p</i> (adjusted)	Significant?
Qwen (no thinking)	0.2200	Qwen	0.1669	← A	726011.5	0.0000	0.0000	Yes
OLMo	0.1780	Qwen (no thinking)	0.2200	B →	583787.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	Qwen	0.1669	← A	362661.0	0.0000	0.0000	Yes
Qwen	0.1669	Llama3-70b	0.3410	B →	48835.0	0.0000	0.0000	Yes
OLMo	0.1780	Llama3-70b	0.3410	B →	56718.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	OLMo	0.1780	← A	401841.5	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	Llama3-70b	0.3410	B →	31847.5	0.0011	0.0016	Yes
GPT4o (no system prompt)	0.2320	Qwen (no thinking)	0.2200	← A	334382.0	0.1277	0.1597	No
Qwen (no thinking)	0.2200	Llama3-70b	0.3410	B →	71338.5	0.1561	0.1734	No
OLMo	0.1780	Qwen	0.1669	← A	631566.5	0.4339	0.4339	No

Table 26: **doublev-2 Model by Model Two-Sample Kolmogorov-Smirnov Tests on Cosine Similarity** (where `qw_is_verb = True`, **Benjamini-Hochberg correction**, $\alpha = 0.05$).

Model A	Median A	Model B	Median B	D statistic	<i>p</i> (raw)	<i>p</i> (adjusted)	Significant?
Qwen	0.1669	Llama3-70b	0.3410	0.3624	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	Qwen	0.1669	0.1879	0.0000	0.0000	Yes
Qwen (no thinking)	0.2200	Qwen	0.1669	0.1539	0.0000	0.0000	Yes
OLMo	0.1780	Llama3-70b	0.3410	0.3242	0.0000	0.0000	Yes
OLMo	0.1780	Qwen (no thinking)	0.2200	0.1443	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	Qwen (no thinking)	0.2200	0.1712	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	OLMo	0.1780	0.1465	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.2320	Llama3-70b	0.3410	0.2598	0.0000	0.0000	Yes
Qwen (no thinking)	0.2200	Llama3-70b	0.3410	0.2192	0.0000	0.0000	Yes
OLMo	0.1780	Qwen	0.1669	0.0475	0.1564	0.1564	No

Table 27: Within-Model Two-Sample Kolmogorov-Smirnov Tests: doublev-1 vs. doublev-2 on Cosine Similarity (where `qw_is_verb = True`, **Benjamini-Hochberg correction**, $\alpha = 0.05$; composite row uses no correction). If models appropriately follow these rules, cosine similarity will be lower (closer to 0) for doublev-2 which emphasizes irrelevance between the two verbs used in each clause. All models display this behavior, indicated by significant differences in median Cosine Similarity scores.

Model	Med doublev-1	<i>n</i> doublev-1	Med doublev-2	<i>n</i> doublev-2	D stat	<i>p</i> (raw)	<i>p</i> (adj.)	Sig.?
OLMo	0.9034	1285	0.1780	1186	0.9480	0.0000	0.0000	Yes
Qwen (no thinking)	0.9016	1079	0.2200	1178	0.8571	0.0000	0.0000	Yes
Qwen	0.9063	1209	0.1669	1045	0.9310	0.0000	0.0000	Yes
GPT-4o (no system prompt)	0.9073	1085	0.2320	594	0.9113	0.0000	0.0000	Yes
Llama3-70b	0.8705	68	0.3410	131	0.8284	0.0000	0.0000	Yes
<i>All models (pooled)</i>	0.9042	4726	0.1970	4134	0.9075	0.0000	—	Yes

Table 28: Within-Model Two-Sample Wilcoxon Rank-Sum Tests: doublev-1 vs. doublev-2 on Cosine Similarity (where `qw_is_verb = True`, **Benjamini-Hochberg correction**, $\alpha = 0.05$; composite row uses no correction). If models appropriately follow these rules, cosine similarity will be lower (closer to 0) for doublev-2 which emphasizes irrelevance between the two verbs used in each clause. All models display this behavior, indicated by significant directional differences in median Cosine Similarity scores. All models display higher cosine similarity for doublev-1 output.

Model	Med doublev-1	<i>n</i> doublev-1	Med doublev-2	<i>n</i> doublev-2	Higher	U stat	<i>p</i> (raw)	<i>p</i> (adj.)	Sig.?
OLMo	0.9034	1285	0.1780	1186	doublev-1 ↑	1462464.5	0.0000	0.0000	Yes
Qwen	0.9063	1209	0.1669	1045	doublev-1 ↑	1189345.0	0.0000	0.0000	Yes
Qwen (no thinking)	0.9016	1079	0.2200	1178	doublev-1 ↑	1108701.0	0.0000	0.0000	Yes
GPT-4o (no system prompt)	0.9073	1085	0.2320	594	doublev-1 ↑	598113.0	0.0000	0.0000	Yes
Llama3-70b	0.8705	68	0.3410	131	doublev-1 ↑	7508.5	0.0000	0.0000	Yes
<i>All models (pooled)</i>	0.9042	4726	0.1970	4134	doublev-1 ↑	18013125.5	0.0000	—	Yes

Table 29: All 50 cleaned outputs for dan-1 for 5 models, truncated to 40 characters.

	gpt4o_no_sp	llama	olmo	qwen	qwen_no_thinking
1	354057805959:0340867590...	958054750956805670958470956805675095684...	800 5700 9 51... 300 800 43 700 9 38...	35075680995:34086076095...	593754801460378...
2	3540659590:03408606960...	-9906756895:3486790...	cat-five-four-zero-cat-to-be-exclamation...	35756890...37486790...	35756890...37486790...
3	7590280498475007-3zerofcatzero75be7...	907500348475300-90340-980-903-980-975-9...	803:1 8009-345-6-0-009...	354075680995:3408607095...	3507508095:340807095...
4	53407809595:03407680790...	5-8-5-7-5-8-0-7-0-7-5-5-3-4-4-8-0-5-3-4-...	cat-seven-five-zero-zero-four-three-five...	3507568095:3408607095...	35756890...3486790...
5	35-7606-6-0550:34-8606-6-790...	57-4780-47-8-57-4780-57-4780-9-4789-9-7-...	80 65-3-1-4-80-901-48-80-3-4-7-5-90-8-0...	35075680995:3408607095...	3-5-7-6-8-0-5-3-1-3-8-6-7-9-5-4-0
6	708ef5mohor07lveka4thbe50-0hisadcat...	9508-76509-346809-76509...	80 53 460 900 30 480 70 908 35 480 70 9...	35075680995:3408607095...	34075080995:360807095...
7	3575785959:34-8676-76-790...	350-967-83-9-134-87-9-4...	8-0-3-5-6-0-4-0-9-0-1:3-0-8-0-4-0-7-0-9...	3507508095:30840709...	3507508095:30840709...
8	5430768659:034086769...	00...	800-31 34-6-8000-3-5-1...	35075680995:3408607095...	59375680995:3480758095...
9	35768567590:3487980790...	5063-5063-8-6-9-13-8-6-7-9-4...	8000-9-5000-6-8000-9000-4-8000-4-3000-50...	35075680995:3408607095...	Three-five-seven-five-zero-eight-53-zer...
10	7054cat5be50:34cat5bevatbe50...	753-4058-357-0-6-357-0-4-0-4-0-5-8-0-3-0-...	Cat nine be happy8 Pronoun three-five z...	354075680995:03408607095...	597-08-95-190-367-08-95-400...
11	Zero0BeFivZedFemalPantFivCatZero...	5-8-5-6-8-0-7-5-13-8-0-6-7-0-5-8-#...	800-9-835-4000-9-8800-9-3000-700-9...	35075408095:308407095...	35-75-4-8-85-0-3-8-4-7-785-0
12	3575785959:34-8676-76-790...	800-9-835-4000-9-8800-9-3000-700-9...	00-He00-moth0r0-fives-cat00-cat0-eng...	354075680995:3408607095...	59375680995:3480758095...
13	3575785959:34-8676-76-790...	800-9-835-4000-9-8800-9-3000-700-9...	800-53-89 600 8000 51... 3000 800 43 700...	354075680995:3408607095...	59375680995:3480758095...
14	3575785959:34-8676-76-790...	800-53-89 600 8000 51... 3000 800 43 700...	800 53-89 600 8000 51... 3000 800 43 700...	35075680995:3408607095...	3507508095:30840709...
15	5340768659:034086769...	00...	0-He0f-five-moth0r0-eight0-wa0r0e0-5-#...	35075680995:3408607095...	35756890...37486790...
16	7054cat5be50:34cat5bevatbe50...	357458935745893:356489458937458...	8800-9991 33400-70055-800-999-55-#...	3-5-4-0-7-5-6-0-8-0-9-5-1-3-4-0-8-6-0-7-...	3578095:308607095...
17	7054cat5be50:34cat5bevatbe50...	050-908-6-905-904-905-908-0-90-905-904-9...	800-9-800-5-800-6-800-8-800-9: 800-3-800...	354075680995:3408607095...	35075080995:308607095...
18	5340768659:034086769...	-00578601490876601...	08-9-57-0-4-0-1 03-9-08-0-4-7-0-8-9-5-0...	35475680995:0340867095...	35075080995:308607095...
19	3575785959:34-8676-76-790...	05-830-980-91-308-980-876-98...	380-000-9-000-8-600-000-5-000-91-30-400...	354075680995:3408607095...	3-5-7-6-8-9-5-1-0-3-7-8-6-7-9-5-4-0...
20	7054cat5be50:34cat5bevatbe50...	9-5-6-7-8-0-3-4-8-6-7-1-9-5-6-8-3-4-6-7-...	89000351-3400...	35075408095:0308407095...	3507503548095:308307095...
21	7054cat5be50:34cat5bevatbe50...	756567-3-8-7-3-0-9-0-0-736567-3-4-8-7-3-...	3-5-4-6-8-8-0-ko-1e-1 8-4-7-0-ko-5-0...	35075680995:3408607095...	35754859:384759...
22	7054cat5be50:34cat5bevatbe50...	7-56-88-50-40-0-3-5-5-0-0-8-0-0-0-0-0-0...	800-57-600 9000 51... 3000 800 43 700...	35075680995:3408607095...	35756890...37486790...
23	7054cat5be50:34cat5bevatbe50...	#58-557056-309-081588-586-3057-509-08...	8-35-0-3-6-4-9-0-0-1-0-8-3-0-3-4-9-3-5-...	35075680995:3408607095...	Three-five-seven-five-zero-eight-53-zer...
24	7054cat5be50:34cat5bevatbe50...	7057-69-804-3-5-7-8-9-5-769-8-6-3-7-8-...	8-0-3-5-6-0-4-0-9-8-3-4-0-3-7-0-5-9-0-9...	35075680995:308407095...	59375680995:3480758095...
25	7054cat5be50:34cat5bevatbe50...	5-7-6-8-0-9-5-7-6-8-0-3-5-0-9-8-5-7-6-...	800 557... 3000 4800 9000... 31000 47000 9...	35-0-756-8-0-4-951:3-8-86-0-7-0-958...	35-75-8-5-0-5-1-6-6-7-5-9-9...
26	7054cat5be50:34cat5bevatbe50...		8003: 3400...	34075680995:3408607095...	597-08-95-397-08-95-400...
27	7054cat5be50:34cat5bevatbe50...		8-0-9-5-6-3-4-0-8-0-9-0-1-0-3-0-9-4-7-5-0...	3-4-0-7-5-6-0-8-0-9-5-1-0-3-4-0-8-6-0-7...	35075080995:30840709509...
28	7054cat5be50:34cat5bevatbe50...		0-He0f-five-moth0r0-eight0-wa0r0e0-5-#...	35075680995:3408607095...	35075080995:308607095...
29	7054cat5be50:34cat5bevatbe50...		0-moth0r-05-eight-eight-wa0r0e0-5-#...	354075680995:3408607095...	35756890...37486790...
30	7054cat5be50:34cat5bevatbe50...		8-0-53-9-0-51-3-4-8-0-7-5-0-9-0-6-8...	35075408095:0308407095...	34-75-8-9-5-3-8-4-7-985...
31	7054cat5be50:34cat5bevatbe50...		8000319000500400...300090009000...	35075680995:3408607095...	35-75-4-8-95-8-3-8-4-7-98-98...
32	7054cat5be50:34cat5bevatbe50...		8000 570006 9000 51... 3000 800 43 700...	35075680995:3408607095...	3568095:3086075...
33	7054cat5be50:34cat5bevatbe50...		8-0-3-5-6-0-9-0-1-3-1-8-0-3-4-0-7-3-0-9...	35075680995:3408607095...	35075680995:3408607095...
34	7054cat5be50:34cat5bevatbe50...		8-0-3-5-6-0-4-0-8-0-9-0-1-3-0-8-0-4-3-0-7...	34075680995:3408607095...	35-75-4-8-95-8-3-8-4-7-98-98...
35	7054cat5be50:34cat5bevatbe50...		8-0-3-5-6-0-4-0-8-0-9-0-1-3-0-8-0-4-3-0-7...	35075680995:3408607095...	35-75-4-8-95-8-3-8-4-7-98-98...
36	7054cat5be50:34cat5bevatbe50...		800 57-91 3000 48-91 83000 47-8-9000...	34-754-8-9-5-3-8-4-7-985...	35075080995:308607095...
37	7054cat5be50:34cat5bevatbe50...		80000003: 34064000034...	3-4-507-5-6809-5-0-3-408-60709-580...	5937680995:3076809509...
38	7054cat5be50:34cat5bevatbe50...		0-He0f-five-moth0r0-eight0-wa0r0e0-5-#...	35075680995:3408607095...	35075680995:308607095...
39	7054cat5be50:34cat5bevatbe50...		80 53 60-9 3900 84 700-9-#3 80 54 70...	35075680995:3408607095...	35756890...37486790...
40	7054cat5be50:34cat5bevatbe50...		00-cat-to-be-happy! 00-three-four-seven...	35075680995:3408607095...	59375458408-09-54-066-7-03-089-3+06-7+0...
41	7054cat5be50:34cat5bevatbe50...		800 57000 9000 51... 3000 800 43 700...	35075680995:3408607095...	Five-eight-zero-seven-zero-zero-eight-53-zer...
42	7054cat5be50:34cat5bevatbe50...		80345... 300-890-8060...	34075680995:3408607095...	35754859:384759...
43	7054cat5be50:34cat5bevatbe50...		800 53-880 403 700 5-#800 53-880 43 70...	35075680995:3408607095...	35075080995:308607095...
44	7054cat5be50:34cat5bevatbe50...		00-Cat-to-be-five-five-exclamation-mark...	35075680995:3408607095...	59375680995:3480758095...
45	7054cat5be50:34cat5bevatbe50...		800 5700 83 901: 3400 800 700 908...	35075680995:3408607095...	597-068-0-3497-0688...
46	7054cat5be50:34cat5bevatbe50...		800 5700 9 51... 300 800 43 700 9 51...	35075680995:3408607095...	35756890...37486790...
47	7054cat5be50:34cat5bevatbe50...		800-900-5300-600-400-800-900-1: 300-400...	35075680995:3408607095...	3507508095:3407593...
48	7054cat5be50:34cat5bevatbe50...		08 03 06 05 09 01: 03 08 04 07 05 09 08...	35075680995:3408607095...	35754859:384759...
49	7054cat5be50:34cat5bevatbe50...		80 53-91 32 60-80-9-5-#8-0-3-4-0-7-0-9-1-3-...	34075680995:3408607095...	35075080995:308607095...

Table 30: All 50 cleaned outputs for translate-1 for 5 models, truncated to 50 characters.

	gpt4o_no_sp	llama	olmo	qwen	qwen_no_thinking
1	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp-at-mq-mq Marysp. ZkZk Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
2	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
3	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
4	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
5	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
6	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
7	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
8	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
9	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
10	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
11	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
12	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
13	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
14	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
15	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
16	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
17	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
18	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
19	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
20	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
21	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
22	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
23	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
24	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
25	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
26	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
27	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
28	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
29	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
30	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
31	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
32	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
33	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
34	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
35	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
36	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
37	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
38	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
39	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
40	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
41	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
42	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
43	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
44	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
45	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
46	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
47	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
48	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
49	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...
50	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	of Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...	Marysp Marysp Marysp Marysp Marysp Marysp...

Table 31: **translate-1 Model by Model Wilcoxon Rank-Sum Tests on Cosine Similarity Scores (Benjamini-Hochberg correction, $\alpha = 0.05$).**

Model A	Median A	Model B	Median B	Higher	U statistic	p (raw)	p (adjusted)	Significant?
GPT4o (no system prompt)	0.6533	Random string (all)	0.0000	$\leftarrow$ A	2500.0	0.0000	0.0000	Yes
OLMo	0.4027	Random string (all)	0.0000	$\leftarrow$ A	2500.0	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Random string (all)	0.0000	$\leftarrow$ A	2500.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	Random string (select)	0.0000	$\leftarrow$ A	2500.0	0.0000	0.0000	Yes
OLMo	0.4027	Random string (select)	0.0000	$\leftarrow$ A	2499.0	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Random string (select)	0.0000	$\leftarrow$ A	2500.0	0.0000	0.0000	Yes
Qwen	0.9154	Random string (all)	0.0000	$\leftarrow$ A	2475.0	0.0000	0.0000	Yes
Qwen	0.9154	Random string (select)	0.0000	$\leftarrow$ A	2473.0	0.0000	0.0000	Yes
Llama3-70b	0.3924	Random string (all)	0.0000	$\leftarrow$ A	2350.0	0.0000	0.0000	Yes
Llama3-70b	0.3924	Random string (select)	0.0000	$\leftarrow$ A	2338.0	0.0000	0.0000	Yes
OLMo	0.4027	Qwen	0.9154	B $\rightarrow$	84.0	0.0000	0.0000	Yes
Qwen	0.9154	Llama3-70b	0.3924	$\leftarrow$ A	2400.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	OLMo	0.4027	$\leftarrow$ A	2338.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	Llama3-70b	0.3924	$\leftarrow$ A	2285.5	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Llama3-70b	0.3924	$\leftarrow$ A	2200.0	0.0000	0.0000	Yes
OLMo	0.4027	Qwen (no thinking)	0.7341	B $\rightarrow$	307.5	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	Qwen	0.9154	B $\rightarrow$	496.0	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Qwen	0.9154	B $\rightarrow$	758.5	0.0007	0.0008	Yes
Random string (select)	0.0000	Random string (all)	0.0000	tie	1350.0	0.0433	0.0479	Yes
GPT4o (no system prompt)	0.6533	Qwen (no thinking)	0.7341	B $\rightarrow$	1014.5	0.1052	0.1104	No
OLMo	0.4027	Llama3-70b	0.3924	$\leftarrow$ A	1397.0	0.3125	0.3125	No

Table 32: **translate-1 Model by Model Kolmogorov-Smirnov Tests on Cosine Similarity Scores (Benjamini-Hochberg correction, $\alpha = 0.05$).**

Model A	Median A	Model B	Median B	D statistic	p (raw)	p (adjusted)	Significant?
GPT4o (no system prompt)	0.6533	Random string (select)	0.0000	1.0000	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	Random string (all)	0.0000	1.0000	0.0000	0.0000	Yes
OLMo	0.4027	Random string (all)	0.0000	1.0000	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Random string (select)	0.0000	1.0000	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Random string (all)	0.0000	1.0000	0.0000	0.0000	Yes
OLMo	0.4027	Random string (select)	0.0000	0.9800	0.0000	0.0000	Yes
Qwen	0.9154	Random string (select)	0.0000	0.9800	0.0000	0.0000	Yes
Qwen	0.9154	Random string (all)	0.0000	0.9800	0.0000	0.0000	Yes
Llama3-70b	0.3924	Random string (select)	0.0000	0.8800	0.0000	0.0000	Yes
Llama3-70b	0.3924	Random string (all)	0.0000	0.8800	0.0000	0.0000	Yes
OLMo	0.4027	Qwen	0.9154	0.8600	0.0000	0.0000	Yes
Qwen	0.9154	Llama3-70b	0.3924	0.8600	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	OLMo	0.4027	0.8000	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	Llama3-70b	0.3924	0.7400	0.0000	0.0000	Yes
OLMo	0.4027	Qwen (no thinking)	0.7341	0.7000	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Llama3-70b	0.3924	0.6600	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.6533	Qwen	0.9154	0.5600	0.0000	0.0000	Yes
Qwen (no thinking)	0.7341	Qwen	0.9154	0.3600	0.0028	0.0033	Yes
GPT4o (no system prompt)	0.6533	Qwen (no thinking)	0.7341	0.3400	0.0058	0.0065	Yes
OLMo	0.4027	Llama3-70b	0.3924	0.2400	0.1124	0.1180	No
Random string (select)	0.0000	Random string (all)	0.0000	0.0800	0.9977	0.9977	No

Table 33: **dan-1 Model by Model Wilcoxon Rank-Sum Tests on Cosine Similarity Scores (Benjamini-Hochberg correction, $\alpha = 0.05$).**

Model A	Median A	Model B	Median B	Higher	U statistic	p (raw)	p (adjusted)	Significant?
Qwen	0.6455	Random string (all)	0.0000	$\leftarrow$ A	2425.0	0.0000	0.0000	Yes
Qwen	0.6455	Random string (select)	0.0000	$\leftarrow$ A	2394.5	0.0000	0.0000	Yes
OLMo	0.0000	Qwen	0.6455	B $\rightarrow$	120.0	0.0000	0.0000	Yes
Qwen	0.6455	Llama3-70b	0.0000	$\leftarrow$ A	2365.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	Qwen	0.6455	B $\rightarrow$	178.0	0.0000	0.0000	Yes
Qwen (no thinking)	0.0545	Qwen	0.6455	B $\rightarrow$	249.5	0.0000	0.0000	Yes
Qwen (no thinking)	0.0545	Random string (all)	0.0000	$\leftarrow$ A	2100.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	Random string (all)	0.0000	$\leftarrow$ A	2075.0	0.0000	0.0000	Yes
Qwen (no thinking)	0.0545	Random string (select)	0.0000	$\leftarrow$ A	1966.0	0.0000	0.0000	Yes
Llama3-70b	0.0000	Random string (all)	0.0000	tie	1750.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	Random string (select)	0.0000	$\leftarrow$ A	1853.5	0.0000	0.0000	Yes
OLMo	0.0000	Qwen (no thinking)	0.0545	B $\rightarrow$	645.0	0.0000	0.0000	Yes
OLMo	0.0000	Random string (all)	0.0000	tie	1650.0	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	OLMo	0.0000	$\leftarrow$ A	1779.0	0.0001	0.0001	Yes
Qwen (no thinking)	0.0545	Llama3-70b	0.0000	$\leftarrow$ A	1784.5	0.0001	0.0001	Yes
Random string (select)	0.0000	Random string (all)	0.0000	tie	1575.0	0.0001	0.0002	Yes
GPT4o (no system prompt)	0.0615	Llama3-70b	0.0000	$\leftarrow$ A	1629.0	0.0058	0.0072	Yes
GPT4o (no system prompt)	0.0615	Qwen (no thinking)	0.0545	$\leftarrow$ A	955.5	0.0390	0.0455	Yes
Llama3-70b	0.0000	Random string (select)	0.0000	tie	1467.0	0.0743	0.0821	No
OLMo	0.0000	Llama3-70b	0.0000	tie	1131.5	0.3436	0.3608	No
OLMo	0.0000	Random string (select)	0.0000	tie	1338.0	0.4516	0.4516	No

Table 34: **dan-1 Model by Model Kolmogorov-Smirnov Tests on Cosine Similarity Scores (Benjamini-Hochberg correction, $\alpha = 0.05$).**

Model A	Median A	Model B	Median B	D statistic	p (raw)	p (adjusted)	Significant?
Qwen	0.6455	Random string (all)	0.0000	0.9400	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	Qwen	0.6455	0.9000	0.0000	0.0000	Yes
OLMo	0.0000	Qwen	0.6455	0.9000	0.0000	0.0000	Yes
Qwen	0.6455	Random string (select)	0.0000	0.9000	0.0000	0.0000	Yes
Qwen	0.6455	Llama3-70b	0.0000	0.8800	0.0000	0.0000	Yes
Qwen (no thinking)	0.0545	Qwen	0.6455	0.7400	0.0000	0.0000	Yes
Qwen (no thinking)	0.0545	Random string (all)	0.0000	0.6800	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	Random string (all)	0.0000	0.6600	0.0000	0.0000	Yes
Qwen (no thinking)	0.0545	Random string (select)	0.0000	0.6200	0.0000	0.0000	Yes
GPT4o (no system prompt)	0.0615	Random string (select)	0.0000	0.5000	0.0000	0.0000	Yes
OLMo	0.0000	Qwen (no thinking)	0.0545	0.4600	0.0000	0.0001	Yes
GPT4o (no system prompt)	0.0615	OLMo	0.0000	0.4200	0.0002	0.0004	Yes
Qwen (no thinking)	0.0545	Llama3-70b	0.0000	0.4200	0.0002	0.0004	Yes
Llama3-70b	0.0000	Random string (all)	0.0000	0.4000	0.0006	0.0009	Yes
GPT4o (no system prompt)	0.0615	Qwen (no thinking)	0.0545	0.3600	0.0028	0.0040	Yes
GPT4o (no system prompt)	0.0615	Llama3-70b	0.0000	0.3400	0.0058	0.0077	Yes
OLMo	0.0000	Random string (all)	0.0000	0.3200	0.0115	0.0142	Yes
Random string (select)	0.0000	Random string (all)	0.0000	0.2600	0.0678	0.0791	No
Llama3-70b	0.0000	Random string (select)	0.0000	0.2200	0.1786	0.1974	No
OLMo	0.0000	Random string (select)	0.0000	0.1600	0.5487	0.5761	No
OLMo	0.0000	Llama3-70b	0.0000	0.1000	0.9667	0.9667	No

Composite scores		Qwen-3 (T)	GPT-4	Qwen-3 (No T)	OLMo-2	Llama-3
		0.960	0.697	0.677	0.723	0.459
	Family-weighted mean	0.975	0.655	0.661	0.654	0.437

Table 35: Normalized performance scores per model and metric (all scores oriented: higher = better, top-normalized per row). Metrics are grouped into four families, each contributing equally ($w = 1/4$) to the family-weighted composite. Bigram metrics use cleaned outputs where available (GPT-4, OLMo-2, Qwen3 Thinking), uncleaned otherwise. Levenshtein distances are log-normalized prior to top-normalization to reduce compression from the Llama-3 dan-1 outlier. Metrics are grouped into four rule families: the *bigram family* (comparison-1, comparison-2, sidequest-1), the *dan-1 family*, the *translate-1 family*, and the *doublev family* (doublev-1, doublev-2), each contributing equally ($w = 1/4$) to the family-weighted composite. Elsewhere, we discuss dan-1 and translate-1 as both being part of the translation rules family, but performance here is anchored to each independently, rather than relative (as it is with the story rules and doublev rules), which is why we count them separately. However, again, reasonable variations are totally possible.

Table 36: Best-fitting linear combinations of linguistic rule features for predicting mean model performance across rules, by number of features (exhaustive subset search, $n = 7$ rules). All coefficients are negative, indicating each constraint type reduces performance. Results are exploratory only.

n	Coefficients	Intercept	R^2
1	uses_no_typical_english = -0.6096	0.908000	0.898000
2	uses_no_typical_english = -0.5795, morphology_constraint = -0.1802	0.998000	0.975000
3	uses_no_typical_english = -0.5427, morphology_constraint = -0.1557, bizarre_required_output_other_than_rare_bigrams = -0.1227	0.986000	0.988000

Table 37: Top 10 feature subsets by R^2 from exhaustive subset search ($n = 7$ rules, subsets of size 1-3). uses_no_typical_english is the strongest single predictor. Results are exploratory only.

n	Features	Coefficients	Intercept	R^2
3	uses_no_typical_english, morphology_constraint, bizarre_required_output_other_than_rare_bigrams	uses_no_typical_english = -0.5427, morphology_constraint = -0.1557, bizarre_required_output_other_than_rare_bigrams = -0.1227	0.986000	0.988000
3	no_prompt_answer_overlap, lexical_restriction, morphology_constraint	no_prompt_answer_overlap = -0.1227, lexical_restriction = -0.5427, morphology_constraint = -0.1557	0.986000	0.988000
3	no_prompt_answer_overlap, uses_no_typical_english, morphology_constraint	no_prompt_answer_overlap = -0.1227, uses_no_typical_english = -0.5427, morphology_constraint = -0.1557	0.986000	0.988000
3	lexical_restriction, morphology_constraint, bizarre_required_output_other_than_rare_bigrams	lexical_restriction = -0.5427, morphology_constraint = -0.1557, bizarre_required_output_other_than_rare_bigrams = -0.1227	0.986000	0.988000
3	lexical_restriction, morphology_constraint, rare_bigrams_required_output	lexical_restriction = -0.5872, morphology_constraint = -0.2111, rare_bigrams_required_output = -0.0769	0.975000	0.986000
3	uses_no_typical_english, morphology_constraint, rare_bigrams_required_output	uses_no_typical_english = -0.5872, morphology_constraint = -0.2111, rare_bigrams_required_output = -0.0769	0.975000	0.986000
3	uses_no_typical_english, decoding_possible_is_at_least_partially_enforced, semantic_constraint	uses_no_typical_english = -0.7199, decoding_possible_is_at_least_partially_enforced = -0.2313, semantic_constraint = -0.1811	1.185000	0.983000
3	lexical_restriction, decoding_possible_is_at_least_partially_enforced, semantic_constraint	lexical_restriction = -0.7199, decoding_possible_is_at_least_partially_enforced = -0.2313, semantic_constraint = -0.1811	1.185000	0.983000
3	lexical_restriction, morphology_constraint, rare_bigrams_in_prompt	lexical_restriction = -0.5744, morphology_constraint = -0.1853, rare_bigrams_in_prompt = -0.0515	0.962000	0.980000
3	uses_no_typical_english, morphology_constraint, rare_bigrams_in_prompt	uses_no_typical_english = -0.5744, morphology_constraint = -0.1853, rare_bigrams_in_prompt = -0.0515	0.962000	0.980000