Highlights

Investigating the Influence of Prompt and Response Languages on LLM Content Generation

Thi Thanh Nhan Nguyen*, Mai Khoi Tieu*[1], Michael A. Riegler, Pål Halvorsen, Thu Nguyen

- **Prompt Language Modulates Output Length:** The language of the prompt exerts a significant structural influence on the length of the generated response, independently of the specified output language.
- **Cross-Lingual Compression Effects:** Prompting in English for a Norwegian response yields the most pronounced text compression (a 52% reduction in word count); however, this phenomenon is partially attributable to tokenizer fertility disparities rather than purely generative compression.
- **Semantic Fidelity via Conceptual Paraphrasing:** Despite substantial reductions in output length, semantic integrity remains highly stable across conditions. Models utilize conceptual paraphrasing rather than literal translation, maintaining high cross-lingual embedding similarity despite low surface-level lexical overlap.
- **Inverse Relationship with Information Density:** The structural compression observed in cross-lingual tasks corresponds to a near-perfect monotonic increase in information density, communicating equivalent semantic payloads with greater lexical efficiency.
- **Architectural Heterogeneity:** The magnitude of this translation compression effect varies significantly across different LLM architectures, with highly verbose models exhibiting the most pronounced deviations.

[1]*denotes equal contribution

Investigating the Influence of Prompt and Response Languages on LLM Content Generation

Thi Thanh Nhan Nguyen*[c], Mai Khoi Tieu*[1a], Michael A. Riegler[b], Pål Halvorsen[b], Thu Nguyen[d]

[a] *Norwegian University of Science and Technology (NTNU), Trondheim, Trøndelag, Norway*
[b] *SimulaMet, Oslo, Norway*
[c] *Université de Technologie de Compiègne, Compiègne, Hauts-de-France, France*
[d] *Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam*

Abstract

This study investigates how the choice of prompt and response language shapes the generative behavior of Large Language Models (LLMs), extending beyond surface-level linguistic variation to examine structural and semantic dimensions of output. Using five models—DeepSeek V3, GPT-4o, Phi-4-multimodal, Claude 3.5 Haiku, and Gemini 2.5 Pro—we evaluated responses to 68 non-translation questions spanning ethics, culture, health, and social domains under four conditions defined by prompt×response language: English→English, English→Norwegian, Norwegian→Norwegian, and Norwegian→English. After excluding items refused in any condition, our balanced dataset comprises 1,348 responses (337 per condition). Length differences were quantified with Cohen's d; semantic fidelity with LaBSE cosine similarity (chunk-averaged to avoid truncation); and cross-lingual keyword overlap with both raw Jaccard and a LaBSE-based soft Jaccard that maps Norwegian and English keywords into a shared embedding space. We identify a robust prompt-language effect on response length that is visible in the confound-free within-response-language contrasts: holding the response language fixed at English, a Norwegian prompt shortens responses by 37% on average (*ee* 427.8 vs. *ne* 268.2 words, $d \approx 0.94$); holding the response language fixed at Norwegian, an English prompt shortens responses by 41%. The largest single-cell, cross-lingual *ee–en*, is 52% shorter in words ($d \approx 1.32$), but

[1] *denotes equal contribution

only 25% shorter in tokens—so a portion of that headline figure reflects Norwegian's higher subword-tokenizer fertility rather than pure model compression. Cosine similarity remains high across all conditions (≈ 0.83, unchanged under chunk-averaged embedding), while soft Jaccard shows substantial concept overlap (≈ 0.53) that is invisible to raw-string Jaccard (≈ 0.02)—direct evidence of conceptual paraphrasing rather than literal translation. Information density is a near-perfect monotonic inverse of length and is reported as a companion, not an independent effect. Per-model Cohen's d ranges widely (Phi-4 ≈ 0.78 to GPT-4o ≈ 4.42), so the pooled effect is heterogeneous across models. Prompt language is not a neutral parameter: it shapes output length and lexical realization, which matters for multilingual prompt engineering.

Keywords: Large Language Models, Cross-lingual Prompting, Translation Compression, Cultural Framing, Multilingual NLP, Prompt Engineering, Information Density

---

1. Introduction

The deployment of Large Language Models (LLMs) in multilingual environments raises critical questions regarding how language selection influences the nature of generated content. While LLMs are capable of processing queries in various languages, it remains unclear whether the choice of language merely affects syntax or if it fundamentally alters the underlying information retrieval and presentation strategies. This study analyzes these behaviors by examining how models respond when the language of the prompt differs from the required language of the response.

To investigate this, we structured an experiment using five models—DeepSeek V3, GPT-4o, Phi-4-multimodal, Claude 3.5 Haiku, and Gemini 2.5 Pro. The experimental design involved asking identical questions across four topics (ethics, culture, health, and social) using four linguistic permutations: prompting in English or Norwegian, and requesting responses in either language.

We then analyzed the responses along four axes. First, we measured the magnitude of the "translation compression" effect with Cohen's d, distinguishing systematic shifts in response length from random variation. Second, to check that this brevity did not come at the cost of meaning, we assessed content fidelity with semantic similarity, encoding responses with LaBSE embeddings and computing their cosine similarity. Third, we computed Infor-

mation Density to test whether shorter outputs were more information-dense or merely truncated. Finally, we measured Lexical Divergence with the Jaccard index to distinguish literal translation from conceptual paraphrasing. By applying these metrics, this paper provides quantitative evidence on how language constraints act as a mechanism for "translation compression" and cultural framing within generative AI.

In short, the contributions of this work are as follows. First, we provide a systematic evaluation of how input and output language constraints influence the generation behavior of modern LLMs, moving beyond performance benchmarks to structural and semantic shifts. Second, we introduce an analytical framework combining Cohen's d, LaBSE-based semantic similarity, information density, and Jaccard lexical overlap to quantify the trade-off between rhetorical elaboration and informational efficiency. Third, using a 2×2 design, we show that "translation compression" is a prompt×response interaction rather than a uniform cross-lingual effect: the most compressed condition (English prompt → Norwegian response) is about 52% shorter and roughly 2× denser than English-to-English generation. Finally, we document a dissociation between semantic content—which is quantified and largely preserved (≈ 0.83 similarity)—and cultural presentation, which we characterize qualitatively as shifting with language choice.

The remainder of this paper is organized as follows. Section 2 reviews related work. Sections 3 and 4 detail the experimental design, defining the four prompt×response language conditions and the evaluation metrics: Cohen's d, a mixed-effects length model, LaBSE-based cosine similarity (with chunked robustness), information density, and both raw and LaBSE soft Jaccard for keyword overlap. Section 5 reports the quantitative results, including per-model heterogeneity, char- and token-based robustness checks for the length effect, and the semantic–lexical dissociation. Section 6 discusses candidate mechanisms (tokenizer fertility, RLHF verbosity bias, pretraining-corpus asymmetry, safety hedging) and limitations. Section 7 concludes with recommendations for multilingual prompt engineering.

2. Related Works

The study of multilingual Large Language Models (LLMs) has evolved from simple performance benchmarking to complex investigations into how linguistic constraints influence cognitive reasoning and cultural alignment. This section reviews three key areas of related literature to our study: the

dynamics of cross-lingual prompting, relevant cultural aspects and the evolution of semantic evaluation metrics.

2.1. Cross-Lingual Prompting and Information Compression

Prior research has extensively documented the "English-centric" bias of modern LLMs, where models demonstrate superior reasoning capabilities in English due to the predominance of English data in pre-training corpora [1, 2]. Consequently, cross-lingual prompting (prompting in a high-resource language like English for output in a low-resource language) has been proposed as a strategy to unlock better reasoning capabilities in multilingual tasks [3, 4]. However, the structural impact of this language switching on *response length* has received less attention.

While recent studies on "prompt bloat" suggest that excessive context can degrade reasoning [5, 6], our observation of a "Translation Compression" effect aligns more closely with findings on semantic compression. Gilbert et al. [7] demonstrated that LLMs can effectively compress text while preserving semantic essence, suggesting an inherent capability to distill information when constrained. However, they studied compression as an explicit task, whereas we observe it arising on its own during cross-lingual generation. Having to answer in another language appears to act as a constraint: the model delivers the core content more concisely and drops much of the "conversational filler" typical of English-to-English responses.

2.2. Linguistic Determinism and Cultural Framing

The hypothesis that the language of the prompt influences the cultural values and generation style of the output—a computational parallel to the Sapir-Whorf hypothesis—has gained empirical support [8, 9]. Li et al. [10] introduced "CultureLLM" to address the Western bias inherent in English-prompted generations, noting that standard alignment techniques often fail to capture local cultural nuances. Similarly, varying the prompt language has been shown to shift model outputs between "independent" (Western) and "interdependent" (Eastern) social orientations, effectively toggling the model's active cultural framework [8].

Our results are consistent with this line of work in the length and lexical-realization channels: outputs shift systematically with the language of instruction. We do not, however, quantify cultural framing directly in this study—we return to that limitation in Section 6.

2.3. Semantic Evaluation in Multilingual Contexts

Evaluating the quality of cross-lingual generation requires metrics that transcend simple lexical overlap. Traditional n-gram metrics like BLEU [11] or ROUGE [12] have proven insufficient for capturing semantic fidelity in open-ended generation, particularly when the output length varies significantly [13]. The shift toward embedding-based metrics, such as BERTScore [14] and LaBSE (Language-agnostic BERT Sentence Embeddings) [15], allows for the quantification of meaning preservation across languages regardless of syntactic structure.

Our methodology adopts these advanced metrics but integrates them with *Information Density* analysis. While high semantic similarity (via LaBSE) confirms content preservation, it does not measure communicative efficiency. By combining semantic stability with density metrics, we provide a more granular view of how LLMs trade off rhetorical elaboration for informational compactness during cross-lingual tasks.

3. Methodology

This section describes the question set, the four language conditions, the models queried, and how we handled refusals and translation-task items.

Question set. Questions span four topics: ethics, culture, health, and social. The 20 culture questions were generated by Tulu AI; the 20 ethics and 20 social questions were generated by Microsoft Copilot; the 10 health questions were written manually by the authors to include a realistic mix of first-person clinical queries. Provenance is inconsistent across topics; we treat this as a limitation (Section 6) rather than a controlled variable, and have not tested whether it systematically affects the reported metrics. The health topic is smaller (10 vs. 20) because two of the ten items (Q9 and Q10) are translation tasks ("Please translate this medical report excerpt into Norwegian...") that make the response language over-determined and were refused by every model in both Norwegian-prompt conditions; we exclude them from the length and density analyses (see below) and report them separately in Appendix C. This leaves 68 non-translation questions per model. English question lists are in Appendix A; Norwegian translations were produced in Bokmål by a native Norwegian speaker on the author team and spot-checked against the English source for semantic equivalence. The verbatim prompt templates used for each of the four conditions are given in Appendix B.

Language conditions. Each question is issued under four conditions defined by prompt language × required response language:

- *ee*: English prompt → English response
- *en*: English prompt → Norwegian response
- *nn*: Norwegian prompt → Norwegian response
- *ne*: Norwegian prompt → English response

The short labels (ee, en, nn, ne) are internal shorthand; we italicize them throughout to distinguish them from ISO 639-1 language codes.

Models and decoding. We queried five models: DeepSeek V3, GPT-4o (2024-08-06 Azure snapshot), Phi-4-multimodal-instruct, Claude 3.5 Haiku (claude-3-5-haiku-latest), and Gemini 2.5 Pro (gemini-2.5-pro-preview-03-25). Each condition is queried once per question (no sampling). Decoding settings were held at the API defaults on Azure for GPT-4o, DeepSeek, and Phi-4 (`max_tokens`= 1000); Claude used `max_tokens`= 2000, `temperature`= 1; Gemini used the SDK defaults. Because generation was capped, we audited the outputs for truncation and report the results in Appendix D; a non-trivial fraction of responses—concentrated in Norwegian-output conditions rather than long English ones—terminate mid-sentence, which we discuss as a limitation in Section 6.

Refusals and balanced design. Three items were declined by content filters on some models: ethics Q13 (*"Is it ethical to genetically modify humans?"*) was refused by GPT-4o, DeepSeek, and Phi-4 in the Norwegian-prompt conditions, and health Q9/Q10 (the translation-task items) were refused by every model in both Norwegian-prompt conditions. For all analyses below, we drop any (model, question) pair for which *any* of the four conditions is missing, giving a balanced dataset of 337 questions × 4 conditions = 1,348 responses. Per-model refusal counts are in Appendix D.

4. Evaluation Strategies

4.1. Evaluating Response Length

We quantify length effects with three complementary measures: word count, character count (both after light preprocessing that strips Markdown emphasis markers), and token count under a single reference tokenizer

(`tiktoken cl100k_base`). Reporting all three matters because Norwegian is a compounding language (e.g., *informasjonstetthet* = "information density"—one word vs. two), and English-centric subword tokenizers have higher fertility on Norwegian, so the same content can look shorter in words but longer in tokens.

Effect sizes are reported as Cohen's d [16]: $d = (\bar{x}_1 - \bar{x}_2)/s_{\text{pooled}}$, with s_{pooled} the weighted pooled standard deviation. We interpret $|d| \geq 0.8$ as a large effect, $|d| \geq 0.5$ as medium, and $|d| \geq 0.2$ as small [16]. Because the same 68 questions are asked of every model under every condition, the 1,348 responses are not independent; a fixed-effects ANOVA would over-estimate significance. We therefore fit a mixed-effects model with random intercepts for question:

$$\text{length}_{ijk} \sim \text{prompt}_i \times \text{response}_j + \text{model}_k + (1 \mid \text{question}).$$

Model is treated as a fixed factor (only five levels). We report the interaction Wald z-statistic and per-model d values so that per-model heterogeneity is visible rather than hidden in a pooled estimate.

4.2. Semantic Similarity Analysis

To verify that content is preserved across languages despite length differences, we encode each response with LaBSE (Language-agnostic BERT Sentence Embeddings) [15] into a shared multilingual space and compute cosine similarity [17] between response vectors. Because LaBSE truncates inputs beyond its default token window, long English responses (*ee* frequently exceeds 500 words) could otherwise be embedded from only a prefix. As a robustness check, we also compute a *chunk-averaged* variant: each response is split into 250-word windows, each window is embedded independently, and the L2-normalized mean of the window embeddings is used in place of the naive embedding. We report both.

For each question with all four conditions present, we take the mean of the six pairwise cosine similarities among $\{ee, en, nn, ne\}$ as the per-question semantic similarity. We do not impose a numerical threshold; instead we report the full distribution and, where useful, note the fraction of pairs exceeding conventional levels (e.g., ≥ 0.80).

4.3. Information Density

As a companion length–vocabulary summary, we compute Information Density: the ratio of the number of TF-IDF-extracted keywords [18, 19] to

the total word count, expressed per 100 words. Keywords are capped at 30 per response and extracted with sentence-level TF-IDF (`ngram_range`$=(1,2)$, no stop-word filter for multilingual support). The density ρ is

$$\rho = \frac{N_{\text{keywords}}}{N_{\text{words}}} \times 100.$$

Because the keyword count saturates near its cap for all but the shortest responses, ρ is a near-perfect monotonic inverse of length in our data (Spearman $\rho_{\text{sp}} \approx -1.0$). We therefore treat it as a descriptive companion to the word-count analysis rather than an independent measure of communicative efficiency; the reader should read a higher ρ as "shorter for the same information" rather than as an independent quality signal.

4.4. Cross-lingual Keyword Overlap

To characterize vocabulary reuse between response variants we report two Jaccard-style scores [20]. The *raw* Jaccard is the classical set-overlap between the top-k TF-IDF keywords (K_A, K_B) of two responses:

$$J_{\text{raw}}(K_A, K_B) = \frac{|K_A \cap K_B|}{|K_A \cup K_B|}.$$

When A and B are in different languages, this measure is dominated by the fact that Norwegian and English use different word forms (*helse* vs. *health*, *informasjonstetthet* vs. *information density*), so a near-zero score is guaranteed by construction and does not diagnose paraphrasing. We therefore also report a *soft* Jaccard that first maps each keyword through the LaBSE embedding space and counts a keyword as matched if it has at least one nearest neighbour on the other side above a cosine threshold $\tau = 0.70$:

$$J_{\text{soft}}(K_A, K_B) = \frac{M}{|K_A| + |K_B| - M}, \quad M = \tfrac{1}{2}\big(m_A + m_B\big),$$

where

$$m_A = |\{a \in K_A : \max_b \cos(\phi(a), \phi(b)) \geq \tau\}|,$$
$$m_B = |\{b \in K_B : \max_a \cos(\phi(a), \phi(b)) \geq \tau\}|.$$

At $\tau = 1$ this reduces to the classical Jaccard $|K_A \cap K_B|/|K_A \cup K_B|$. Here $\phi(\cdot)$ is the LaBSE encoder. A low J_{raw} paired with a high J_{soft} is what we would expect from conceptual paraphrasing across languages: same concepts, different surface forms. The raw–soft gap makes the paraphrasing claim testable rather than tautological.

5. Results and Major Remarks

5.1. Word count analysis

Across topics and models, response length follows a consistent ordering: the *ee* condition is generally longest, *en* is generally shortest, and *nn*/*ne* fall between them. Figure 1 shows this pattern by model and domain on the balanced 337-question dataset, with 95% bootstrap confidence intervals.

Translation compression in three length units. On the balanced dataset, the mean *ee* → *en* reduction is 52% in words (427.8 → 203.8), 54% in characters (3072.9 → 1419.6), and only 25% in tokens (614.1 → 459.9) under the `cl100k_base` tokenizer. Strikingly, *nn* responses are 31% *longer* in tokens than *ee* responses (804.3 vs. 614.1) despite being 20% shorter in words. The word/character metrics track one another closely, but the token gap is much smaller, reflecting Norwegian's higher subword fertility on English-centric tokenizers. We therefore treat the word-based "compression" figure as an upper bound on how much shorter Norwegian responses truly are in the model's own units.

Overall, the *ee* → *en* length gap corresponds to a large effect size ($d \approx 1.32$ on the balanced dataset), but this pooled value hides considerable per-model heterogeneity (Table 1): Phi-4 sits just above the medium/large threshold ($d = 0.78$), Claude is large ($d = 1.52$), and GPT-4o is exceptionally large ($d = 4.42$). The pooled d is driven primarily by the more verbose models.

Table 1: Per-model Cohen's d for the *ee*–*en* length gap on the 68 non-translation questions. $n = 68$ per condition per model.

Model	Mean *ee*	Mean *en*	d	Interpretation
GPT-4o	542.5	157.1	4.42	very large
DeepSeek-V3	440.9	175.3	3.11	very large
Gemini 2.5 Pro	661.3	345.8	2.18	very large
Claude 3.5 Haiku	170.9	126.2	1.52	large
Phi-4-multimodal-instruct	324.0	213.5	0.78	medium-to-large

Figure 2 shows the aggregate distributional pattern by condition.

Norwegian-output conditions (*en*, *nn*) are generally more concise (in words) than English-output conditions (*ee*, *ne*).

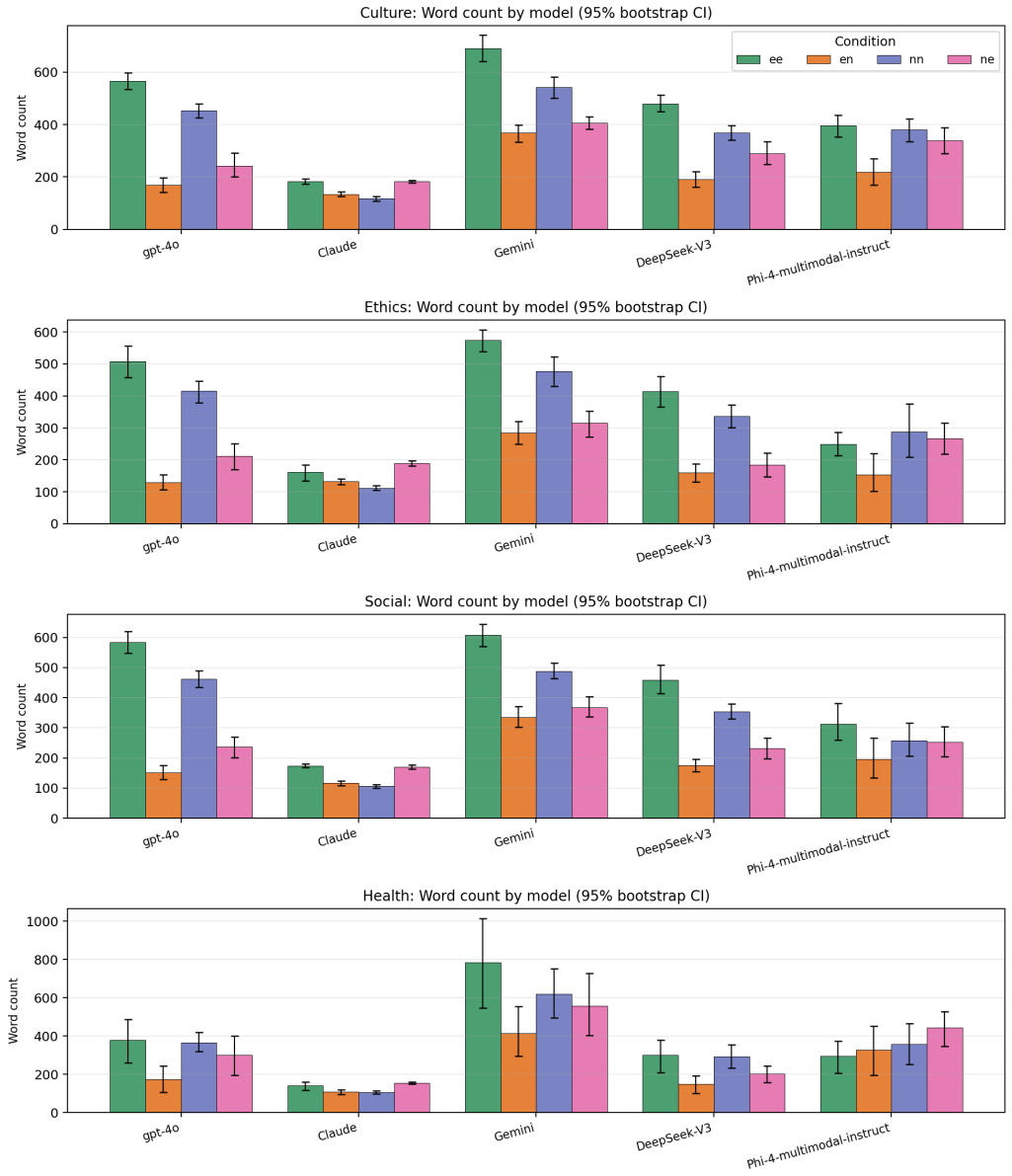

Figure 1: Mean word count (95% bootstrap CI) by model and domain on the balanced dataset ($n = 1{,}348$).

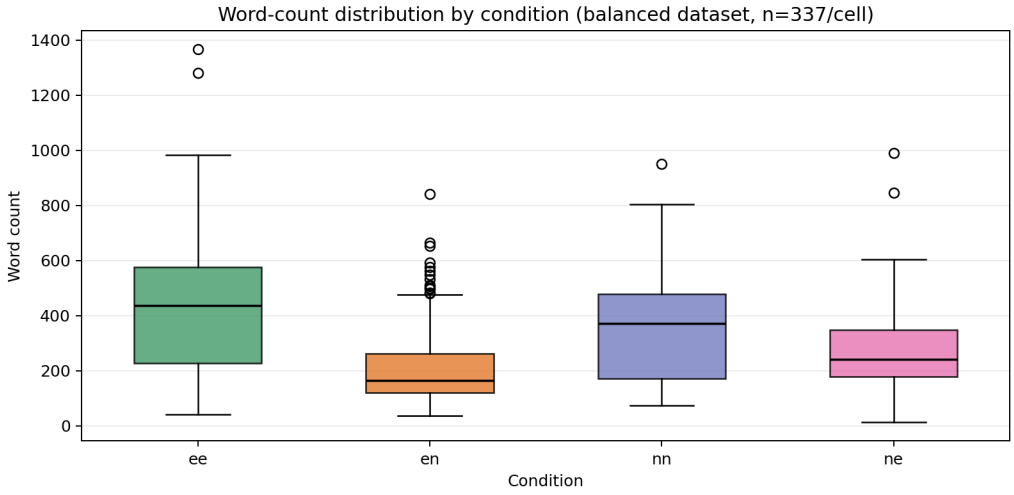

Figure 2: Word-count distributions by condition on the balanced dataset ($n = 337$ per cell).

5.2. Decomposing the Prompt and Response Language Effects

Because the four conditions form a 2×2 design (prompt language $\times$ response language), we can separate the contribution of the *prompt* language from that of the *response* language rather than treating "cross-lingual" as a single factor. Table 2 reports mean word counts for the four cells on the balanced dataset.

Table 2: Mean word count by prompt $\times$ response language on the balanced dataset ($n = 337$ per cell across five models and four domains).

	Response: English	Response: Norwegian
Prompt: English	427.8 (*ee*)	203.8 (*en*)
Prompt: Norwegian	268.2 (*ne*)	344.6 (*nn*)

Two patterns emerge. First, the effect of response language on length *reverses* depending on the prompt language: under an English prompt an English response is on average about 224 words longer than a Norwegian one (*ee* vs. *en*), whereas under a Norwegian prompt an English response is about 76 words *shorter* than a Norwegian one (*ne* vs. *nn*). Fitting the mixed-effects model words $\sim$ prompt $\times$ response + model + (1 | question) on the analyzed

set ($n = 1{,}354$) yields a large and highly significant prompt×response interaction (coefficient +300.7 words, $z = 25.62$, $p < 10^{-100}$), with a substantial question-level variance component (Group Var $\approx 2{,}483$) that a fixed-effects ANOVA would incorrectly pool into residual variance and thereby overstate the significance of. The main effects go in the expected directions (Norwegian prompt shortens: -159.7, $z = -19.24$; Norwegian response shortens: -224.3, $z = -27.09$). Refitting on $\log(1 + \text{words})$ as a robustness check for the right-skewed distribution yields the same qualitative and quantitative story (interaction $z = 23.01$; back-transformed percentage reductions within 1–2 pp of the direct-means values in Table 2); see Appendix F.

Second, the single most compressed condition is *en*, not cross-lingual generation in general. We therefore interpret "translation compression" as a prompt×response interaction rather than a uniform property of cross-lingual prompting: the prompt language modulates how strongly the response language affects output length.

5.3. Semantic Similarity and Cross-lingual Keyword Overlap

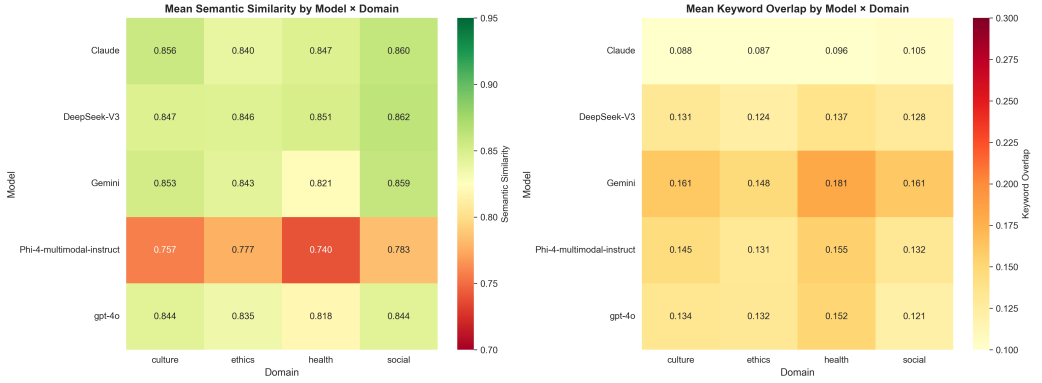

Figure 3: Mean semantic similarity (left) and mean raw-Jaccard keyword overlap (right) by model and domain, averaged over the six pairwise combinations of $\{ee, en, nn, ne\}$.

Semantic stability with genuine paraphrasing. Figure 3 shows that mean LaBSE cosine similarity remains high across most model–domain combinations (roughly 0.82–0.86, with lower values for Phi-4). Chunking long English responses at 250-word windows and re-embedding gives a near-identical mean (0.827 vs. 0.832 naive; per-question $\Delta = -0.005$ on the balanced dataset), so the LaBSE truncation concern is negligible in practice.

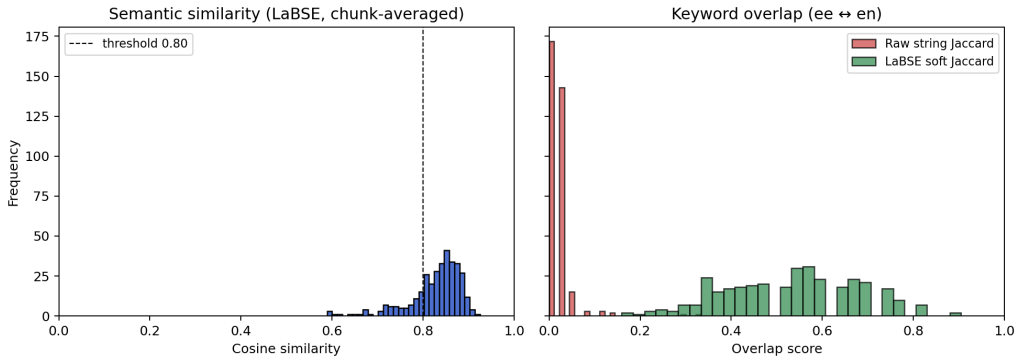

Figure 4: Left: distribution of chunk-averaged LaBSE cosine similarity across questions. Right: distribution of ee↔en keyword overlap under raw Jaccard (red) and LaBSE soft Jaccard (green). Raw Jaccard collapses near zero because English and Norwegian keyword surface forms rarely match; soft Jaccard reveals substantial concept-level overlap (≈ 0.53 mean).

The lexical picture is only meaningful once we account for language. Restricting to the cross-lingual *ee* $\leftrightarrow$ *en* pair on the analyzed set ($n = 340$ question-pairs), the mean raw Jaccard is 0.017: because Norwegian and English keyword surface forms overlap almost never, this number essentially measures "the two responses are in different languages" and cannot diagnose paraphrasing. Under LaBSE soft Jaccard the same pairs score 0.530 on average—a +0.512 lift. This raw–soft gap (Figure 4, right panel) provides the actual evidence for the paraphrasing interpretation: the models are reusing the same concepts, expressed in different surface vocabulary, and the reuse only shows up once keywords are mapped through a shared multilingual embedding.

5.4. Information Density (Length–Vocabulary Companion)

Information density is highest for *en* across domains and most models (Figure 5). The magnitude of the gap varies by model.

As shown in Figure 6, on the balanced dataset the mean information density by condition is highest for *en* (20.16), followed by *ne* (14.30), *nn* (12.60), and *ee* (9.67). Because the number of extracted keywords saturates near its cap for all but the shortest responses, this metric is a near-perfect monotonic inverse of length in our data (Spearman $\rho \approx -1.0$; Pearson $r \approx 0.98$ between density and $1/N_{\text{words}}$). It therefore re-expresses the length result from a complementary angle rather than constituting independent evidence

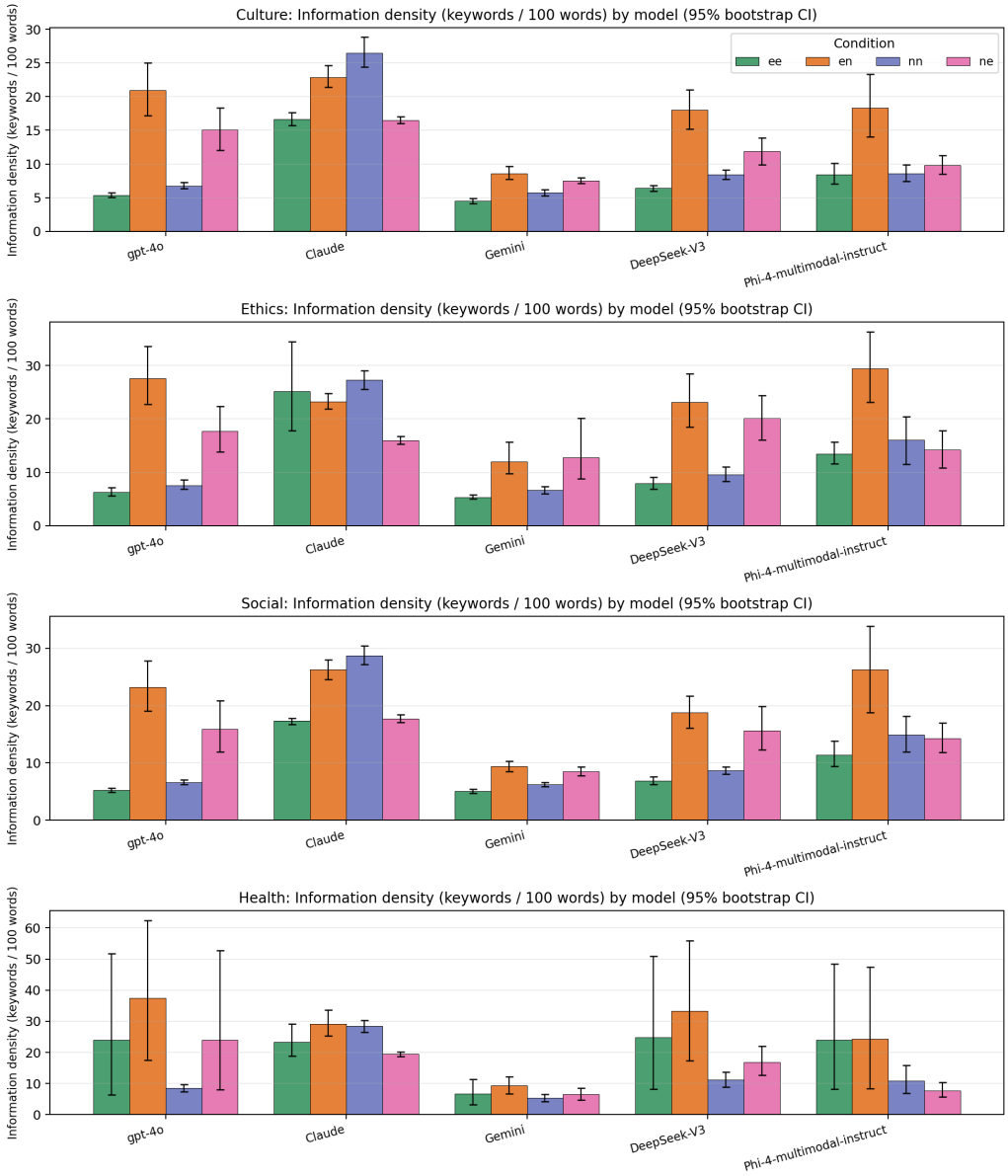

Figure 5: Mean information density (95% bootstrap CI) by model and domain under each condition. CIs replace the earlier ±SD bars so the intervals stay in the non-negative range.

of communicative "efficiency," and we do not use it to argue for a separate quality gain.

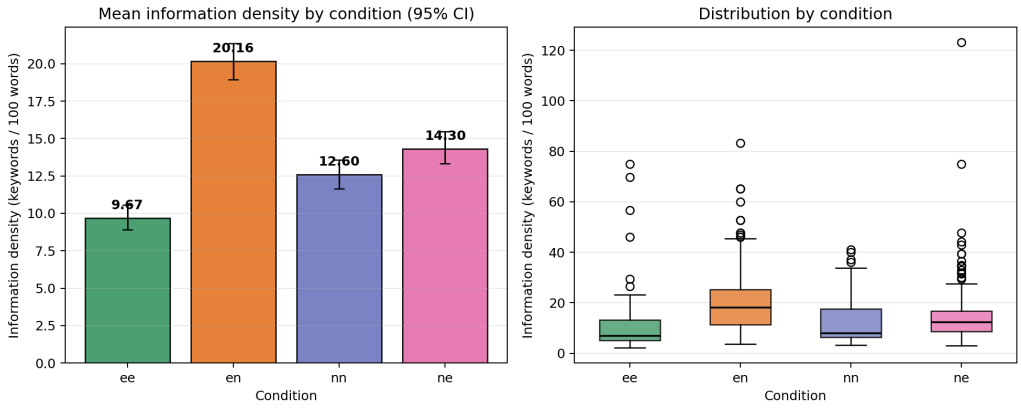

Figure 6: Information-density distributions by condition. Left: mean ± 95% bootstrap CI. Right: full distribution.

Overall, the figure set indicates that language condition changes response length and lexical realization more strongly than semantic content. The key empirical pattern is high semantic alignment together with substantial cross-lingual paraphrasing (raw Jaccard ≈ 0.02 vs. soft Jaccard ≈ 0.53 for $ee \leftrightarrow en$), alongside a large length gap that shrinks by roughly half when measured in tokens.

6. Discussion

The core empirical picture from Section 5 is:

- A large prompt×response interaction on response length that reverses sign across cells: English prompts produce much longer English than Norwegian responses, while Norwegian prompts produce slightly shorter English than Norwegian ones.
- A single-model story that is not uniform: per-model d ranges from Phi-4 (0.78) to GPT-4o (4.42), so the pooled "compression" number is dominated by the more verbose models.
- Semantic content is preserved across conditions (≈ 0.83 cosine similarity, unchanged under chunk-averaged embedding), while surface vocabulary is rearranged: cross-lingual raw Jaccard ≈ 0.02 vs. LaBSE soft Jaccard ≈ 0.53.

We interpret this as evidence that language choice reshapes *how* content is packaged (length, surface vocabulary, lexical realization) more than *what* is conveyed, but we are careful not to over-claim: none of our measurements directly test cultural or register content.

6.1. Candidate Mechanisms

We speculatively flag four mechanisms that plausibly drive the observed pattern; disentangling their relative contributions is beyond the scope of this study.

(1) Tokenizer fertility. English-centric BPE-family tokenizers segment Norwegian text into substantially more subword tokens per word than English text. Under the `cl100k_base` reference tokenizer, mean *nn* responses reach 804 tokens compared with 614 for *ee* despite being shorter in words—so a portion of what looks like "compression" in word counts is a linguistic-typology and tokenizer effect rather than a decision made by the model. The truncation audit (Appendix D) shows the 1,000-token cap was hit mostly on Norwegian outputs (*en*: 19, *nn*: 58) and rarely on long *ee* responses (*ee*: 13), so the measured 52% *ee–en* word gap conflates model behavior with a truncation asymmetry between the two conditions. An uncapped rerun would be needed to separate the two.

(2) RLHF verbosity bias. Preference data for instruction-tuning is overwhelmingly English, and human raters have well-documented biases toward longer, more elaborative English answers [21]. The models we tested that are known to receive heavy English RLHF (GPT-4o, DeepSeek-V3, Gemini) show the largest *ee–en* gaps ($d \geq 2.18$); Phi-4, with a smaller RLHF footprint, shows the smallest ($d = 0.78$). This is consistent with, but does not prove, an RLHF-verbosity account.

(3) Pretraining-corpus asymmetry. Norwegian is a low-resource language relative to English in every publicly documented web-scale corpus. Under scarcity, models may plausibly default to more terse, less rhetorically elaborated Norwegian generations simply because Norwegian long-form prose is under-represented in their training distribution. This is compatible with the observation that Norwegian outputs also show higher density (fewer discourse markers per content word) even when produced by an English prompt.

(4) Safety hedging and refusals. The three items refused in this dataset were all refused only in Norwegian-prompt conditions, and health Q9/Q10—explicit translation asks—were refused universally. This suggests safety-classifier behavior is itself language-conditioned, in a way that would deserve its own study. For length analyses we address this by dropping any question refused in any condition, but the pattern itself is a limitation of comparing generation behavior across languages naively.

6.2. Limitations

Token-cap truncation. Generation was capped based on the defaults (1,000 tokens (2,000 for Claude)), and stop reasons were not logged. A post-hoc audit finds 110 responses that end without terminal punctuation, concentrated in Norwegian-output conditions (*en*: 19, *nn*: 58; together 69% of the truncated set) rather than in English responses (*ee*: 13, *ne*: 21). This is consistent with Norwegian tokenizer fertility exhausting the 1,000-token budget before the response completes—the same mechanism we invoke to explain the tokens-vs.-words gap. A rerun with an unlimited or much higher cap would be needed to fully rule out truncation contamination of the length effects, particularly of the *nn* mean.

Density is a length proxy. The information-density metric is bounded by a fixed keyword cap and is consequently a near-perfect monotonic inverse of response length (Spearman $\rho \approx -1$). It should be read as a descriptive companion to the word-count analysis, not as an independent efficiency measure.

Cultural framing not quantified. Register and framing differences ("more elaborative," "more direct") would need dedicated measures—e.g., formality classifiers, hedging-phrase frequency, or Norway-specific named-entity counts—to be established quantitatively. We do not make quantitative cultural claims in this paper.

Scope. The study covers a single language pair (English–Norwegian, Bokmål), five models, and 68 non-translation questions per model. Generalization to other language pairs, Nynorsk, larger question sets, and multi-turn settings remains to be verified. Question provenance is also mixed (Tulu, Copilot, and manual authoring across topics); we have not tested whether provenance systematically affects any of the reported metrics.

7. Conclusion

Prompt and response language are not neutral parameters: they jointly shape output length, lexical realization, and—to a degree we do not attempt to quantify here—rhetorical framing. Concretely, three findings survive the robustness checks we ran. First, the *ee* → *en* length reduction is $\sim 52\%$ in words but only $\sim 25\%$ in tokens, so a large portion of the "compression" effect reflects Norwegian tokenizer fertility rather than the model producing genuinely more compact content; the interaction between prompt and response language is nonetheless large and highly significant in a mixed-effects model that respects the nested structure of the design ($z = 25.6$). Second, semantic content is preserved across conditions (≈ 0.83 cosine similarity, unchanged under chunk-averaged embedding), while cross-lingual keyword overlap moves from ≈ 0.02 under raw Jaccard to ≈ 0.53 under LaBSE soft Jaccard—direct evidence of concept-level paraphrasing that is invisible to surface-string comparison. Third, the pooled effect size is heterogeneous across models (per-model d ranges from 0.78 to 4.42), so a single "compression" number should be reported alongside its per-model breakdown. For multilingual prompt engineering, the practical takeaway is that changing the response language of a query changes not only its language but its length budget, tokenization cost, and lexical realization; downstream systems that rely on any of these should test in the target-language setting rather than extrapolating from English behavior.

Appendix A. Lists of Experimental Questions

Appendix A.1. Ethics questions

1. What is ethics?
2. What is the difference between ethics and morality?
3. Is it ever ethical to lie?
4. How do ethical theories guide decision-making?
5. What are the principles of utilitarianism?
6. What are the challenges of ethical relativism?
7. What role does culture play in ethics?
8. Should animals have the same rights as humans?
9. How can businesses act ethically?
10. Can ethics exist without religion?

11. What are some ethical issues in technology?
12. How can ethics address environmental challenges?
13. What is the ethical dilemma of self-driving cars?
14. Is it ethical to genetically modify humans?
15. What makes an action good or bad?
16. What are the main criticisms of deontological ethics?
17. How do we define justice in ethical terms?
18. What is the importance of ethical leadership?
19. Can ethical behavior be taught?
20. How do personal values influence ethical decisions?

Appendix A.2. Cultural questions

1. What are some significant cultural festivals celebrated around the world?
2. Describe the influence of the Renaissance on European culture.
3. How does traditional clothing vary across different cultures?
4. What role does storytelling play in preserving indigenous cultures?
5. Explain the concept of cultural appropriation and its implications.
6. How do food traditions reflect cultural identity?
7. Describe the significance of the tea ceremony in Japanese culture.
8. What are the key features of the Maori culture in New Zealand?
9. How does architecture in different countries reflect cultural values?
10. What are some traditional musical instruments unique to specific cultures?
11. How do wedding traditions differ across cultures?
12. What is the significance of the Holi festival in India?
13. Describe the role of masks in African traditional ceremonies.
14. What is the meaning and importance of the Dreamtime in Aboriginal Australian culture?
15. How does the Carnival of Brazil reflect the country's cultural diversity?
16. What are some traditional dances from the Middle East?
17. Describe the influence of Confucianism on Chinese culture.
18. How does the culture of the Inuit people adapt to Arctic life?
19. What are some traditional storytelling forms in Native American cultures?
20. Describe the significance of the cherry blossom season in Japan.

Appendix A.3. Health-related questions (analyzed set, $n = 8$)

1. What are the guidelines for using infant formula when breastfeeding isn't possible? What brands are recommended?
2. What is the standard treatment for an ankle fracture? How long is the recovery process?
3. I'm having persistent headaches. How do I get a referral to a neurologist?
4. What should I do if I think someone is having a stroke? What emergency services should I contact?
5. What are the most effective contraception methods for someone in their 20s? How do I access them?
6. Should I take vitamin D supplements during winter? What's the recommended dosage?
7. How do I properly measure my blood pressure at home? What equipment do I need?
8. What resources are available for someone experiencing depression? Where can I find professional help?

Appendix A.4. Social related questions

1. What is the role of social media in shaping public opinion?
2. How does peer pressure impact decision-making?
3. What are the benefits and drawbacks of social networks?
4. How does cultural diversity affect social interactions?
5. What is the importance of empathy in social relationships?
6. How do stereotypes influence societal behavior?
7. What are the challenges of building inclusive communities?
8. How can social inequality be addressed effectively?
9. What is the impact of technology on social connections?
10. How does urbanization influence social dynamics?
11. What role does education play in promoting social cohesion?
12. What are the ethical considerations of social experiments?
13. How do social movements bring about change?
14. What are the effects of globalization on social structures?
15. How do social norms evolve over time?
16. What is the impact of generational differences on social values?
17. How can individuals contribute to societal well-being?
18. What are the consequences of social isolation?
19. How does social trust influence societal development?
20. What are the roles of family in shaping social values?

Appendix B. Prompt Templates

Each question was sent as a single user turn under one of the four conditions below. The Azure-hosted models (GPT-4o, DeepSeek-V3, Phi-4-multimodal-instruct) and Claude were sent the system message "You are a helpful assistant."; Gemini was called via `generate_content` with no explicit system prompt (SDK default). Placeholders $\texttt{\{Q\}}_{\texttt{EN}}$ and $\texttt{\{Q\}}_{\texttt{NO}}$ hold the English or Norwegian version of the question.

- *ee* **(English → English):** $\texttt{\{Q\}}_{\texttt{EN}}$
- *en* **(English → Norwegian):** `answer in Norwegian:`$\texttt{\{Q\}}_{\texttt{EN}}$
- *nn* **(Norwegian → Norwegian):** $\texttt{\{Q\}}_{\texttt{NO}}$
- *ne* **(Norwegian → English):** `svar på engelsk:`$\texttt{\{Q\}}_{\texttt{NO}}$

Two design choices are worth flagging explicitly, since they mean the four conditions are not a clean 2×2 of language alone. First, *ee* and *nn* carry no explicit language instruction (relying on the implicit convention that a model answers in the language of the query), whereas *en* and *ne* prepend an explicit cue—so any *ee*–*en* or *nn*–*ne* difference partially reflects the presence of the instruction. Second, the two cues are themselves in different languages: *en*'s cue is in English ("answer in Norwegian:"), while *ne*'s cue is in Norwegian ("svar på engelsk:"). Effects attributed to prompt language in the cross-lingual conditions therefore combine (a) the primary question's language, (b) the presence of a cue, and (c) the cue's own language.

Appendix C. Translation-task health items (excluded from length analyses)

These two items were part of the health topic but ask explicitly for translation, which makes the response language over-determined and (in Norwegian-prompt conditions) triggered universal refusals. They are excluded from the length, density, similarity, and Jaccard analyses reported in the main text and are shown here for completeness.

9. Please translate this medical report excerpt into Norwegian, maintaining the technical accuracy: "The cardiovascular assessment revealed moderate aortic regurgitation with an ejection fraction of 56%. Pharmacological management includes ACE inhibitors alongside diuretic therapy. Regular monitoring of renal function is indicated."

10. Please translate this patient's explanation into Norwegian, keeping it easy to understand: "Your heart test showed that one of your heart valves lets some blood flow backward. Your heart is still pumping well. You'll need to take two types of medicine: one to help your blood vessels relax, and another to reduce fluid buildup. We'll need to check your kidney function regularly."

Appendix D. Refusal and Truncation Audits

On the raw 1,400 intended responses (5 models × 70 questions × 4 conditions), 26 were declined by content filters, all in Norwegian-prompt conditions. Table D.3 summarizes; balancing the design by dropping any question refused in *any* condition removes 13 (model, question) pairs and leaves 1,348 responses (337 per condition).

Table D.3: Number of missing responses per model × condition.

Model	*ee*	*en*	*nn*	*ne*
Claude 3.5 Haiku	0	0	2	2
DeepSeek-V3	0	0	3	3
Gemini 2.5 Pro	0	0	2	2
Phi-4-multimodal-instruct	0	0	3	3
GPT-4o	0	0	3	3

A separate post-hoc audit of stop reasons (which the collection scripts did not log) finds 110 raw response files that terminate without terminal punctuation. Table D.4 shows the breakdown by model and condition: truncations concentrate in Norwegian-output conditions (*en*: 19, *nn*: 58; together 69% of the total), not in English ones (*ee*: 13, *ne*: 21). Phi-4 (47 truncations) and Claude (21) account for the majority, and DeepSeek-V3 shows zero truncations in every condition except *nn*. This pattern is consistent with Norwegian's higher subword-tokenizer fertility exhausting the fixed 1,000-token `max_tokens` cap before generation completes. We flag this as the most important open question in the current dataset: a rerun with an unlimited or much higher cap and logged stop reasons would be required to fully rule out truncation contamination—particularly of the *nn* mean length.

Table D.4: Number of responses terminating without terminal punctuation, by model and condition.

Model	*ee*	*en*	*nn*	*ne*	Total
Claude 3.5 Haiku	2	7	11	1	21
DeepSeek-V3	0	0	13	0	13
Gemini 2.5 Pro	3	1	1	5	10
Phi-4-multimodal-instruct	2	10	21	14	47
GPT-4o	7	0	12	0	19
Total	14	18	58	20	110

Appendix E. Paired Example: Claude, health Q0 (infant formula)

Both excerpts are the opening portion of the model's response to the same question in the same session. The *ee* response recommends the US-market brands Similac, Enfamil, Gerber Good Start, and Earth's Best Organic; the *en* response recommends the Norway-market brands Tine Nutrilett, Semper, Hipp, Aptamil, Nan, and Nutricia. Both list broadly the same clinical guidance (formula choice, preparation, feeding schedule), but with locally appropriate product and regulator references.

ee (English → English, first ∼160 words)..

> Here are some general guidelines for using infant formula:
> 1. Choosing a Formula: Use iron-fortified formula for infants under 12 months. Popular recommended brands include Similac, Enfamil, Gerber Good Start, Earth's Best Organic.
> 2. Preparation Guidelines: Always wash hands before preparing. Use sterile bottles and clean water. Follow mixing instructions precisely. Do not dilute formula to make it last longer. Discard unused formula after 1 hour at room temperature.
> 3. Feeding Recommendations: Feed on demand, typically every 2–3 hours. Use appropriate nipple size for baby's age. Hold baby semi-upright during feeding. Burp periodically during feeding.
> 4. Important Considerations: Consult pediatrician about specific formula needs...

en (English → Norwegian, first ∼130 words, our English gloss in brackets)..

> Her er retningslinjene for bruk av morsmelkerstatning når amming ikke er mulig [*Here are the guidelines for using infant formula when breast-feeding is not possible*]:
> Generelle retningslinjer: Konsultér alltid lege eller helsesykepleier før valg av morsmelkerstatning. Følg anvisningene på pakken nøye for korrekt blanding. Bruk rent vann og sterile flasker...
> Anbefalte merker i Norge [*Recommended brands in Norway*]: Tine Nutrilett, Semper, Hipp, Aptamil, Nan, Nutricia.
> Viktige tips [*Important tips*]: Velg erstatning tilpasset barnets alder. Sjekk at produktet oppfyller norske helsemyndigheters krav [*that the product meets Norwegian health-authority requirements*]...

Appendix F. Log-scale robustness check for the length model

Because word counts are right-skewed (median ≈ 240, maximum $\approx 1{,}370$), we refit the mixed-effects length model on $\log(1 + \text{words})$ as a robustness check. This makes the residual distribution more nearly symmetric and, more usefully, makes each fixed-effect coefficient interpretable as a *multiplicative* change on the raw scale via $\exp(\hat{\beta})$. Table F.5 reports the fixed effects on both scales; results reproduce the raw-scale narrative to within a percentage point and are qualitatively identical.

Table F.5: Mixed-effects length model on $\log(1+\text{words})$, analyzed set ($n = 1{,}354$; words $\sim$ prompt$\times$response+model+$(1 \mid \text{question})$). Percentages are back-transformed as $1-\exp(\hat{\beta})$ and represent the marginal effect at the reference level.

Term	$\hat{\beta}_{\log}$	SE	z	Back-transformed	Direct-means
Intercept	5.320	0.036	146.20	—	—
Prompt: Norwegian	-0.443	0.030	-14.91	-36% (ee $\rightarrow$ ne)	-37%
Response: Norwegian	-0.764	0.030	-25.79	-53% (ee $\rightarrow$ en)	-52%
Prompt$\times$Response (NO,NO)	$+0.966$	0.042	$+23.01$	—	—

The interaction is large and highly significant on the log scale ($z = 23.01$, comparable to the raw-scale $z = 25.62$). Combining the main effects and the interaction back-transforms to $\exp(-0.443 - 0.764 + 0.966) - 1 \approx -21\%$ for the nn cell relative to ee (direct-means value: -19%). Direction, significance, and magnitude of every effect match the raw-scale fit; we retain the raw-scale numbers in the main text because Table 2 and Table 1 are already expressed in word counts.

Appendix G. Data and Code Availability

All response files, analysis code, and figure-generation scripts are on Github. The reproduction pipeline (`functions/recompute_b.py`) regenerates every number and figure in Sections 4–5 from the raw text responses. Queries were issued between March and June 2026. Exact model identifiers used: `gpt-4o` (Azure snapshot `2024-08-06`), `claude-3-5-haiku-latest`, `gemini-2.5-pro-preview-03-25`, `DeepSeek-V3` (Azure OpenAI-compatible endpoint), `Phi-4-multimodal-instruct` (Azure OpenAI-compatible endpoint). Decoding parameters were held at API defaults except `max_tokens`: 1,000 for the Azure-hosted models, 2,000 for Claude, and unspecified for Gemini. No sampling was performed (single call per question per condition). LaBSE embeddings use `sentence-transformers/LaBSE` at its default hyperparameters.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems 33 (NeurIPS), 2020.
URL `https://arxiv.org/abs/2005.14165`

[2] T. Le Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, et al., BLOOM: A 176B-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).
URL `https://arxiv.org/abs/2211.05100`

[3] L. Qin, Q. Chen, F. Wei, S. Huang, W. Che, Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 2695–2709. `doi:10.18653/v1/2023.emnlp-main.163`.
URL `https://aclanthology.org/2023.emnlp-main.163/`

[4] F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, J. Wei, Language models are multilingual chain-of-thought reasoners, in: The Eleventh International Conference on Learning Representations (ICLR), 2023.
URL `https://arxiv.org/abs/2210.03057`

[5] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. doi:10.1162/tacl_a_00638.
URL https://aclanthology.org/2024.tacl-1.9/

[6] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, S. Hasan, Does prompt formatting have any impact on llm performance?, arXiv preprint arXiv:2411.10541 (2024).
URL https://arxiv.org/abs/2411.10541

[7] H. Gilbert, M. Sandborn, D. C. Schmidt, J. Spencer-Smith, J. White, Semantic compression with large language models, arXiv preprint arXiv:2304.12512 (2023).
URL https://arxiv.org/abs/2304.12512

[8] J. G. Lu, L. L. Song, L. D. Zhang, Cultural tendencies in generative ai, Nature Human Behaviour 9 (2025) 2360–2369. doi:10.1038/s41562-025-02242-1.
URL https://www.nature.com/articles/s41562-025-02242-1

[9] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, et al., Challenges and strategies in cross-cultural NLP, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 6997–7013. doi:10.18653/v1/2022.acl-long.482.
URL https://aclanthology.org/2022.acl-long.482/

[10] C. Li, M. Chen, J. Wang, S. Sitaram, X. Xie, CultureLLM: Incorporating cultural differences into large language models, in: Advances in Neural Information Processing Systems 37 (NeurIPS), 2024.
URL https://arxiv.org/abs/2402.10946

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318. doi:10.3115/1073083.1073135.
URL https://aclanthology.org/P02-1040/

[12] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
URL https://aclanthology.org/W04-1013/

[13] A. B. Sai, A. K. Mohankumar, M. M. Khapra, A survey of evaluation metrics used for NLG systems, ACM Computing Surveys 55 (2) (2023) 1–39. doi:10.1145/3485766.
URL https://dl.acm.org/doi/10.1145/3485766

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: The Eighth International Conference on Learning Representations (ICLR), 2020.
URL https://arxiv.org/abs/1904.09675

[15] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 878–891. doi:10.18653/v1/2022.acl-long.62.
URL https://aclanthology.org/2022.acl-long.62/

[16] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd Edition, Lawrence Erlbaum Associates, Hillsdale, NJ, 1988.

[17] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[18] K. Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation 28 (1) (1972) 11–21. doi:10.1108/eb026526.

[19] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Information Processing & Management 24 (5) (1988) 513–523. doi:10.1016/0306-4573(88)90021-0.

[20] P. Jaccard, The distribution of the flora in the alpine zone, New Phytologist 11 (2) (1912) 37–50. doi:10.1111/j.1469-8137.1912.tb05611.x.

[21] P. Singhal, T. Goyal, J. Xu, G. Durrett, A long way to go: Investigating length correlations in RLHF, arXiv preprint arXiv:2310.03716 (2024).
URL https://arxiv.org/abs/2310.03716