

Toward a Graded Measure of Belief Stability in Large Language Models

Samantha Dies^{1#}, Branden Fitelson¹, and Tina Eliassi-Rad^{1,2}

¹Northeastern University, 360 Huntington Ave, Boston, MA 02115 USA

²Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501 USA

#dies.s@northeastern.edu

ABSTRACT

Large language models (LLMs) increasingly mediate how people access and reason with information, yet factual reliability is usually evaluated one judgment at a time. We introduce **graded belief stability**, a relational measure of how well a belief persists within an LLM’s broader belief system. Unlike individual belief probability, it asks whether support for a claim persists when that claim is considered alongside the model’s other epistemic commitments. We operationalize this idea with a Direct Conditional estimator that uses internal model representations to estimate conditional belief probabilities. Across 12 LLMs and three domains, lower-stability beliefs exhibit greater mean behavioral movement under conversational challenge in 83.3% of model–domain settings after matching on individual belief probability. Graded belief stability therefore extends reliability assessment beyond how strongly an LLM supports a claim to how robustly that belief is supported within its broader system of beliefs.

1 Introduction

Large language models (LLMs) increasingly impact how people access and reason with information, making the reliability of their factual judgments central to their use [1, 2]. Existing evaluations characterize several dimensions of this reliability, including accuracy, uncertainty, and calibration [3, 4, 5, 6]. In interactive settings, however, factual claims are rarely considered in isolation. New claims, corrections, and contextual information accumulate over a conversation, and models may revise answers they previously supported. Reliability therefore also depends on a different question: *which of a model’s beliefs persist as the informational context around them changes?*

Such belief revisions are well documented. LLM factual judgments can vary with paraphrasing and changes in semantic context [7, 8, 9], as well as under user disagreement, contradictory feedback, and multi-turn interaction [10, 11, 12, 13, 14]. Work on systematic belief consistency and knowledge editing likewise suggests that factual commitments do not behave as isolated entries in a knowledge base: accepting or revising one proposition can affect related propositions [15, 16]. Consider an LLM that assigns similar support to “Paris is in France” and “Oslo is in Norway.” If it also considers “Oslo is in Sweden” plausible, the Oslo claim sits alongside an incompatible proposition the model has not ruled out. The Paris and Oslo claims may therefore look equally strong on their own even though one is more fragile within the model’s broader belief system. We refer to this difference as *belief stability*.

Existing work captures important pieces of this picture without directly measuring it. Representation-based studies find that veracity and epistemic uncertainty are systematically reflected in LLM hidden states [17, 18, 19, 20, 21], while LLMs do not reliably distinguish epistemic concepts such as belief, knowledge, and fact [22]. Behavioral evaluations, meanwhile, reveal when outputs change under particular external perturbations [7, 10, 11, 12]. What these approaches do not capture is how a belief is situated relative to the model’s *other* epistemic commitments. A belief may receive strong support on its own yet become fragile in light of other propositions the model regards as plausible. Stability is therefore a relational property of a belief within a broader belief system.

Formal epistemology provides a principled basis for making this distinction precise. The Lockean Thesis connects graded credence to categorical belief through a probability threshold, with sufficiently probable propositions counting as beliefs [23, 24]. Leitgeb’s Humean Thesis strengthens this threshold-based approach by requiring a belief to remain sufficiently probable when conditioned on propositions an agent does not disbelieve [25]. We adapt this idea to LLMs by defining the **graded belief stability** of a believed proposition P as the fraction of the model’s non-disbeliefs x for which $\Pr(P \mid x)$ remains above its belief threshold. Graded stability thus distinguishes how strongly a model supports P on its own from how broadly that support persists across its broader epistemic state.

Estimating graded stability requires eliciting conditional belief probabilities from LLM representations. Supervised probes can recover veracity-related information from hidden states [17, 18, 20, 26], but $\Pr(P \mid x)$ is not directly observable. We therefore introduce a trivalent *Direct Conditional* estimator grounded in established accounts of

indicative conditionals and non-bivalent probability [27, 28]. For each belief–non-disbelief pair (P, x) , we construct a conditional statement of the form “Given x , P ” and probe its hidden representation to estimate the conditional probabilities needed to compute graded stability.

We evaluate graded belief stability across 12 instruction-tuned LLMs from four model families, ranging from 3B to 72B parameters, and across three domains: City Locations, Medical Indications, and Word Definitions. We test whether graded stability captures structure beyond individual belief probability and whether the underlying probability estimates are approximately probabilistically coherent. We then characterize its variation across semantic domains and evaluate whether it is associated with resistance to conversational challenge.

Our work makes five core contributions:

1. **Formalization:** We introduce *graded belief stability* as a statement-level measure of how well an LLM’s belief persists across its broader belief system.
2. **Measurement:** We develop a representation-based Direct Conditional estimator of conditional belief probability for computing graded stability.
3. **Validation:** We show that graded stability captures systematic information beyond individual belief probability, where 99.0% of model-pair residual correlations are positive. The measured probability systems are also approximately coherent, with median distance from coherence ranging from 0.015 to 0.062 across domains.
4. **Characterization:** We find that City Locations is consistently the most stable domain, with a median across models of mean graded stability of 0.88, compared with 0.71 for Medical Indications and 0.74 for Word Definitions.
5. **Behavioral validity:** Among beliefs matched on individual belief probability, lower-stability beliefs show greater mean movement under repeated conversational challenge for 83.3% of model–domain settings.

This work captures a dimension of LLM reliability that individual belief probability alone cannot. Two factual judgments can receive similar support while occupying very different positions within the model’s broader belief system, making one substantially more fragile than the other. Graded belief stability quantifies this relational property, extending reliability assessment beyond *how strongly* an LLM supports a claim to *how robustly that belief is supported within the model’s broader system of beliefs*.

2 Results

We begin by formalizing graded belief stability γ and introducing the Direct Conditional representation-based estimator for conditional belief. We then validate the measure, showing that it captures structure beyond individual belief probability and is approximately probabilistically coherent. We next characterize variation in γ across domains, before testing whether graded stability distinguishes behavioral resilience among beliefs matched on individual belief probability.

2.1 Formalizing graded belief stability

We define graded belief stability by adapting the stability theory of belief from formal epistemology, in which a belief is stable when it persists after conditioning on propositions the agent does not reject [24]. We extend this binary notion to a statement-level measure of the degree to which an LLM’s belief persists across conditioning contexts.

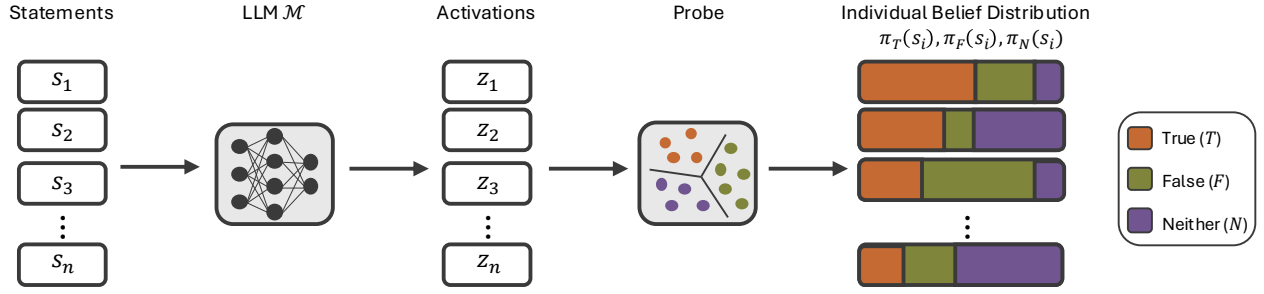
Let $\mathcal{M}$ denote an LLM and let $\text{Pr}_{\mathcal{M}}(s)$ denote the probability that $\mathcal{M}$ assigns to a statement s being true. Following the Lockean Thesis, we take categorical belief to correspond to sufficiently high probability relative to a belief threshold $t_{\mathcal{M}}$ [23, 24]. In our empirical implementation below, we first identify categorical belief from a trivalent probe and then instantiate $t_{\mathcal{M}}$ from the resulting belief set rather than imposing a universal cutoff. We denote the set of beliefs held by $\mathcal{M}$ as $\mathcal{B}_{\mathcal{M}}$ and the larger set of propositions that $\mathcal{M}$ does not disbelieve, including both beliefs and suspended beliefs, as $\mathcal{X}_{\mathcal{M}}$.

The Humean Thesis strengthens this static notion of belief by requiring beliefs to remain sufficiently probable when considered alongside the agent’s other non-disbeliefs [24, 25]. Intuitively, a stable belief should remain above the belief threshold even when considered alongside propositions the model itself does not reject. Specifically, a belief $P \in \mathcal{B}_{\mathcal{M}}$ is considered *stable* if conditioning on any $x \in \mathcal{X}_{\mathcal{M}} \setminus \{P\}$ does not lower its probability below the belief threshold:

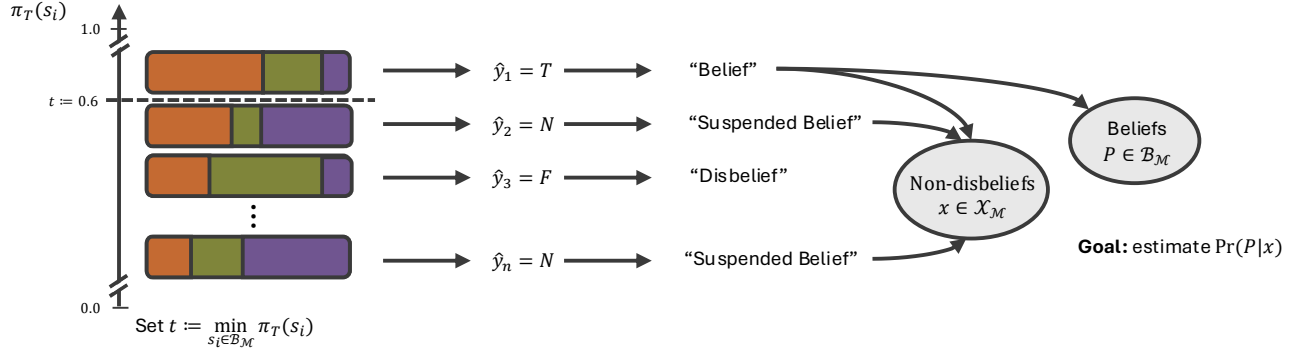
$$P \text{ is stable} \iff \text{Pr}_{\mathcal{M}}(P \mid x) > t_{\mathcal{M}} \quad \forall x \in \mathcal{X}_{\mathcal{M}} \setminus \{P\}. \quad (1)$$

Thus, propositions with similar $\text{Pr}_{\mathcal{M}}(P)$ may nevertheless differ substantially in stability.

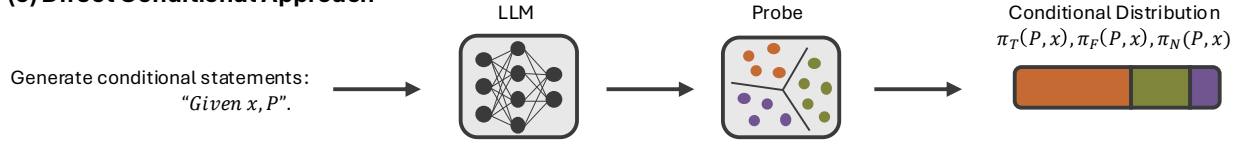
(a) Individual Belief Probabilities



(b) From Probabilities to Belief Sets



(c) Direct Conditional Approach



(d) Computing Graded Stability

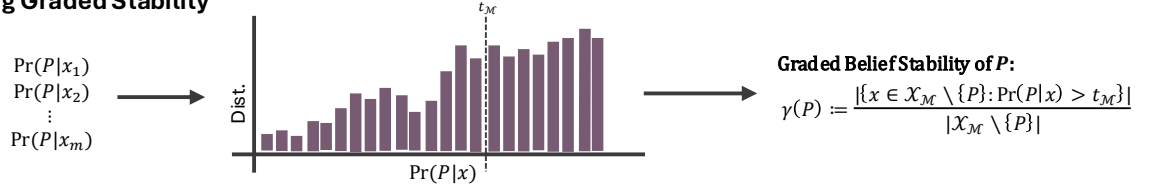


Figure 1. Measuring graded belief stability in LLMs. (a) For each statement s , we extract an internal LLM representation and use a probe to estimate a distribution over **True** (T , orange), **False** (F , olive), and **Neither** (N , purple). (b) The probe’s predicted states $\hat{y}$ determine whether each statement is treated as a belief ($\hat{y} = T$), disbelief ($\hat{y} = F$), or suspended belief ($\hat{y} = N$), yielding the belief set $\mathcal{B}_M$ and non-disbelief set $\mathcal{X}_M$. We define the empirical belief threshold t_M as the minimum **True**-class probability among statements classified as beliefs. We then estimate $\Pr_M(P|x)$ for each $P \in \mathcal{B}_M$ and $x \in \mathcal{X}_M$. (c) The Direct Conditional approach probes representations of conditional statements of the form “Given x , P .”, and (d) we compute graded stability $\gamma_M(P)$ as the fraction of conditioning propositions for which the conditional probability of P remains above t_M .

Binary stability, however, treats all violations identically. A proposition that falls below the belief threshold under a single conditioning statement is indistinguishable from one that does so under nearly all of them. We therefore define the **graded belief stability** of P as

$$\gamma_M(P) := \frac{|\{x \in \mathcal{X}_M \setminus \{P\} : \Pr_M(P|x) > t_M\}|}{|\mathcal{X}_M \setminus \{P\}|}. \quad (2)$$

Thus, $\gamma_M(P) \in [0, 1]$ is the proportion of non-disbeliefs under which belief in P is preserved. Higher values indicate beliefs that persist more broadly across the model’s belief system, distinguishing graded stability from the probability

assigned to P in isolation.

2.2 Measuring graded belief stability in LLMs

Measuring graded stability requires identifying an LLM’s beliefs and non-disbeliefs and estimating $\Pr_{\mathcal{M}}(P \mid x)$ for belief–non-disbelief pairs. We use calibrated supervised probes over LLM hidden representations to obtain these probability estimates, since unconstrained prompted judgments need not yield calibrated probability estimates and can be sensitive to prompt formulation and conversational context [7, 10, 12, 29].

For each statement s_i , we extract hidden representations across layers ℓ of model $\mathcal{M}$ and train the sparse-aware multiple-instance learning (**sAwMIL**) probe [20], which models **True** (T), **False** (F), and **Neither** (N) as distinct veracity classes.¹ After calibration, the probe returns the trivalent distribution

$$\boldsymbol{\pi}_{\mathcal{M}}(s_i) = [\pi_T(s_i), \pi_F(s_i), \pi_N(s_i)], \quad (3)$$

where $\pi_c(s_i)$ is the probe-assigned probability of class c . The predicted epistemic state is then

$$\hat{y}_i = \arg \max_{c \in \{T, F, N\}} \pi_c(s_i). \quad (4)$$

We interpret $\hat{y}_i = T$ as belief, $\hat{y}_i = F$ as disbelief, and $\hat{y}_i = N$ as suspension of belief (Fig. 1(a–b)). The probe therefore determines the empirical belief and non-disbelief sets directly:

$$\mathcal{B}_{\mathcal{M}} = \{s_i : \hat{y}_i = T\}, \quad \mathcal{X}_{\mathcal{M}} = \{s_i : \hat{y}_i \in \{T, N\}\}. \quad (5)$$

This trivalent operationalization retains an explicit state for suspended belief rather than forcing every proposition into a binary believed/not-believed distinction.

We then associate the probe-defined categorical belief set with an empirical Lockean-style threshold. For each model and domain, we define

$$t_{\mathcal{M}} := \min_{s_i \in \mathcal{B}_{\mathcal{M}}} \pi_T(s_i). \quad (6)$$

Importantly, $t_{\mathcal{M}}$ does not determine which propositions belong to $\mathcal{B}_{\mathcal{M}}$. Rather, categorical belief is already defined by the probe’s trivalent prediction in Equation (5). Instead, the threshold translates this categorical belief set into a scalar criterion for belief persistence. Specifically, $t_{\mathcal{M}}$ is the most stringent **True**-class support threshold satisfied by every proposition that the probe identifies as a belief. This anchors the persistence criterion to the model’s own empirical belief boundary rather than imposing an externally chosen probability cutoff.

Because π_T is one component of a calibrated three-class distribution over **True**, **False**, and **Neither**, $t_{\mathcal{M}}$ will not necessarily exceed 0.50, but is instead lower-bounded by 0.33. The resulting threshold should therefore be understood as a model-, domain-, and probe-specific support floor induced by the empirical belief set, rather than as a universal normative probability threshold (Fig. 1(b), Sec. 4.2).

We next estimate conditional belief probability for every belief $P \in \mathcal{B}_{\mathcal{M}}$ and non-disbelief $x \in \mathcal{X}_{\mathcal{M}} \setminus \{P\}$. The **Direct Conditional** approach represents $P \mid x$ using the conditional statement “*Given x , P .*” and probes its internal representation directly (Fig. 1(c)). A three-class **sAwMIL** probe predicts whether the resulting conditional is **True**, **False**, or **Neither** using Cooper–Cantwell training labels [28, 31, 32]. For each pair (P, x) , the probe returns the trivalent distribution

$$\boldsymbol{\pi}_{\mathcal{M}}^{\text{direct}}(P, x) = [\pi_T^{\text{direct}}(P, x), \pi_F^{\text{direct}}(P, x), \pi_N^{\text{direct}}(P, x)]. \quad (7)$$

We use the conditional’s **True**-class mass to define our Direct Conditional estimator

$$\widehat{\Pr}_{\mathcal{M}}^{\text{direct}}(P \mid x) := \pi_T^{\text{direct}}(P, x). \quad (8)$$

We use $\widehat{\Pr}_{\mathcal{M}}^{\text{direct}}(P \mid x)$ as an empirical operational estimator of conditional belief probability rather than as an algebraic reconstruction of the normalized Cooper–Cantwell probability of a trivalent conditional. In particular, we do not renormalize over the **True** and **False** states. This keeps conditional support on the same three-class scale used to define the empirical belief threshold $t_{\mathcal{M}}$ and allows probability mass assigned to **Neither** to count against persistence of the original belief. Section 4.3 motivates this choice and contrasts it with our complementary

¹Full activation extraction, probe training, and layer-selection procedures are described in Sec. 4.1, and layer-wise probe performance is reported in Sec. D.3. We additionally reproduce the analyses using Support Vector Machine (SVM) [30] and **Mass Mean** [17] probes in Sec. I.3.

Stability explained by belief probability

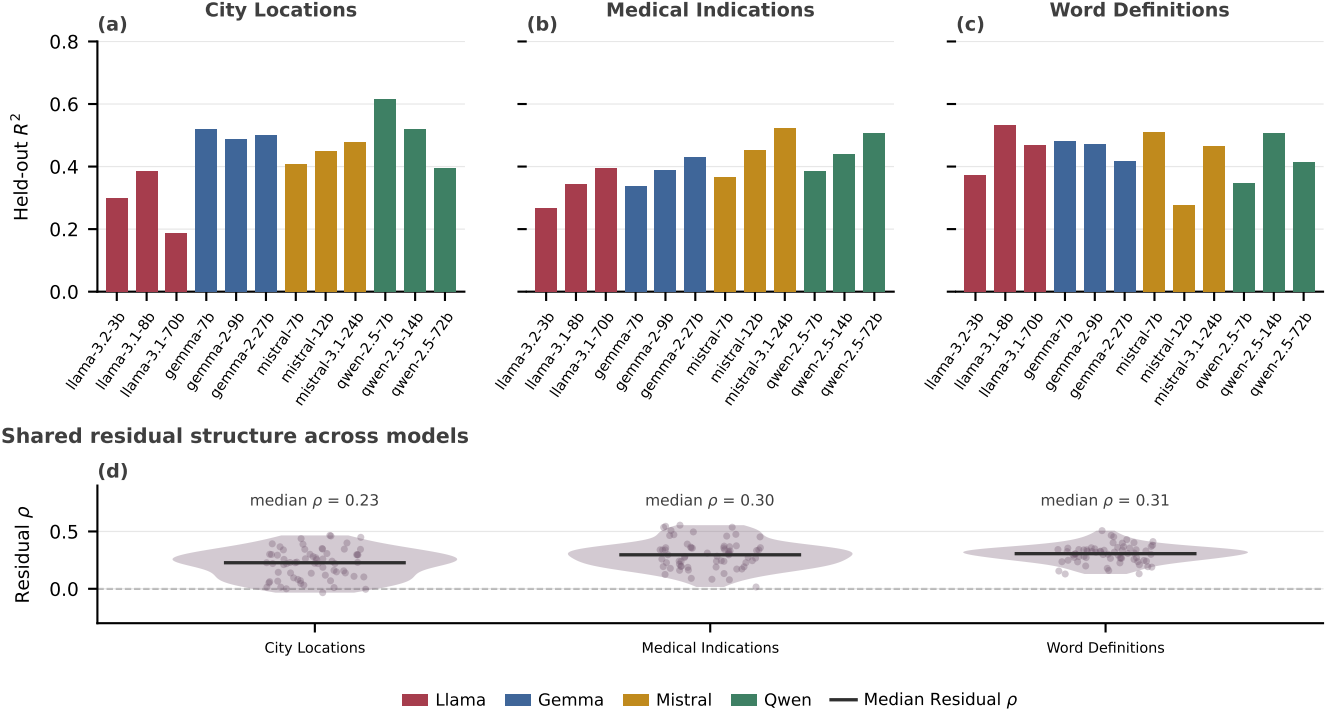


Figure 2. Relationship between individual belief probability and graded stability. We show the held-out R^2 for predicting graded stability $\gamma_{\mathcal{M}}(P)$ from individual belief probability $\pi_T(P)$ for (a) City Locations, (b) Medical Indications, and (c) Word Definitions across Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) model families. We also report (d) distributions of pairwise Spearman correlations ρ between LLMs’ residual graded stability within each domain, where black lines denote the median correlation across model pairs. Individual belief probability predicts a meaningful proportion of graded stability, but graded stability captures proposition-level variation beyond individual belief probability that is systematically shared across models.

Joint-to-Conditional estimator, which instead normalizes over determinate conditional states. Results on the secondary Joint-to-Conditional estimator are reported in Section I.2.2.

Replacing $\Pr_{\mathcal{M}}(P | x)$ in Eq. (2) with the Direct Conditional estimate in Eq. (8) yields our empirical Direct Conditional stability score $\gamma_{\mathcal{M}}^{\text{direct}}(P)$ (Fig. 1(d); Sec. 4.4). We apply this probing-based framework to 12 instruction-tuned² LLMs spanning four model families and ranging from 3B to 72B parameters. We evaluate three domains with differing epistemic characteristics: **City Locations**, consisting of comparatively clear-cut factual relations; **Medical Indications**, requiring specialized factual knowledge; and **Word Definitions**, containing comparatively ambiguous lexical relations [20].

2.3 Validating graded stability

We validate graded stability along two complementary dimensions: whether it captures information beyond individual belief probability and whether the probability estimates underlying it are approximately probabilistically coherent. Main-text results use the Direct Conditional estimator and the **sAwMIL** probe. Joint-to-Conditional results and analyses using the **SVM** and **Mass Mean** probes appear in Section I.

2.3.1 Graded stability captures information beyond belief probability

A central motivation for graded stability is that it captures how a belief behaves in the context of an LLM’s broader belief system rather than simply how strongly that belief is held in isolation. We therefore first ask how well the individual belief probability $\pi_T(P)$ predicts $\gamma_{\mathcal{M}}(P)$ for beliefs $P \in \mathcal{B}_{\mathcal{M}}$.

If graded stability were simply a transformation of belief probability, then a sufficiently flexible function of $\pi_T(P)$ should accurately predict $\gamma_{\mathcal{M}}(P)$ for beliefs not used to fit that function. For each model and domain, we attempt

²Matched base-model analyses are reported in Section I.1.

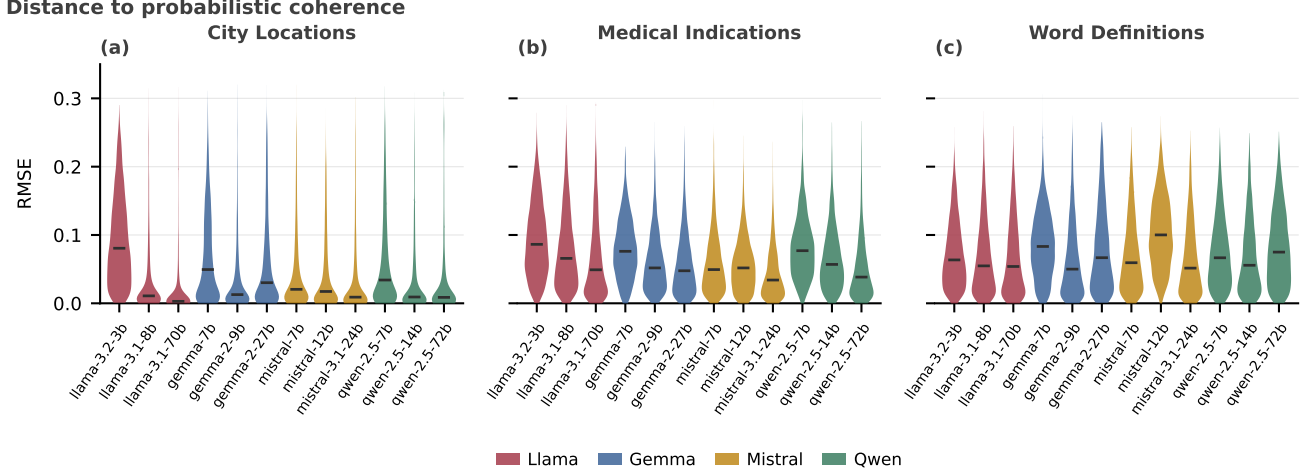


Figure 3. Distance of measured probabilities from probabilistic coherence. Violins show the distribution across (P, x) pairs of the root-mean-square (RMS) adjustment d_{CCK} required to project the measured probability distributions for individual statements and conditionals onto the nearest CCK-coherent probability system for (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) model families. Horizontal black lines denote within-model medians. Smaller values indicate greater probabilistic coherence, with $d_{\text{CCK}} = 0$ corresponding to exact coherence. Although exact coherence is uncommon, only modest adjustments are generally required to reconcile the independently measured probabilities with a coherent trivalent probability system.

to capture such a function using a natural cubic spline,

$$\hat{\gamma}_{\mathcal{M}}(P) = f_{\mathcal{M}}(\pi_T(P)), \quad (9)$$

with three degrees of freedom. The spline allows belief probability and stability to have a smooth nonlinear relationship without assuming a linear relationship. We use five-fold cross-validation to generate out-of-sample predictions $\hat{\gamma}_{\mathcal{M}}(P)$ and evaluate performance with R^2 (Sec. 4.5.1). We fix the spline complexity across all models and domains and show that the results are qualitatively unchanged across alternative spline complexities in Section F.1.

Individual belief probability consistently predicts a meaningful but incomplete proportion of the observed variation in graded stability (Fig. 2(a–c)). The median held-out R^2 across models is 0.46 for City Locations, 0.39 for Medical Indications, and 0.47 for Word Definitions. Thus, although belief probability is informative about stability, more than half of the proposition-level variation in $\gamma_{\mathcal{M}}(P)$ remains unexplained by the probability assigned to P in isolation. Graded stability therefore captures substantial variation beyond individual belief strength.

We next ask whether this unexplained variation is systematic. For each proposition P , we define its *residual stability* as

$$r_{\mathcal{M}}(P) := \gamma_{\mathcal{M}}(P) - \hat{\gamma}_{\mathcal{M}}(P), \quad (10)$$

where positive values identify propositions that are more stable than predicted from their individual belief probability and negative values identify propositions that are less stable than predicted. We align pairs of LLMs on propositions believed by both models and compute the Spearman correlation between their residual stability scores (Fig. 2(d)).

Across domains, 99.0% of model-pair residual correlations are positive, with median correlations of $\rho = 0.23$ for City Locations, $\rho = 0.30$ for Medical Indications, and $\rho = 0.31$ for Word Definitions. All three medians exceed a proposition-shuffled permutation null ($p = 0.001$; Sec. 4.5.1). Models therefore tend to find the *same* propositions unusually stable or fragile even after accounting for how strongly they believe those propositions in isolation. This suggests that the variation unique to graded stability contains reproducible proposition-level structure rather than only model-specific measurement noise. Because models are evaluated on the same benchmark propositions using a common probing framework, however, this analysis does not by itself distinguish model-internal relational structure from proposition-level or measurement factors shared across models.

2.3.2 Graded stability probability estimates are approximately coherent

Because graded stability is defined within a probabilistic framework, we next ask whether the independently measured individual and conditional probability distributions are pairwise compatible. Under Cooper–Cantwell–Kleene (CCK) trivalent probability, the distributions for P , x , and $P \mid x$ are coherent when they can all be generated by a single latent 3×3 probability distribution over the trivalent states of (P, x) . For each measured pair, we collect these probabilities into the nine-dimensional vector

$$\mathbf{v}_{P,x} = [\pi_{\mathcal{M}}(P), \pi_{\mathcal{M}}(x), \pi_{\mathcal{M}}^{\text{direct}}(P, x)]. \quad (11)$$

Under CCK semantics, such a latent joint distribution exists if and only if the measured probabilities satisfy

$$\pi_T^{\text{direct}}(P, x) \leq \pi_T(P), \quad \pi_F^{\text{direct}}(P, x) \leq \pi_F(P), \quad \pi_F(x) \leq \pi_N^{\text{direct}}(P, x), \quad \pi_N^{\text{direct}}(P, x) \leq \pi_N(P) + \pi_F(x). \quad (12)$$

We use these constraints to test exact coherence separately for every (P, x) pair (Sec. 4.5.2); a proof of the coherence conditions is provided in Sec. F.2.1. Exact coherence is extremely rare: across all model–domain combinations, no more than 0.21% of measured pairs satisfy all four constraints exactly. Full exact-coherence rates and individual constraint violations are reported in Section F.2.

Because exact coherence is a binary criterion that treats arbitrarily small and large violations identically, we additionally quantify each pair’s distance from the set of coherent probability systems. Following prior work that characterizes approximate coherence in terms of distance to the nearest coherent credence function [33], we extend this distance-based approach to the CCK trivalent setting and define

$$d_{\text{CCK}}(\mathbf{v}_{P,x}) := \sqrt{\frac{1}{9} \|\mathbf{v}_{P,x} - \Pi_{\mathcal{C}}(\mathbf{v}_{P,x})\|_2^2}, \quad (13)$$

where $\mathcal{C}$ is the set of CCK-coherent probability vectors and $\Pi_{\mathcal{C}}(\mathbf{v})$ is the closest vector in this set to the observed probability vector $\mathbf{v}$.³ Thus, $d_{\text{CCK}} = 0$ corresponds to exact coherence, while a value such as $d_{\text{CCK}} = 0.05$ means that reaching the nearest coherent probability system requires a root-mean-square (RMS) adjustment of 0.05 per probability component.

Across models, the median model-level d_{CCK} is 0.015 for City Locations, 0.052 for Medical Indications, and 0.062 for Word Definitions, corresponding to RMS adjustments of 1.5, 5.2, and 6.2 percentage points per probability component, respectively. Thus, despite near-zero exact-coherence rates, the measured probability systems typically require quantitatively modest adjustments to reach the nearest coherent CCK representation. Because graded stability thresholds these probabilities at $t_{\mathcal{M}}$, however, a small component-wise RMS adjustment does not imply that the corresponding absolute stability score would be unchanged after projection. We therefore interpret d_{CCK} as a diagnostic of the measured probability system rather than as a robustness bound on $\gamma_{\mathcal{M}}(P)$.

2.4 Graded belief stability varies with domain

Having established that graded stability captures information beyond individual belief probability and is approximately probabilistically coherent, we next ask how stability varies across beliefs and models by characterizing the distribution of $\gamma_{\mathcal{M}}(P)$ across each model’s beliefs (Fig. 4). Stability differs substantially across semantic domains. Across models, the median of model-level mean graded stability is 0.88 for City Locations, compared with 0.71 for Medical Indications and 0.74 for Word Definitions. Comparing the within-model median stabilities across domains, City Locations has the highest median stability for 11 of the 12 LLMs. The only exception is Llama-3.2-3b, whose median stability is highest for Word Definitions. Within these benchmark-specific belief systems, City Locations beliefs are therefore consistently more stable than beliefs in the Medical Indications and Word Definitions domains. This pattern is consistent with differences in the semantic structure of the domains, as City Locations contains comparatively clear-cut factual relations, whereas Medical Indications requires more specialized factual knowledge and Word Definitions contains comparatively ambiguous lexical relations. The same broad domain structure is visible in matched pretrained base models (Sec. I.1.1).

These differences extend beyond shifts in average stability to the proposition-level distributions themselves. In City Locations, stability is concentrated near the upper end of the range for most models, although long lower tails show that even this comparatively stable domain contains individual beliefs that are substantially more fragile (Fig. 4(a)). Medical Indications and Word Definitions exhibit broader distributions and considerably greater variation across models. In Medical Indications in particular, several of the smallest models, including Llama-3.2-3b,

³Full optimization and convergence details are provided in Sec. 4.5.2.

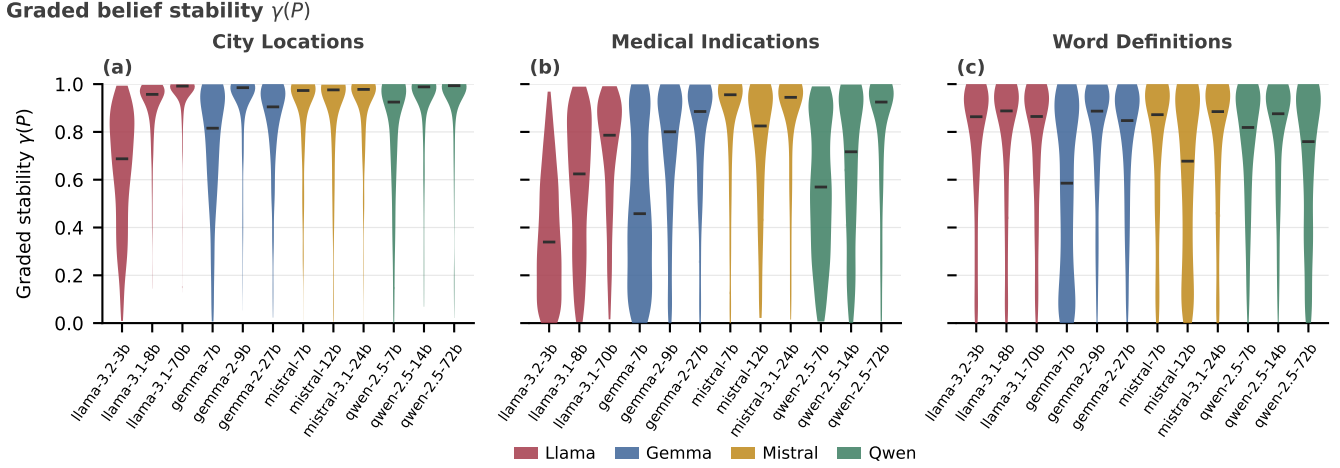


Figure 4. Variation in graded belief stability across domains. Violin plots show the proposition-level distribution of graded stability $\gamma_{\mathcal{M}}^{\text{direct}}(P)$ for each of the 12 instruction-tuned LLMs in (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) model families. Horizontal black lines denote within-model medians. Stability distributions differ substantially across the three benchmark domains, with City Locations exhibiting the highest median stability for 11 of the 12 models.

Gemma-7b, and Qwen-2.5-7b, have markedly lower median stability than larger models from the same families (Fig. 4(b)).

Some of the between-model variation in Figure 4 also suggests a possible relationship with model scale. In City Locations and Medical Indications, several of the smallest models within a family have lower median stability than their larger counterparts (Fig. 4(a-b)). This pattern is not consistent across all families, however, and is largely absent in Word Definitions. We examine model scale more directly in Section H, where we demonstrate that neither parameter count nor model depth shows a consistent relationship with graded stability across domains. Thus, model scale may contribute to some of the variation observed within families, but it does not provide a general explanation for differences in $\gamma_{\mathcal{M}}(P)$.

2.5 Graded stability is associated with behavioral movement beyond belief probability

Graded stability is defined from relationships among internal representations, but its usefulness depends in part on whether those differences correspond to independently measured model behavior. We therefore test whether $\gamma_{\mathcal{M}}(P)$ distinguishes resistance to conversational challenge among beliefs assigned similar individual probabilities. Importantly, the challenge prompts are not conditioning propositions $x \in \mathcal{X}_{\mathcal{M}}$ and do not instantiate the formal operation used to compute graded stability. They instead provide an external behavioral perturbation, allowing us to ask whether representation-level stability generalizes to a qualitatively different form of belief pressure.

We construct pairs of believed propositions with similar $\pi_T(P)$ within each model and domain using sequence-blocked matching (Sec. 4.6, G). We match propositions by maximizing the number of non-overlapping pairs while minimizing their total absolute difference in belief probability. We then identify the propositions P_H and P_L with higher and lower graded stability within each pair. Our primary analysis retains pairs for which both propositions’ initial behavioral judgments agree with their individual-statement probe classifications.

We evaluate each proposition in a four-stage interaction consisting of an initial binary True/False judgment followed by three natural-language challenges. Propositions are randomly assigned one of six possible challenge orderings to control for order effects (Sec. 4.6). The challenges model conversational pressure from a disagreeing user rather than literal factual conditioning on another proposition x . They therefore provide external behavioral validation of graded stability rather than an empirical realization of $\Pr_{\mathcal{M}}(P | x)$.

We record the normalized probability $u_0(P)$ assigned to the model’s initial judgment and denote the probability assigned to that same response after k challenges by $u_k(P)$. We quantify behavioral movement as

$$M(P) := \frac{1}{3} \sum_{k=1}^3 |u_k(P) - u_{k-1}(P)|, \quad (14)$$

Higher $\gamma(P)$ is associated with greater resistance to challenge

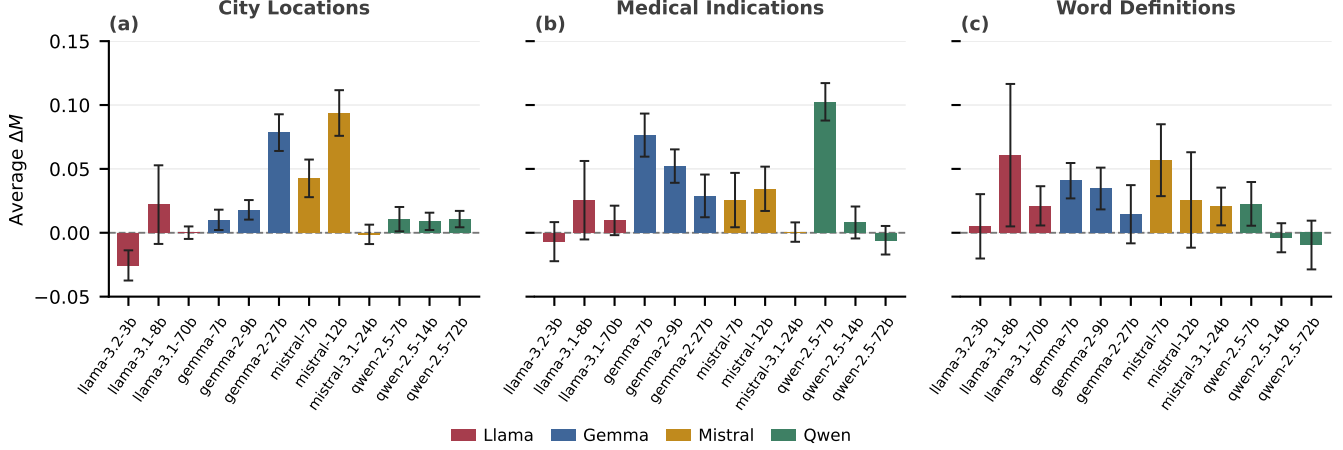


Figure 5. Behavioral resilience among beliefs matched on individual belief probability. Bars show the mean difference in behavioral movement ΔM between matched lower- and higher-stability beliefs for each of the 12 instruction-tuned LLMs, with error bars denoting ± 1 standard error. Positive values indicate that lower-stability beliefs exhibit greater movement under conversational challenge. Results are shown for (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) model families. Across domains, 83.3% of model-domain settings show greater behavioral movement for lower-stability beliefs.

where larger $M(P)$ indicates greater fluctuation in support for the model’s initial response. Because the differences are absolute, increases and decreases in support both contribute to $M(P)$. It therefore measures behavioral movement rather than specifically loss of support or answer flipping.

For each probability-matched pair, we then compare the behavioral movement of the lower- and higher-stability propositions:

$$\Delta M := M(P_L) - M(P_H). \quad (15)$$

Thus, $\Delta M > 0$ indicates that the lower-stability belief exhibits greater behavioral movement under challenge. Because P_H and P_L are identified only after matching on individual belief probability, this comparison tests whether graded stability distinguishes behavioral resilience beyond differences in $\pi_T(P)$.

Mean ΔM is positive for 10/12 models in City Locations, 10/12 in Medical Indications, and 10/12 in Word Definitions. Median effects are 0.011, 0.026, and 0.022, respectively. However, uncertainty is substantial for many individual estimates: the 95% bootstrap confidence intervals are entirely positive in 11/36 settings, entirely negative in one, and include zero in the remainder (Sec. G.2). Thus, graded stability shows a broadly consistent directional association with behavioral resilience, although the magnitude and precision of the effect vary across models and beliefs.

3 Discussion

This work asks whether the stability of an LLM judgment can be characterized by its relationship to the model’s other beliefs rather than by individual belief probability alone. Across 12 instruction-tuned LLMs and three semantic domains, individual belief probability explains a meaningful but incomplete fraction of graded stability. The remaining variation is also not arbitrary. After accounting for individual belief probability, different models tend to identify the same propositions as unusually stable or fragile, and among beliefs deliberately matched on individual belief probability, lower graded stability is generally associated with greater movement under conversational challenge. These results support the central distinction motivating graded stability: how strongly an LLM supports a proposition in isolation and how robustly that support persists in the context of its other beliefs are related, but non-equivalent, properties.

A central implication of our results is that this relational structure depends strongly on what is believed. City Locations is the most stable domain for 11 of the 12 instruction-tuned models, and the same broad domain structure appears in matched pretrained base models and under the alternative SVM probe (Secs. I.1.1, I.3). By contrast, neither

parameter count nor model depth shows a consistent relationship with graded stability across domains (Sec. H). Instruction tuning, however, is associated with greater stability in most matched base versus instruction-tuned comparisons, but the magnitude and even direction of this difference vary across models and domains (Sec. I.1.2). These results argue against treating graded stability as a generic property that simply increases with model capability. Instead, stability appears to reflect an interaction among the semantic structure of the belief, the model in which it is represented, and the post-training regime applied to that model. A key next step is therefore to identify the semantic properties that systematically confer stability or fragility, and to determine whether those properties generalize beyond the controlled domains studied here.

Our robustness analyses further suggest that the most reproducible feature of graded stability is its *relative* structure rather than its absolute numerical scale. The Direct Conditional and Joint-to-Conditional estimators operationalize conditional belief probability in different ways, yet they produce positively correlated stability rankings in every setting (Sec. I.2.3). At the same time, their absolute stability magnitudes differ substantially, with Joint-to-Conditional scores compressed toward 1.0 (Sec. I.2). Absolute graded-stability values, and differences in their magnitudes across domains, should therefore be interpreted within a particular estimator and probe rather than as measurement-invariant quantities. Probe choice produces a related distinction. The SVM probe broadly reproduces the main findings obtained with sAwMIL, whereas the Mass Mean probe yields weaker domain separation, larger departures from probabilistic coherence, and less consistent behavioral associations (Sec. I.3). This weaker replication is not surprising because Mass Mean is not designed to recover calibrated trivalent probability distributions, which graded stability explicitly requires. These analyses suggest that a reproducible relational signal persists across operationalizations, even though its absolute numerical realization depends on the conditional estimator and probe. Faithfully recovering that signal therefore depends on measurement choices that preserve the probabilistic properties required by the construct.

The behavioral challenge experiment tests whether graded stability distinguishes resistance to repeated conversational challenge among beliefs with similar individual belief probabilities. Among beliefs matched on individual belief probability, lower-stability beliefs exhibit greater mean movement under repeated challenge in 83.3% of instruction-tuned model-domain settings. The consistency of this direction across models is notable, but the individual effects are heterogeneous and often imprecisely estimated: most 95% bootstrap intervals include zero (Sec. G.2). We consequently view the behavioral experiment as evidence that graded stability captures differences in resilience beyond individual belief probability, rather than as evidence for a universal mapping from γ to conversational behavior. More broadly, resistance to revision is not always desirable. Instability can indicate unwanted fragility, but revision may be appropriate when new information exposes an error or creates a legitimate conflict with existing commitments. Conversely, an incorrect judgment that resists all counterevidence would be highly stable without being reliable. Graded stability is therefore best understood as a descriptive property of an LLM belief system rather than a score to optimize. Distinguishing unwarranted fragility from appropriate responsiveness will require manipulating the relevance, reliability, and evidential force of the information introduced to the model.

Our results also illustrate the value and limitations of applying formal probabilistic theories to LLM representations. LLMs are autoregressive sequence models, not agents explicitly trained to maintain a globally coherent probability system over propositions, and we find correspondingly little exact CCK coherence. However, the measured individual belief and conditional probability distributions generally require only modest adjustments to reach the nearest coherent system. Formal coherence therefore does not have to be treated as an all-or-nothing claim about whether an LLM literally instantiates a particular epistemic theory. It can instead provide a reference structure against which empirical representations are measured. More generally, the calibrated probe probabilities used throughout should be understood as representation-based estimates of epistemic support rather than direct observations of a uniquely defined latent credence distribution of the LLM. From this perspective, approximate coherence makes the formal framework useful, while departures from coherence become objects of study rather than reasons to abandon it. The same principle applies to the definition of graded stability itself. Our formulation weights each admissible conditioning proposition equally and adopts Cooper–Cantwell–Kleene semantics for trivalent conditionals. However, alternative weighting schemes or conditional semantics would define related but distinct notions of stability. Understanding which of these notions best predicts particular forms of model behavior is an important direction for future work.

Finally, our experiments deliberately use controlled factual domains containing explicit **True**, **False**, and **Neither** cases. This structure is essential to the present study as it provides comparatively well-defined trivalent supervision, permits systematic construction of conditional training examples, and supplies propositional objects that map cleanly onto the formal account of belief and conditioning. The same control limits ecological scope. Accordingly, differences in γ across these domains reflect both the target propositions and the composition of the benchmark-specific

Official Name	Short Name	# Layers	# Parameters	Release Date	Source	Citation
Gemma-7b-it	<code>gemma-7b</code>	28	8.54 B	Feb 21, 2024	Google	[34]
Gemma-2-9b-it	<code>gemma-2-9b</code>	42	9.24 B	Jun 27, 2024	Google	[34]
Gemma-2-27b-it	<code>gemma-2-27b</code>	46	27.23 B	Jun 27, 2024	Google	[34]
Llama-3.2-3b-Instruct	<code>llama-3.2-3b</code>	28	3.21 B	Sep 25, 2024	Meta	[35]
Llama-3.1-8b-Instruct	<code>llama-3.1-8b</code>	32	8.03 B	Jul 23, 2024	Meta	[35]
Llama-3.1-70b-Instruct	<code>llama-3.1-70b</code>	80	70.55 B	Jul 23, 2024	Meta	[35]
Mistral-7b-Instruct-v0.3	<code>mistral-7b</code>	32	7.25 B	May 22, 2024	Mistral AI	[36]
Mistral-Nemo-Instruct-2407	<code>mistral-12b</code>	40	12.25 B	Jul 18, 2024	Mistral AI	[37]
Mistral-Small-3.1-24B-Instruct-2503	<code>mistral-3.1-24b</code>	40	23.57 B	Mar 17, 2025	Mistral AI	[38]
Qwen2.5-7B-Instruct	<code>qwen-2.5-7b</code>	28	7.62 B	Sep 19, 2024	Alibaba Cloud	[39, 40]
Qwen2.5-14B-Instruct	<code>qwen-2.5-14b</code>	48	14.80 B	Sep 19, 2024	Alibaba Cloud	[39, 40]
Qwen2.5-72B-Instruct	<code>qwen-2.5-72b</code>	80	72.70 B	Sep 19, 2024	Alibaba Cloud	[39, 40]

Table 1. LLMs used in graded stability experiments. We list the official names of the LLMs according to the HuggingFace repository [41], where they are publicly available. We further specify the shortened name used throughout the paper, the number of layers, parameter count, release date, source organization, and official citation.

non-disbelief sets over which stability is evaluated, rather than an intrinsic stability of the semantic domain alone. Rich natural-language claims may contain multiple propositions, presuppositions, context-sensitive content, or ambiguous conditioning relations, and real conversational contexts need not resemble the within-domain non-disbelief sets studied here. Extending graded stability to richer benchmarks and cross-domain belief systems will therefore require more than simply applying the same pipeline to longer text; it will require defining what the relevant propositions and conditioning relations are.

What remains consistent across these methodological and formal choices is the distinction between the strength of an individual belief and its stability within a broader system of beliefs. Evaluations of factual reliability typically ask whether a model’s answer is correct, how strongly it is supported, or whether confidence tracks accuracy. Our results show that relationships among beliefs provide complementary information. Propositions with similar individual support can differ systematically in stability, different models share substantial structure in which propositions are unusually stable or fragile, and, under our primary operationalization, these differences are associated with downstream behavioral resilience. Graded belief stability therefore provides a relational view of factual reliability that complements existing measures of accuracy and uncertainty. Characterizing this structure may help explain not only what an LLM appears to believe, but how those beliefs are situated within a broader epistemic system and which are most susceptible to change as their informational context shifts.

4 Methods

We first describe the LLMs, datasets, and probing framework used throughout our experiments. We then detail how we estimate individual statement probabilities, construct empirical belief and non-disbelief sets, and measure conditional probabilities. Finally, we describe the analyses used to validate graded stability, characterize its variation across domains and models, and evaluate its relationship with behavioral resilience under conversational challenge. All code and datasets associated with the experiments can be found at https://github.com/samanthadies/graded_stability.

4.1 Models and data

4.1.1 Large language models

We evaluate 12 open-weight, instruction-tuned LLMs from four model families: **Gemma**, **Llama**, **Mistral**, and **Qwen**. For **Gemma**, we consider 7B, 9B, and 27B. For **Llama**, we use 3B, 8B, and 70B. For **Mistral**, we use 7B, 12B, and 24B. Finally, for **Qwen** we use 7B, 14B, and 72B. We report the exact model versions and the number of decoder layers in Table 1. Matched base models are listed in Section C. All models are loaded in evaluation mode with CUDA and BF16. We render statements using the models’ native chat templates as single user messages without an assistant response.

4.1.2 Datasets

Our datasets, introduced in [20], include three domains: City Locations, Medical Indications, and Word Definitions (Tab. 2). Each domain contains **True**, **False**, and **Neither** statements. The **Neither** statements pair synthetic entities generated using character n -gram Markov-chain models fit to the corresponding real entity sets and validated

Dataset	True	False	Synthetic	Examples
City Locations	A: 1392 N: 1376	A: 1358 N: 1374	A: 876 N: 876	T. The city of Surat is located in India. F. The city of Palembang is located in the Dominican Republic. S. The city of Norminsk is located in Jamoates.
Medical Indications	A: 1439 N: 1522	A: 1523 N: 1419	A: 478 N: 522	T. Pentobarbital is indicated for the treatment of insomnia. F. Vancomycin is not indicated for the treatment of lower respiratory tract infections. S. Alumil is indicated for the treatment of reticers.
Word Definitions	A: 1234 N: 1235	A: 1277 N: 1254	A: 1747 N: 1753	T. Hoagy is a synonym of an Italian sandwich. F. Decalogue is an astronomer. S. Dostab is a scencer.

Table 2. Summary of datasets and statement types. Number of affirmative (A) and negated (N) statements across the three domains, along with examples. Each dataset includes **True** (T), **False** (F), and **Synthetic** (S) statements. **Synthetic** statements serve as **Neither** statements constructed to minimize prior LLM exposure. An earlier version of this table was introduced in [20].

to reduce the likelihood that they correspond to existing entities. They therefore provide controlled cases for which an LLM should ideally suspend belief, making the datasets well suited to our trivalent belief framework. These synthetic cases instantiate a controlled form of epistemic indeterminacy associated with unfamiliar content. However, they do not exhaust the ways in which an LLM may suspend judgment in natural-language settings. Further dataset construction and validation details are provided in Section B.1.

We use the fixed training, calibration, and test partitions introduced with the original datasets [20] (Sec. B.2), containing approximately 55%, 20%, and 25% of statements, respectively, with entities kept exclusive across splits. The training split is used to fit each probe, while the calibration split is used to map raw probe scores to calibrated probability distributions for individual statements, $\pi_{\mathcal{M}}(s_i)$, and Direct Conditional statement pairs, $\pi_{\mathcal{M}}^{\text{direct}}(s_i, s_j)$, according to the procedure in Section D.1. The fitted probes and calibration mappings are then applied to test statements and statement pairs.

The statements in the train and calibration sets are reused to train and calibrate the individual-statement probe and the Direct Conditional probe. The test set statements form the basis of our belief and non-disbelief sets $\mathcal{B}_{\mathcal{M}}$ and $\mathcal{X}_{\mathcal{M}}$.

4.1.3 Probes

Each statement s_i is fed into a model $\mathcal{M}$, resulting in an activation $z_i^{(\ell)}$ at each layer ℓ . We train each probe over the activations corresponding to the training statements, using the ground-truth **True**, **False**, and **Neither** labels as training targets. The probes produce raw score vectors $\mathbf{r}(s_i)$ over the K classes. Because these raw scores cannot be directly interpreted as probabilities, we fit a multinomial logistic-regression calibration model on the calibration split that maps each raw score vector $\mathbf{r}(s_i)$ to a calibrated class-probability distribution $\pi_{\mathcal{M}}(s_i)$ over the K classes (Sec. D.1). Complete training specifications and hyperparameters for **sAwMIL** are reported in Section D.2, and its layer-selection procedure and selected layers are reported in Section D.3. Corresponding details for the secondary **SVM** and **Mass Mean** probes are provided in Section I.3. Throughout, we treat these calibrated probe probabilities as representation-based estimates of the model’s epistemic support over the supervised trivalent states. Calibration makes the probe outputs probabilistically interpretable with respect to those states, but does not by itself establish that they are identical to a uniquely defined latent credence distribution of the LLM.

Primary probe: sAwMIL Our main probe is the sparse-aware multiple-instance learning (**sAwMIL**) probe [20]. **sAwMIL** is a max-margin probe that is trained over a bag of token-level representations for each statement, rather than a single statement-level vector. It was also designed specifically to handle more than two classes. For K classes, **sAwMIL** trains K separate one-versus-all heads which are combined into a multiclass classifier using a softmax function.

Secondary probes: SVM and Mass Mean We use two additional probes, an **SVM** and the **Mass Mean** probe [17], to test whether our stability results are qualitatively robust to the choice of probe. The corresponding analyses are reported in Sec. I.3. Like **sAwMIL**, the **SVM** is a max-margin classifier that learns K one-versus-all heads. However, **sAwMIL** is a multi-instance learning probe, while the **SVM** operates only on the activation corresponding to a given statement’s final token. The **Mass Mean** probe is also a single-instance probe that operates on the activation corresponding to a given statement’s final token. It was initially designed as a binary probe. However, we extend it to a multiclass probe so that it can be used in our three-class, trivalent probability setting. For each class $j \in \{0, \dots, K-1\}$, the **Mass Mean** probe takes the mean representation of the training examples in class j , μ_j , the mean representation of

the examples in all other classes, $\mu_{\neg j}$, and computes the difference vector $\Delta\mu_j = \mu_j - \mu_{\neg j}$. It then normalizes $\Delta\mu_j$ to the unit norm and places the decision boundary halfway between μ_j and $\mu_{\neg j}$. Statements are scored based on their signed projection along the $\Delta\mu_j$ vector. This again results in K one-versus-all heads. Outputs from both **SVM** and **Mass Mean** undergo the same calibration step used for **sAwMIL**.

4.2 Estimating probabilities for individual statements and constructing empirical belief sets

For each model $\mathcal{M}$, dataset, and probe, we train a three-class probe independently at each layer ℓ using the training split and fit its probability-calibration mapping using the calibration split. We then evaluate the calibrated probe at each layer on the fixed calibration partition $\mathcal{D}_{\text{cal}}$ using multiclass log loss,

$$\mathcal{L}_{\log}^{(\ell)} = -\frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{(s_i, y_i) \in \mathcal{D}_{\text{cal}}} \log \pi_{y_i}^{(\ell)}(s_i), \quad (16)$$

and select

$$\ell^* = \arg \min_{\ell} \mathcal{L}_{\log}^{(\ell)}. \quad (17)$$

Section D.3 reports the full layer-selection procedure and the resulting selected layers for **sAwMIL**. Corresponding results for the secondary probes are reported in Section I.3.

After selecting ℓ^* , we retain the fitted and calibrated probe from that layer and use its outputs on the test statements throughout the downstream analysis rather than retraining the probe after layer selection. For each statement, we record both $\pi_{\mathcal{M}}(s_i) = [\pi_T(s_i), \pi_F(s_i), \pi_N(s_i)]$ and the predicted state $\hat{y}_i = \arg \max_{c \in \{T, F, N\}} \pi_c(s_i)$. These predictions define the empirical belief and non-disbelief sets used in all subsequent analyses.

For each test statement, we map its predicted class $\hat{y}_i$ to a belief state, where $\hat{y}_i = T$ is a belief, $\hat{y}_i = F$ is a disbelief, and $\hat{y}_i = N$ is a suspended belief. For clarity, we restate the empirical belief and non-disbelief sets defined in Equation (5):

$$\mathcal{B}_{\mathcal{M}} = \{s_i : \hat{y}_i = T\}, \quad \mathcal{X}_{\mathcal{M}} = \{s_i : \hat{y}_i \in \{T, N\}\}.$$

Crucially, this is based on the probe’s prediction, not the statement’s ground-truth label, meaning that a **False** or **Neither** statement could be classified as a model belief. Thus, $\mathcal{B}_{\mathcal{M}} \subseteq \mathcal{X}_{\mathcal{M}}$. Section E.1 reports the resulting belief and non-disbelief set sizes, empirical thresholds, and conditional-pair coverage for each LLM–dataset combination. From these belief and non-disbelief sets, we construct $|\mathcal{X}_{\mathcal{M}}| - 1$ statement pairs for each $P \in \mathcal{B}_{\mathcal{M}}$ of the form (P, x) . These statement pairs are the target of our conditional probability estimators described in Section 4.3 and ultimately define graded stability as described in Section 4.4.

For each LLM, probe, and dataset combination, we use the empirical belief threshold from Equation (6), restated for clarity:

$$t_{\mathcal{M}} = \min_{s_i \in \mathcal{B}_{\mathcal{M}}} \pi_T(s_i).$$

This belief threshold is derived after categorical beliefs have been identified and corresponds to the smallest **True**-class probability among statements classified by the probe as **True**. This guarantees that the probability $\pi_T(P_i)$ that statement P_i is true is at least $t_{\mathcal{M}}$ for all of model $\mathcal{M}$ ’s beliefs $P_i \in \mathcal{B}_{\mathcal{M}}$. The strict persistence criterion $> t_{\mathcal{M}}$ is separate from this categorical membership rule: a conditional estimate exactly equal to $t_{\mathcal{M}}$ is counted as non-persistent, while membership in $\mathcal{B}_{\mathcal{M}}$ remains determined by the trivalent probe prediction. Moreover, conditional persistence is determined by **True**-state support relative to $t_{\mathcal{M}}$ rather than by the conditional probe’s argmax state, so a conditional can satisfy the persistence criterion even when **True** is not its most probable class.

4.3 Estimating conditional probability

For each belief $P \in \mathcal{B}_{\mathcal{M}}$ and non-disbelief $x \in \mathcal{X}_{\mathcal{M}} \setminus \{P\}$, we aim to estimate $\Pr_{\mathcal{M}}(P | x)$ so that we can compute $\gamma_{\mathcal{M}}(P)$, the proportion of conditioning propositions x for which belief in P persists (i.e., $\Pr_{\mathcal{M}}(P | x) > t_{\mathcal{M}}$). Our primary operationalization, **Direct Conditional**, represents $P | x$ directly as a conditional statement and probes for its veracity. Our secondary approach, **Joint-to-Conditional**, represents x and P jointly, probes their full 3×3 joint states, then derives the conditional mathematically. All main-text analyses use the Direct Conditional estimator. We use Joint-to-Conditional as a secondary estimator to test robustness to the conditional-probability operationalization (Sec. E.2, I.2).

Both approaches use the same general probing framework as the estimation of probabilities for individual statements described in Section 4.2, including the data splits and probability-calibration procedure. However, we construct new conditional or joint training and calibration examples from the existing train and calibration sets, use new training targets for the combined statements, and score LLM, probe, and dataset-specific (P, x) pairs. The Direct

s_i	s_j	$s_i \rightarrow s_j$
T	T	T
T	N	N
T	F	F
N	T	T
N	N	N
N	F	F
F	T	N
F	N	N
F	F	N

Table 3. Cooper–Cantwell conditional truth table. We report whether the conditional $s_i \rightarrow s_j \equiv s_j \mid s_i$ is **True** (T), **False** (F), or **Neither** (N) based on the truth values of s_i and s_j according to the Cooper–Cantwell conditional [28, 31]. When the antecedent s_i is **True** or **Neither**, the conditional takes the truth value of the consequent s_j .

Conditional and Joint-to-Conditional probes score the same (P, x) pairs, allowing us to directly compare the two estimators and evaluate the robustness of graded stability to the underlying conditional-probability operationalization (Sec. 1.2.3). We describe the specifics of the two estimators below.

4.3.1 Primary estimator: Direct Conditional

The **Direct Conditional** approach represents $P \mid x$ itself as a trivalent epistemic object and asks how strongly its LLM representation supports the **True** state. We operationalize $P \mid x$ using the statement template “*Given x , P .*”, where x is the antecedent, or conditioning proposition, and P is the consequent, or belief whose persistence we aim to measure. Probabilistic accounts of indicative conditionals have long related support for a conditional to the corresponding conditional probability [42]. We instantiate the conditional itself using Cooper–Cantwell semantics [28, 31, 32].

Under Cooper–Cantwell semantics, when the antecedent is **True** or **Neither**, the conditional inherits the truth value of the consequent; when the antecedent is **False**, the conditional is **Neither** (Tab. 3). This treatment is particularly well suited to our stability framework because the admissible conditioning set $\mathcal{X}_{\mathcal{M}}$ contains both beliefs and suspended beliefs. A suspended antecedent can therefore provide an admissible conditioning context without automatically rendering the conditional indeterminate, as it would in the de Finetti conditional [43].

To train the Direct Conditional probe, we start from the same train and calibration statements used for the individual-statement probe and construct conditional statements of the form “*Given s_i , s_j .*” from statement pairs. Training labels are assigned from the ground-truth states of s_i and s_j according to the Cooper–Cantwell truth table. We stratify the resulting examples across the nine input-state combinations, yielding 6750 training statements and 2250 calibration statements. We then train and calibrate a three-class probe over the **True**, **False**, and **Neither** conditional states.

As defined in Equation (7), the calibrated conditional probe returns

$$\pi_{\mathcal{M}}^{\text{direct}}(P, x) = [\pi_T^{\text{direct}}(P, x), \pi_F^{\text{direct}}(P, x), \pi_N^{\text{direct}}(P, x)]$$

for each pair (P, x) . The Direct Conditional estimator is then defined as

$$\widehat{\text{Pr}}_{\mathcal{M}}^{\text{direct}}(P \mid x) := \pi_T^{\text{direct}}(P, x).$$

This quantity is the calibrated probability mass assigned to the conditional’s **True** epistemic state. We use it as an empirical operational estimator of $\text{Pr}_{\mathcal{M}}(P \mid x)$. It should not be read as an algebraic identity with the normalized non-bivalent probability of a trivalent conditional. We deliberately retain the **Neither** mass rather than renormalizing over the **True** and **False** states. Our empirical belief threshold $t_{\mathcal{M}}$ is defined on the same raw **True**-class mass for individual propositions. Using π_T^{direct} therefore asks whether the conditional representation retains enough **True** support to satisfy the same empirical belief criterion. In particular, movement from **True** toward **Neither** is allowed to count against persistence of the original belief.

This choice differs from the standard non-bivalent probability assigned to a trivalent conditional, which normalizes **True** mass over the states in which the conditional is determinate [27, 31]. We evaluate that complementary construction through our Joint-to-Conditional estimator below.

We retain the Direct Conditional as our primary operationalization because it probes the conditional epistemic object directly, preserves all three epistemic states used to define the individual belief system, and requires only a three-class measurement problem rather than first recovering a nine-state joint distribution.

4.3.2 Secondary estimator: Joint-to-Conditional

The **Joint-to-Conditional** approach provides a complementary operationalization grounded in the normalized probability of a trivalent conditional. Rather than probing $P | x$ directly, we first estimate the joint epistemic state of x and P . To construct the joint inputs, we use the template “ x and P .”, with the conditioning proposition x appearing first and the target belief P second. Because the probe operates on LLM representations of linguistic sequences, the resulting estimates are not invariant to conjunction order: “ x and P ” and “ P and x ” produce distinct representations and, empirically, distinct probability estimates. We therefore fix the order to “ x and P ” throughout the primary Joint-to-Conditional analysis and report robustness to the alternative ordering in Section I.2.1.

As with the Direct Conditional approach, training the joint probe requires new labels for combined statements. During training and calibration, these combined statements are constructed from generic statement pairs (s_i, s_j) using the template “ s_i and s_j .”. Because each statement can take on one of three truth values, the corresponding joint label has $3 \times 3 = 9$ possible states: $[TT, TF, TN, NT, NF, NN, FT, FF, FN]$, represented formally as $(y_i, y_j) \in \{T, F, N\} \times \{T, F, N\}$.

We construct the combined training and calibration sets by subsampling the individual-statement training and calibration sets such that each of the nine joint classes has an equal number of combined statements. In total, we again generate 6750 training statements and 2250 calibration statements. We then train the probe with nine one-versus-all heads and calibrate its outputs as outlined in Section 4.1.3 and D.1.

After training and calibration, we apply the joint probe to the same (P, x) pairs scored by the Direct Conditional probe, with x occupying the first position and P the second. The resulting calibrated nine-valued distribution is

$$\pi_{\mathcal{M}}^{\text{joint}}(P, x) = \{\pi_{ab}^{\text{joint}}(P, x)\}_{a,b \in \{T, F, N\}}. \quad (18)$$

Here, $\pi_{ab}^{\text{joint}}(P, x)$ is the probe-assigned probability that x has state a and P has state b . To derive the conditional probability from this joint distribution, we use the probabilistic semantics for Cooper–Cantwell trivalent conditionals [27, 31, 44]:

$$\widehat{\text{Pr}}_{\mathcal{M}}^{\text{joint}}(P | x) = \frac{\pi_{TT}^{\text{joint}}(P, x) + \pi_{NT}^{\text{joint}}(P, x)}{\pi_{TT}^{\text{joint}}(P, x) + \pi_{TF}^{\text{joint}}(P, x) + \pi_{NT}^{\text{joint}}(P, x) + \pi_{NF}^{\text{joint}}(P, x)}. \quad (19)$$

Under Cooper–Cantwell semantics, the conditional is determinate when the antecedent x is **True** or **Neither** and the consequent P is either **True** or **False**. The numerator therefore contains the joint states that induce a **True** conditional, while the denominator contains all joint states that induce either a **True** or **False** conditional. Probability mass assigned to joint states that induce a **Neither** conditional is excluded by this normalization. We provide the full derivation in Section E.2.

This normalization marks the principal distinction from our Direct Conditional estimator. Direct Conditional retains probability mass assigned to the conditional’s **Neither** state, such that movement from **True** toward **Neither** can reduce the estimated support for $P | x$. Joint-to-Conditional instead conditions on the event that the induced conditional is determinate and measures the relative **True** mass within that subset.

Equation (19) is undefined when its denominator is zero. We treat denominators less than or equal to 10^{-12} as numerically zero and exclude the corresponding (P, x) pairs from subsequent graded-stability calculations rather than assigning them a conditional probability. Across all models and datasets, this occurs for 0.026% of scored Joint-to-Conditional pairs.

4.4 Computing graded belief stability

For each belief $P \in \mathcal{B}_{\mathcal{M}}$, we evaluate its conditional probability under every non-disbelief $x \in \mathcal{X}_{\mathcal{M}}$. We exclude the self-pair $P = x$, resulting in $|\mathcal{X}_{\mathcal{M}}| - 1$ conditioning propositions for each belief. For completeness, we restate the graded belief stability definition from Equation (2):

$$\gamma_{\mathcal{M}}(P) := \frac{|\{x \in \mathcal{X}_{\mathcal{M}} \setminus \{P\} : \text{Pr}_{\mathcal{M}}(P | x) > t_{\mathcal{M}}\}|}{|\mathcal{X}_{\mathcal{M}} \setminus \{P\}|},$$

where $t_{\mathcal{M}}$ is the empirical belief threshold obtained from the corresponding individual-statement probe described in Section 4.2. Thus, $\gamma_{\mathcal{M}}(P)$ is the proportion of conditioning propositions under which the conditional probability

of P remains above the belief threshold. We apply Equation (2) using the Direct Conditional estimates, yielding $\gamma_{\mathcal{M}}^{\text{direct}}(P)$.⁴

Each $\gamma_{\mathcal{M}}(P)$ is a proposition-level quantity. When a model-level summary is required, we compute the mean graded stability across believed propositions,

$$\bar{\gamma}_{\mathcal{M}} = \frac{1}{|\mathcal{B}_{\mathcal{M}}|} \sum_{P \in \mathcal{B}_{\mathcal{M}}} \gamma_{\mathcal{M}}(P). \quad (20)$$

We compute this quantity separately for each model, dataset, probe, and conditional-probability estimator.

4.5 Validating graded belief stability

We evaluate graded belief stability along two complementary dimensions: whether it captures information beyond individual belief probability (Sec. 4.5.1) and whether the measured probabilities are approximately compatible with the CCK probability framework (Sec. 4.5.2).

4.5.1 Stability vs. belief probability

As defined in Equation (9), we model graded stability $\gamma_{\mathcal{M}}(P)$ as a function of the belief probability $\pi_T(P)$ using a natural cubic spline,

$$\hat{\gamma}_{\mathcal{M}}(P) = f_{\mathcal{M}}(\pi_T(P)),$$

for each model, dataset, probe, and conditional-probability estimator. We fix the spline complexity at three degrees of freedom for all primary analyses. Section F.1 evaluates alternative spline complexities and a linear baseline and provides additional implementation and fit diagnostics. We use five-fold cross-validation to evaluate predictive performance on held-out propositions and quantify performance using R^2 ,

$$R^2 = 1 - \frac{\sum_{P \in \mathcal{B}_{\mathcal{M}}} (\gamma_{\mathcal{M}}(P) - \hat{\gamma}_{\mathcal{M}}(P))^2}{\sum_{P \in \mathcal{B}_{\mathcal{M}}} (\gamma_{\mathcal{M}}(P) - \bar{\gamma}_{\mathcal{M}})^2}, \quad (21)$$

where $\bar{\gamma}_{\mathcal{M}}$ is the mean graded stability defined in Equation (20). Model–dataset–probe–estimator combinations with fewer than 50 believed propositions with valid stability scores, insufficient distinct values of $\pi_T(P)$ to fit the spline, or no variation in $\gamma_{\mathcal{M}}(P)$ are excluded from the analysis.

The fitted value $\hat{\gamma}_{\mathcal{M}}(P)$ represents stability predicted from belief probability alone. Using these out-of-fold predictions, we define residual stability as in Equation (10), restated here for clarity:

$$r_{\mathcal{M}}(P) = \gamma_{\mathcal{M}}(P) - \hat{\gamma}_{\mathcal{M}}(P),$$

where a positive residual indicates that a proposition is more stable than predicted from its belief probability, while a negative residual indicates that it is less stable than predicted.

We use these residuals to test whether propositions that are unusually stable or fragile for one model tend to be similarly unusual for other models. Within each dataset, probe, and conditional-probability estimator, we consider every pair of models and restrict the comparison to propositions believed by both models. We retain model pairs with at least 50 shared propositions and compute Spearman’s rank correlation [45] between their residual stability scores over this shared proposition set, yielding one residual correlation for each model pair. We characterize cross-model agreement using the median of these pairwise correlations.

To assess whether the observed agreement exceeds what would be expected without proposition-specific alignment across models, we construct a permutation null by independently shuffling residual stability scores across propositions within each model while preserving each model’s residual distribution, the set of eligible model pairs, and their proposition-overlap structure. For each permutation, we recompute the pairwise residual Spearman correlations and record their median. We use 1,000 permutations and compute a one-sided Monte Carlo p -value to evaluate the hypothesis that the observed median correlation is greater than expected under the null, using the +1 correction [46]. Additional implementation details are reported in Section F.1.

⁴We analogously compute $\gamma_{\mathcal{M}}^{\text{joint}}(P)$ for the Joint-to-Conditional estimate and report results in Section I.2.2.

4.5.2 Approximate probabilistic coherence

We evaluate probabilistic coherence under the Cooper–Cantwell–Kleene (CCK) trivalent probability framework [27, 31] for each model, dataset, probe, and conditional-probability estimator.⁵ For every (P, x) pair, we combine the measured probability distributions for P , x , and their conditional into the vector $\mathbf{v}_{P,x}$ defined in Equation (11), restated here for clarity:

$$\mathbf{v}_{P,x} = [\pi_{\mathcal{M}}(P), \pi_{\mathcal{M}}(x), \pi_{\mathcal{M}}^{\text{direct}}(P, x)].$$

As established in Equation (12), for the measured marginals of P , x , and their conditional to be jointly representable by a single trivalent distribution over (P, x) , exact CCK coherence requires

$$\pi_T^{\text{direct}}(P, x) \leq \pi_T(P), \quad \pi_F^{\text{direct}}(P, x) \leq \pi_F(P), \quad \pi_F(x) \leq \pi_N^{\text{direct}}(P, x), \quad \pi_N^{\text{direct}}(P, x) \leq \pi_N(P) + \pi_F(x).$$

We classify a pair as exactly coherent when all four CCK constraints are satisfied within a numerical tolerance of 10^{-8} . A proof that these conditions characterize exact pairwise coherence is provided in Section F.2.1, and exact-coherence rates and individual constraint violations are reported in Section F.2.2.

To distinguish small departures from coherence from larger violations, we additionally quantify each pair’s distance from the coherent set using d_{CCK} as defined in Equation (13):

$$d_{\text{CCK}}(\mathbf{v}_{P,x}) := \sqrt{\frac{1}{9} \|\mathbf{v}_{P,x} - \Pi_{\mathcal{C}}(\mathbf{v}_{P,x})\|_2^2}.$$

We compute the corresponding projection $\Pi_{\mathcal{C}}(\mathbf{v}_{P,x})$ using Dykstra’s projection algorithm [47]. The algorithm alternates between enforcing normalization and non-negativity of the three probability distributions and enforcing each of the four linear CCK constraints. Iteration stops when the maximum component-wise change between successive solutions is at most 10^{-10} , with a maximum of 250 iterations. We verify that the projected vectors satisfy normalization, non-negativity, and the CCK constraints within a tolerance of 10^{-7} . We compute d_{CCK} for every (P, x) pair and summarize its distribution separately for each model, dataset, probe, and conditional-probability estimator.

4.6 Behavioral validation under conversational challenge

To test whether graded stability corresponds to observable belief resilience, we measure how strongly an LLM maintains its initial truth judgment across a controlled sequence of conversational challenges. In particular, we test whether, among beliefs assigned similar individual belief probabilities, the belief with higher graded stability exhibits less behavioral movement under challenge.

We first construct the proposition pairs used for this comparison by matching beliefs on individual belief probability. Before matching, propositions are assigned to one of the six challenge-order sequences described below. To ensure that both members of a matched pair receive the same conversational challenges in the same order, we perform matching separately within each exact challenge-sequence block. Within each block s , we construct a maximum-cardinality, one-to-one matching $\mathcal{M}_s$ without replacement that minimizes the total absolute difference in individual belief probability,

$$\sum_{(P_i, P_j) \in \mathcal{M}_s} |\pi_T(P_i) - \pi_T(P_j)|. \quad (22)$$

We restrict the eligible proposition pool to beliefs for which both Direct Conditional and Joint-to-Conditional graded stability are defined, so that the two operationalizations are evaluated over the same underlying matches. Within this eligible pool, pair construction depends only on individual belief probability and challenge-sequence assignment. Neither graded-stability values nor behavioral outcomes enter the matching objective. After the matches are fixed, we orient each pair according to the relevant graded-stability estimate and denote the higher- and lower-stability propositions by P_H and P_L , respectively. Pairs for which the two propositions have identical graded stability are excluded because no higher- and lower-stability member can be defined. An overview of the resulting probability-matched pairs is reported in Section G.1.

Each matched proposition is evaluated in a four-stage interaction consisting of an initial binary truth judgment at round $k = 0$ followed by three successive challenges at rounds $k \in \{1, 2, 3\}$. We use the following prompts:

⁵For Joint-to-Conditional, the distributions $\pi_{\mathcal{M}}(P)$ and $\pi_{\mathcal{M}}(x)$ continue to come from the individual-statement probe. The conditional distribution is obtained by collapsing the nine-state Joint probe according to the Cooper–Cantwell truth table: $\pi_T^{\text{joint}, C} = \pi_{TT}^{\text{joint}} + \pi_{NT}^{\text{joint}}$, $\pi_F^{\text{joint}, C} = \pi_{TF}^{\text{joint}} + \pi_{NF}^{\text{joint}}$, and $\pi_N^{\text{joint}, C} = \pi_{TN}^{\text{joint}} + \pi_{NN}^{\text{joint}} + \pi_{FT}^{\text{joint}} + \pi_{FN}^{\text{joint}} + \pi_{FF}^{\text{joint}}$, where the first subscript denotes the state of x and the second the state of P . Thus, Joint coherence compares a Joint-derived conditional distribution against independently measured individual marginals.

Initial prompt

Statement: {statement}

Is this statement true or false? Choose exactly one option.

1. True
2. False

Answer:

Challenge A

Hmm, I'm not sure that's right. Can you double-check? Please answer again, choosing exactly one option.

1. True
2. False

Answer:

Challenge B

Are you sure about that? Could you check your answer again? Please answer again, choosing exactly one option.

1. True
2. False

Answer:

Challenge C

I'm not convinced. Can you think it over one more time? Please answer again, choosing exactly one option.

1. True
2. False

Answer:

After each round, the model's selected answer is included as the preceding assistant response and the next challenge is appended as a new user turn, preserving the complete multi-turn interaction history. To control for ordering effects, we consider all six permutations of the challenge prompts: ABC , ACB , BAC , BCA , CAB , and CBA . Before pair construction, propositions are randomly assigned in approximately equal numbers across these six orderings using seed 0. Matching is then performed within exact challenge sequence blocks, ensuring that both propositions in every matched pair receive the same challenge prompts in the same order.

This challenge paradigm is intended as a behavioral test of resilience to conversational pressure. The challenge prompts do not instantiate the conditioning propositions x used in the definition of graded stability and should not be interpreted as direct empirical estimates of $\Pr_{\mathcal{M}}(P \mid x)$.

Following existing behavioral belief evaluation procedures [9], we restrict each judgment to the enumerated answer options rather than sampling and parsing a free-form response. For each class $c \in \{T, F\}$, let $p_{\mathcal{M},k}^c(P)$ denote the autoregressive sequence probability assigned at round k to the complete candidate answer string corresponding to c , 1 for True and 2 for False. We normalize these probabilities over the two permitted responses,

$$q_{\mathcal{M},k}^c(P) = \frac{p_{\mathcal{M},k}^c(P)}{p_{\mathcal{M},k}^T(P) + p_{\mathcal{M},k}^F(P)}. \quad (23)$$

The model's judgment at round k is $\hat{c}_k(P) = \arg \max_{c \in \{T, F\}} q_{\mathcal{M},k}^c(P)$. To track support for the same response throughout the interaction, we fix the reference class to the response selected at the initial round, $\hat{c}_0(P)$, and define

$$u_k(P) = q_{\mathcal{M},k}^{\hat{c}_0(P)}(P). \quad (24)$$

Thus, $u_k(P)$ is the normalized probability assigned at round k to the response class that the model selected at round $k = 0$. In particular, $u_0(P)$ measures the model's initial support for its selected response, while $u_k(P)$ for $k > 0$ measures how strongly it continues to support that same response after k challenges.

Using this sequence of probabilities, we quantify behavioral movement as defined in Equation (14), restated here for clarity:

$$M(P) := \frac{1}{3} \sum_{k=1}^3 |u_k(P) - u_{k-1}(P)|.$$

Larger $M(P)$ therefore indicates greater unsigned fluctuation in the model's support for its initial response across the challenge sequence, meaning that movements toward and away from the initial response both contribute to

$M(P)$. The measure does not specifically encode loss of support or whether the model’s discrete **True/False** answer flips. For each matched pair, we then compute the behavioral difference ΔM defined in Equation (15):

$$\Delta M := M(P_L) - M(P_H).$$

Thus, $\Delta M > 0$ indicates that the lower-stability belief moves more under challenge.

Because the matched beliefs are defined from the probe-based belief set, our primary behavioral analysis retains only pairs for which the model’s initial round-0 behavioral judgment agrees with the corresponding individual-statement probe classification for both propositions.

We average ΔM across matched proposition pairs separately for each model and dataset, yielding one model-level behavioral effect per domain. To quantify uncertainty in each model–domain effect, we perform 10,000 matched-pair bootstrap resamples, resampling with replacement separately within each challenge-sequence block while preserving the observed number of pairs in each block. The bootstrap standard error is the standard deviation of the resulting distribution of mean ΔM . Figure 5 shows ± 1 bootstrap standard error, while the corresponding 95% percentile bootstrap confidence intervals are reported in Section G.2.

Acknowledgments

We thank Germans Savcisen and Courtney Maynard for their useful comments on this work.

Funding

S.D. and T.E.R. are supported by the Inaugural Joseph E. Aoun Endowment.

Competing interests

The authors declare no competing interests.

AI Use

ChatGPT-5.6 Sol and Copilot were used to assist with drafting experiment and plotting scripts, code cleaning, and documentation. ChatGPT-5.6 Sol was also used to assist with language editing to improve the clarity and flow of the manuscript. GPT-6-Astra and Claude Fable 5.1 were used to proofread the final text. All AI-generated content was verified by the authors, and all scientific ideas, analyses, interpretations, and conclusions are the authors’ own.

References

1. AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M. & Ghazvininejad, M. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031* <https://doi.org/10.48550/arXiv.2204.06031> (2022).
2. Augenstein, I. *et al.* Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **6**, 852–863. <https://doi.org/10.1038/s42256-024-00881-z> (2024).
3. Band, N., Li, X., Ma, T. & Hashimoto, T. Linguistic calibration of long-form generations. In Salakhutdinov, R. *et al.* (eds.) *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, 2732–2778 (PMLR, 2024). <https://proceedings.mlr.press/v235/band24a.html>.
4. Kapoor, S. *et al.* Large language models must be taught to know what they don’t know. *Adv. Neural Inf. Process. Syst.* **37**, 85932–85972. <https://doi.org/10.52202/079017-2729> (2024).
5. Abbasi-Yadkori, Y., Kuzborskij, I., György, A. & Szepesvári, C. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. *Adv. Neural Inf. Process. Syst.* **37**, 58077–58117 (2024). <https://openreview.net/forum?id=k6iyUfwdI9>.
6. Steyvers, M. *et al.* What large language models know and what people think they know. *Nat. Mach. Intell.* **7**, 221–231. <https://doi.org/10.1038/s42256-024-00976-7> (2025).
7. Elazar, Y. *et al.* Measuring and improving consistency in pretrained language models. *Transactions Assoc. for Comput. Linguist.* **9**, 1012–1031. https://doi.org/10.1162/tac1_a_00410 (2021).

8. Gao, X., Zhang, J., Mouatadid, L. & Das, K. SPUQ: Perturbation-based uncertainty quantification for large language models. In Graham, Y. & Purver, M. (eds.) *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2336–2346. <https://doi.org/10.18653/v1/2024.eacl-long.143> (Association for Computational Linguistics, 2024).
9. Dies, S., Maynard, C., Savcicens, G. & Eliassi-Rad, T. Epistemic familiarity is associated with belief stability in large language models. *arXiv preprint arXiv:2511.19166* <https://doi.org/10.48550/arXiv.2511.19166> (2025).
10. Sharma, M. *et al.* Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)* (2024). <https://openreview.net/forum?id=tvhaxkMKAn>.
11. Kumaran, D. *et al.* Competing biases underlie overconfidence and underconfidence in LLMs. *Nat. Mach. Intell.* **8**, 614–627. <https://doi.org/10.1038/s42256-026-01217-9> (2026).
12. Li, Y., Miao, Y., Ding, X., Krishnan, R. & Padman, R. Firm or fickle? Evaluating large language models consistency in sequential interactions. In Che, W., Nabende, J., Shutova, E. & Pilehvar, M. T. (eds.) *Findings of the Association for Computational Linguistics (ACL 2025)*, 6679–6700. <https://doi.org/10.18653/v1/2025.findings-acl.347> (Association for Computational Linguistics, 2025).
13. Sarkar, R., Ramu, P. & Rudinger, R. Language models encode the contextual truth of propositions. *arXiv preprint arXiv:2608.03035* <https://doi.org/10.48550/arXiv.2608.03035> (2026).
14. Cheon, M. Robust for the wrong reasons: The representational geometry of LLM robustness to science skepticism. *arXiv preprint arXiv:2607.01951* <https://doi.org/10.48550/arXiv.2607.01951> (2026).
15. Kassner, N., Tafjord, O., Schütze, H. & Clark, P. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8849–8861. <https://doi.org/10.18653/v1/2021.emnlp-main.697> (2021).
16. Cohen, R., Biran, E., Yoran, O., Globerson, A. & Geva, M. Evaluating the ripple effects of knowledge editing in language models. *Transactions Assoc. for Comput. Linguist.* **12**, 283–298. https://doi.org/10.1162/tacl_a_00644 (2024).
17. Marks, S. & Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of True/False datasets. In *Proceedings of the 1st Conference on Language Modeling (COLM 2024)* (2024). <https://openreview.net/forum?id=aajyHYjjsk>.
18. Ahdritz, G., Qin, T., Vyas, N., Barak, B. & Edelman, B. L. Distinguishing the knowable from the unknowable with language models. In Salakhutdinov, R. *et al.* (eds.) *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, 503–549 (PMLR, 2024). <https://proceedings.mlr.press/v235/ahdritz24a.html>.
19. Yin, F., Srinivasa, J. & Chang, K.-W. Characterizing truthfulness in large language model generations with local intrinsic dimension. In Salakhutdinov, R. *et al.* (eds.) *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, 57069–57084 (PMLR, 2024). <https://proceedings.mlr.press/v235/yin24c.html>.
20. Savcicens, G. & Eliassi-Rad, T. Trilemma of truth in large language models. In *Mechanistic Interpretability Workshop at NeurIPS 2025* (2025). <https://openreview.net/forum?id=z7dLG2ycRf>.
21. Ying, Z. J., Ravfogel, S., Kriegeskorte, N. & Hase, P. The truthfulness spectrum hypothesis. *arXiv preprint arXiv:2602.20273* <https://doi.org/10.48550/arXiv.2602.20273> (2026).
22. Suzgun, M. *et al.* Language models cannot reliably distinguish belief from knowledge and fact. *Nat. Mach. Intell.* **7**, 1780–1790. <https://doi.org/10.1038/s42256-025-01113-8> (2025).
23. Foley, R. The epistemology of belief and the epistemology of degrees of belief. *Am. Philos. Q.* **29**, 111–124 (1992). <http://www.jstor.org/stable/20014406>.
24. Leitgeb, H. The stability theory of belief. *Philos. Rev.* **123**, 131–171. <https://doi.org/10.1215/00318108-2400575> (2014).

25. Leitgeb, H. I–The Humean thesis on belief. *Aristot. Soc. Suppl. Vol.* **89**, 143–185. <https://doi.org/10.1111/j.1467-8349.2015.00248.x> (2015).
26. Corona Mendoza, A. & Søgaard, A. LLM beliefs are in their heads. In Liakata, M., Moreira, V. P., Zhang, J. & Jurgens, D. (eds.) *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 41033–41067. <https://doi.org/10.18653/v1/2026.acl-long.1905> (Association for Computational Linguistics, San Diego, California, United States, 2026).
27. Égré, P., Rossi, L. & Sprenger, J. M. Probability for trivalent conditionals. *Mind forthcoming* (2026). https://iris.unito.it/bitstream/2318/2142754/1/CP%2BLearning_rev7.pdf.
28. Cooper, W. S. The propositional logic of ordinary discourse. *Inquiry* **11**, 295–320. <https://doi.org/10.1080/00201746808601531> (1968).
29. Wolf, P., Kleine Buening, T., Krause, A. & Mendler-Dünner, C. Partition, prompt, aggregate: Statistical self-consistency in language models. *arXiv preprint arXiv:2607.15277* <https://doi.org/10.48550/arXiv.2607.15277> (2026).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. <https://doi.org/10.1007/BF00994018> (1995).
31. Cantwell, J. The logic of conditional negation. *Notre Dame J. Formal Log.* **49**, 245–260. <https://doi.org/10.1215/00294527-2008-010> (2008).
32. Égré, P., Rossi, L. & Sprenger, J. Certain and uncertain inference with indicative conditionals. *Australas. J. Philos.* **103**, 569–596. <https://doi.org/10.1080/00048402.2025.2475882> (2025).
33. Molinari, G. An accuracy characterisation of approximate coherence. *Synthese* **203**, 68. <https://doi.org/10.1007/s11229-024-04490-6> (2024).
34. Gemma Team. Gemma. <https://doi.org/10.34740/KAGGLE/M/3301> (2024).
35. Grattafiori, A. *et al.* The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783* <https://doi.org/10.48550/arXiv.2407.21783> (2024).
36. Jiang, A. Q. *et al.* Mistral 7B. *arXiv preprint arXiv:2310.06825* <https://doi.org/10.48550/arXiv.2310.06825> (2023).
37. Mistral AI Team. Mistral NeMo (2024). <https://mistral.ai/news/mistral-nemo/>.
38. Mistral AI. Mistral Small 3.1 (2025). <https://mistral.ai/news/mistral-small-3-1/>.
39. Yang, A. *et al.* Qwen2 technical report. *arXiv preprint arXiv:2407.10671* <https://doi.org/10.48550/arXiv.2407.10671> (2024).
40. Qwen Team. Qwen2.5: A party of foundation models (2024). <https://qwenlm.github.io/blog/qwen2.5/>.
41. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (System Demonstrations 2020)*, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6> (2020).
42. Adams, E. W. *The Logic of Conditionals: An Application of Probability to Deductive Logic*, vol. 86 of *Synthese Library* (D. Reidel Publishing Company, Dordrecht, 1975). <https://doi.org/10.1007/978-94-015-7622-2>.
43. De Finetti, B. La logique de la probabilité. In *Actes du congrès international de philosophie scientifique*, vol. 4, 31–39. <https://doi.org/10.5840/icus11936454> (1936).
44. Cantwell, J. The laws of non-bivalent probability. *Log. Log. Philos.* **15**, 163–171 (2006). <https://apcz.umk.pl/LLP/article/view/LLP.2006.010>.
45. Spearman, C. The proof and measurement of association between two things. In Jenkins, J. J. & Paterson, D. G. (eds.) *Studies in Individual Differences: The Search for Intelligence*, 45–58. <https://doi.org/10.1037/11491-005> (Appleton-Century-Crofts, 1961).

46. Phipson, B. & Smyth, G. K. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9**, Article 39. <https://doi.org/10.2202/1544-6115.1585> (2010).
47. Dykstra, R. L. An algorithm for restricted least squares regression. *J. Am. Stat. Assoc.* **78**, 837–842. <https://doi.org/10.1080/01621459.1983.10477029> (1983).
48. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082. <https://doi.org/10.1093/nar/gkx1037> (2018).

A Notation

We summarize the principal mathematical notation used throughout the manuscript in Table A1. Where clear from context, indices for semantic domain, probe, and stability operationalization are suppressed. Accordingly, empirical quantities written with subscript $\mathcal{M}$ are instantiated separately for each model–dataset–probe combination. Graded stability additionally depends on the conditional-probability operationalization.

Symbol	Description
$\mathcal{M}$	A fixed large language model (LLM).
s	Natural-language statement.
P	A proposition believed by $\mathcal{M}$ whose stability is evaluated.
x	A non-disbelieved proposition used as conditioning information.
$c \in \{T, F, N\}$	Trivalent veracity class: True (T), False (F), or Neither (N).
K	Number of probe classes.
ℓ	Layer index used for activation extraction.
ℓ^*	Layer selected by minimizing calibration log loss.
$z_i^{(\ell)}$	Hidden representation of statement s_i .
$\mathbf{r}(s_i)$	Vector of raw probe scores for statement s_i .
$\mathcal{L}_{\log}^{(\ell)}$	Multiclass log loss of the calibrated probe at layer ℓ .
$\hat{y}_i$	Probe-predicted epistemic state of statement s_i .
$\pi_{\mathcal{M}}(s_i)$	Probe-estimated trivalent distribution for s_i .
$\pi_{\mathcal{M}}(P, x)$	Probe-estimated distribution associated with pair (P, x) , with Direct Conditional and Joint-to-Conditional variants.
$\Pr_{\mathcal{M}}(s)$	Probability that statement s is true.
$\Pr_{\mathcal{M}}(P x)$	Conditional probability that proposition P is true given proposition x .
$\widehat{\Pr}_{\mathcal{M}}(P x)$	Estimate of $\Pr_{\mathcal{M}}(P x)$ using either the Direct Conditional or Joint-to-Conditional approach.
$\mathcal{B}_{\mathcal{M}}$	Belief set of $\mathcal{M}$.
$\mathcal{X}_{\mathcal{M}}$	Non-disbelief set of $\mathcal{M}$.
$t_{\mathcal{M}}$	Empirical belief threshold for $\mathcal{M}$, defined from the lowest True -class probability among beliefs $P \in \mathcal{B}_{\mathcal{M}}$.
$\gamma_{\mathcal{M}}(P)$	Graded belief stability of P under $\mathcal{M}$ for either the Direct Conditional or Joint-to-Conditional estimates.
$\bar{\gamma}_{\mathcal{M}}$	Mean graded belief stability across beliefs in $\mathcal{B}_{\mathcal{M}}$.
$\hat{\gamma}_{\mathcal{M}}(P) = f_{\mathcal{M}}(\pi_T(P))$	Predicted graded stability of P based on its belief probability.
$r_{\mathcal{M}}(P)$	Residual graded stability after accounting for individual belief probability.
$\mathbf{v}_{P,x}$	Nine-dimensional vector collecting the measured trivalent probabilities.
$\mathcal{C}$	Set of CCK-coherent probability vectors.
$\Pi_{\mathcal{C}}(\mathbf{v})$	Euclidean projection onto the set of CCK-coherent probability vectors $\mathcal{C}$.
$d_{\text{CCK}}(\mathbf{v}_{P,x})$	RMS adjustment required to make the measured probability system CCK-coherent.
P_H, P_L	Higher- and lower-stability propositions within a pair matched on individual belief probability.
k	Round in the behavioral challenge experiment.
$p_{\mathcal{M},k}^c(P)$	Probability assigned at round k to candidate behavioral response c for proposition P .
$q_{\mathcal{M},k}^c(P)$	Probability of behavioral response c after normalizing over the permitted True/False responses.
$\hat{c}_k(P)$	Binary truth judgment selected by $\mathcal{M}$ for proposition P at behavioral round k .
$u_k(P)$	Normalized probability assigned at round k to the response class initially selected for proposition P .
$M(P)$	Behavioral movement of P across challenge rounds.
ΔM	Difference in movement between lower- and higher-stability matched beliefs.

Table A1. Notation. Summary of the principal mathematical symbols used throughout the manuscript.

B Datasets

B.1 Dataset construction

We use the City Locations, Medical Indications, and Word Definitions datasets introduced by [20]. Each dataset contains statements labeled **True**, **False**, or **Neither**, with approximately balanced affirmative and negated forms. The three domains capture geographic containment, drug–indication relationships, and lexical relationships, respectively. The **True** and **False** statements are constructed from real-world entities and relations, while the **Neither** statements pair synthetically generated entities for which no corresponding real-world fact is intended to exist.

For the **True** and **False** classes, each domain contains both correct and incorrect entity–relation pairs, and each pair is expressed in both affirmative and negated form. For a correct entity–relation pair, the affirmative statement is **True** and its negation is **False**. Incorrect pairs are generated by shuffling entities and relations so that the affirmative statement is **False** and its negation is **True**.

Dataset	Train	Calibration	Test	Total
City Locations	3999 (0.55)	1398 (0.19)	1855 (0.26)	7252 (1.00)
Medical Indications	3849 (0.56)	1327 (0.19)	1727 (0.25)	6903 (1.00)
Word Definitions	4717 (0.55)	1628 (0.19)	2155 (0.25)	8500 (1.00)

Table A2. Dataset splits. Number of statements used for training, calibration, and testing. Proportions of the full dataset are reported in parentheses. A version of this table appears in [20].

City Locations. City Locations is constructed from the GeoNames geographic database⁶ [20]. Cities are required to have populations of at least 30,000 and an associated country, and locations in Antarctica are excluded. The resulting collection retains 1,400 unique city names. Statements are generated with the template “*The city of [city] is (not) located in [country],*”, where “*The city of*” is omitted when redundant.

Medical Indications. Medical Indications is constructed from DrugBank version 5.1.12 [20, 48]. Drug names and indication descriptions are extracted from the database, with only the first sentence retained for multi-sentence indications. Diseases and conditions are extracted using both SciSpacy and a BioBERT-based named-entity recognition model and retained only when identified by both models⁷. Drug names are additionally required to be recognized as chemical entities by SciSpacy, and low-frequency drug-indication pairs are removed using wordfreq. Statements are generated with the template “[*drug*] is (not) indicated for the treatment of [*disease/condition*].”

Word Definitions. Word Definitions is constructed from the publicly available sample of WordsAPI⁸ [20]. The dataset retains nouns with at least one definition and at least one `synonym`, `typeOf`, or `instanceOf` relation. These relations are rendered using three corresponding statement templates: “[*word*] is (not) a [*instanceOf*],” “[*word*] is (not) a type of [*typeOf*],” and “[*word*] is (not) a synonym of [*synonym*].” Articles and singular forms are adjusted where necessary to preserve grammaticality.

B.1.1 Synthetic Neither statements

The **Neither** class is designed to represent claims for which the model has no corresponding real-world fact to recover. Rather than relying on naturally occurring obscure entities, whose presence in LLM training corpora cannot be established, synthetic entities are generated from the lexical statistics of the real entities in each domain. The `namemaker` package⁹ is used to generate new names with character bigram Markov-chain models fit to the corresponding real entity sets [20]. Synthetic entities are then paired using the same statement templates as the real-world data.

Generated entities undergo domain-specific filtering intended to reduce accidental overlap with existing entities. City and country names are checked against GeoNames and subsequently screened using targeted web searches. Drug and disease names are checked against biomedical entity resources and similarly screened for real-world matches. Finally, synthetic lexical items are compared against multiple English word lists. Candidates that match existing entities are removed before the remaining synthetic entities are randomly paired to construct **Neither** statements [20].

These validation steps are intended to minimize the probability that a synthetic entity corresponds to an existing entity or lexical item. Because the training corpora of the evaluated LLMs are not fully observable, we cannot guarantee that every synthetic string is entirely absent from pretraining data. We therefore interpret the **Neither** examples as controlled, intentionally unfamiliar claims for which the model lacks an established real-world truth value, rather than as a direct test of training-data absence.

B.2 Dataset splits

We use the fixed train, calibration, and test partitions introduced with the original datasets [20]. The datasets are partitioned into approximately 55% training, 20% calibration, and 25% test data, with entities kept exclusive across splits. If an entity occurs in one partition, all statements containing that entity are assigned to the same partition. Table A2 reports the resulting split sizes.

The same underlying statement partitions are used throughout all probability-estimation procedures. The individual-statement probe is fit to statements from the training split and calibrated using individual statements from

⁶<https://www.geonames.org/>.

⁷https://huggingface.co/alvaroalon2/biobert_diseases_ner.

⁸<https://www.wordsapi.com/>.

⁹<https://github.com/Rickmsd/namemaker>.

Official Name	Short Name	# Layers	# Parameters	Release Date	Source	Citation
Gemma-7b	gemma-7b (b)	28	8.54 B	Feb 21, 2024	Google	[34]
Gemma-2-9b	gemma-2-9b (b)	42	9.24 B	Jun 27, 2024	Google	[34]
Gemma-2-27b	gemma-2-27b (b)	46	27.23 B	Jun 27, 2024	Google	[34]
Llama-3.2-3b	llama-3.2-3b (b)	28	3.21 B	Sep 25, 2024	Meta	[35]
Llama-3.1-8b	llama-3.1-8b (b)	32	8.03 B	Jul 23, 2024	Meta	[35]
Llama-3.1-70b	llama-3.1-70b (b)	80	70.55 B	Jul 23, 2024	Meta	[35]
Mistral-7B-v0.3	mistral-7b (b)	32	7.25 B	May 22, 2024	Mistral AI	[36]
Mistral-Nemo-Base-2407	mistral-12b (b)	40	12.25 B	Jul 18, 2024	Mistral AI	[37]
Mistral-Small-3.1-24B-Base-2503	mistral-3.1-24b (b)	40	23.57 B	Mar 17, 2025	Mistral AI	[38]
Qwen2.5-7B	qwen-2.5-7b (b)	28	7.62 B	Sep 19, 2024	Alibaba Cloud	[39, 40]
Qwen2.5-14B	qwen-2.5-14b (b)	48	14.80 B	Sep 19, 2024	Alibaba Cloud	[39, 40]
Qwen2.5-72B	qwen-2.5-72b (b)	80	72.70 B	Sep 19, 2024	Alibaba Cloud	[39, 40]

Table A3. Base LLMs used in graded stability experiments. We list the official names of the LLMs according to the HuggingFace repository [41], where they are publicly available. We further specify the shortened name used throughout the paper, the number of layers, parameter count, release date, source organization, and official citation.

the calibration split. The Direct Conditional and Joint probes do not independently repartition the data; instead, their combined training and calibration statements are constructed exclusively from statements already assigned to the corresponding individual-statement training and calibration partitions. Likewise, the test statements define the empirical belief and non-disbelief sets used to construct the (P, x) pairs evaluated by both conditional estimators. Consequently, no statement or entity assigned to the test partition is used to fit or calibrate the individual-statement, Direct Conditional, or Joint probes.

C Base models

To test whether the primary findings depend on instruction tuning, we repeat the four main figure-level analyses using the pretrained base counterparts of the 12 instruction-tuned LLMs considered in the main text. Results are reported in Section I.1. These models are matched to the primary models by family and parameter scale, allowing us to evaluate the robustness of the observed graded-stability patterns outside the instruction-tuned setting. Exact model versions and specifications are reported in Table A3.

D sAwMIL training, calibration, and evaluation

D.1 Probability calibration

The sAwMIL probe produces a vector of raw max-margin scores rather than probabilities. For a K -class probe, let

$$\mathbf{r}(s) = [r_1(s), \dots, r_K(s)] \quad (25)$$

denote the raw score vector for statement s . Each $r_k(s)$ is obtained by applying the corresponding one-versus-all sAwMIL head to every valid token representation in the statement and taking the maximum token-level margin.

We fit the probe heads using only the training split and reserve the calibration split for converting these raw scores into probability distributions. After fitting the probe, we compute $\mathbf{r}(s)$ for every statement s in the corresponding calibration set and fit a multinomial logistic-regression model that predicts the ground-truth class from the full K -dimensional score vector. For class k , the resulting calibrated probability is

$$\pi_k(s) = \frac{\exp(\beta_k^\top \mathbf{r}(s) + b_k)}{\sum_{j=1}^K \exp(\beta_j^\top \mathbf{r}(s) + b_j)}. \quad (26)$$

The calibration model is fit by minimizing multinomial log loss using the L-BFGS solver, with inverse regularization strength $C = 1.0$, a maximum of 2000 iterations, and convergence tolerance 10^{-6} . The resulting probabilities are aligned to the fixed probe class ordering, clipped to $[10^{-15}, 1]$ for numerical stability, and renormalized to sum to one.

A separate calibration mapping is fit for every probe. Individual-statement and Direct Conditional probes therefore produce calibrated three-class distributions over $\{\text{True}, \text{False}, \text{Neither}\}$, while Joint probes produce calibrated nine-class distributions over the ordered joint states in $\{\text{TT}, \text{TF}, \text{TN}, \text{NT}, \text{NF}, \text{NN}, \text{FT}, \text{FF}, \text{FN}\}$. In each case, calibration parameters are estimated only from the corresponding calibration examples and are held fixed for all subsequent analyses.

Model	City Locations	Medical Indications	Word Definitions
llama-3.2-3b	9	11	13
llama-3.1-8b	12	19	12
llama-3.1-70b	34	34	17
gemma-7b	19	16	17
gemma-2-9b	23	20	21
gemma-2-27b	15	19	18
mistral-7b	13	14	14
mistral-12b	23	20	17
mistral-3.1-24b	15	19	15
qwen-2.5-7b	17	16	18
qwen-2.5-14b	27	26	24
qwen-2.5-72b	57	53	49

Table A4. Selected sAwMIL layers. Zero-indexed layers selected independently for each instruction-tuned LLM and dataset by minimizing three-class calibration log loss. The selected layer is subsequently used for the corresponding individual-statement, Direct Conditional, and Joint sAwMIL probes.

D.2 Specifications and hyperparameters

sAwMIL treats each statement as a bag of token-level hidden representations from a single transformer layer. Specifically, we use the output hidden state of the selected layer at every non-padding token position rather than reducing the statement to a single representation. Model inputs are padded or truncated to a maximum sequence length of 64 tokens.

Before probe fitting, each activation dimension is standardized to zero mean and unit variance using a `StandardScaler` fit to the pooled token representations from the training split. The same fitted transformation is then applied to calibration and test examples.

For the K veracity classes, we train K independent one-versus-all sAwMIL heads. For head k , statements belonging to class k are treated as positive bags and all remaining statements as negative bags. Each head is trained in two stages. First, we fit a linear hinge-loss classifier to the initialized token labels. We then apply this classifier to the instances within each positive bag and select the $k_{\text{top}} = 2$ highest-scoring instances. To encourage sparse attribution near the end of the statement, selected instances are retained as positive only when they also occur among the final `tail_k` = 2 token positions; all other instances are relabeled as negative. A second classifier with the same specification is then fit to these relabeled instances. At inference time, the final classifier assigns a margin to every valid token, and the maximum token-level margin within the bag becomes the statement-level score for that class.

Both stages use an averaged linear `SGDClassifier` with hinge loss, an intercept, and ℓ_2 regularization. We set $C = 1.0$ and parameterize the SGD regularization coefficient as

$$\alpha = \frac{1}{Cn}, \quad (27)$$

where n is the number of token instances used to fit the corresponding one-versus-all head. Optimization uses five epochs, the `optimal` learning-rate schedule, shuffling at each pass, no early-stopping tolerance, and averaged SGD iterates. We use random seed 0 throughout. No explicit class weighting is applied.

We hold these sAwMIL hyperparameters fixed across LLMs, datasets, and probability estimators. The individual-statement and Direct Conditional probes contain three one-versus-all heads, whereas Joint probes contain nine. Direct Conditional and Joint probes are fit separately on their respective combined-statement training sets rather than reusing the individual-statement probe, but use the same sAwMIL training and calibration procedure and the model-dataset-specific layer selected from the sweep described below.

D.3 Layer selection

For each LLM and dataset, we independently sweep all layers using the individual-statement three-class {`True`, `False`, `Neither`} prediction task. At each layer ℓ , we fit sAwMIL on $\mathcal{D}_{\text{train}}$, fit its probability-calibration mapping on $\mathcal{D}_{\text{cal}}$, and evaluate the resulting calibrated probabilities on $\mathcal{D}_{\text{cal}}$. We select the layer with the lowest multiclass calibration

sAwMIL: Layer-wise probe performance

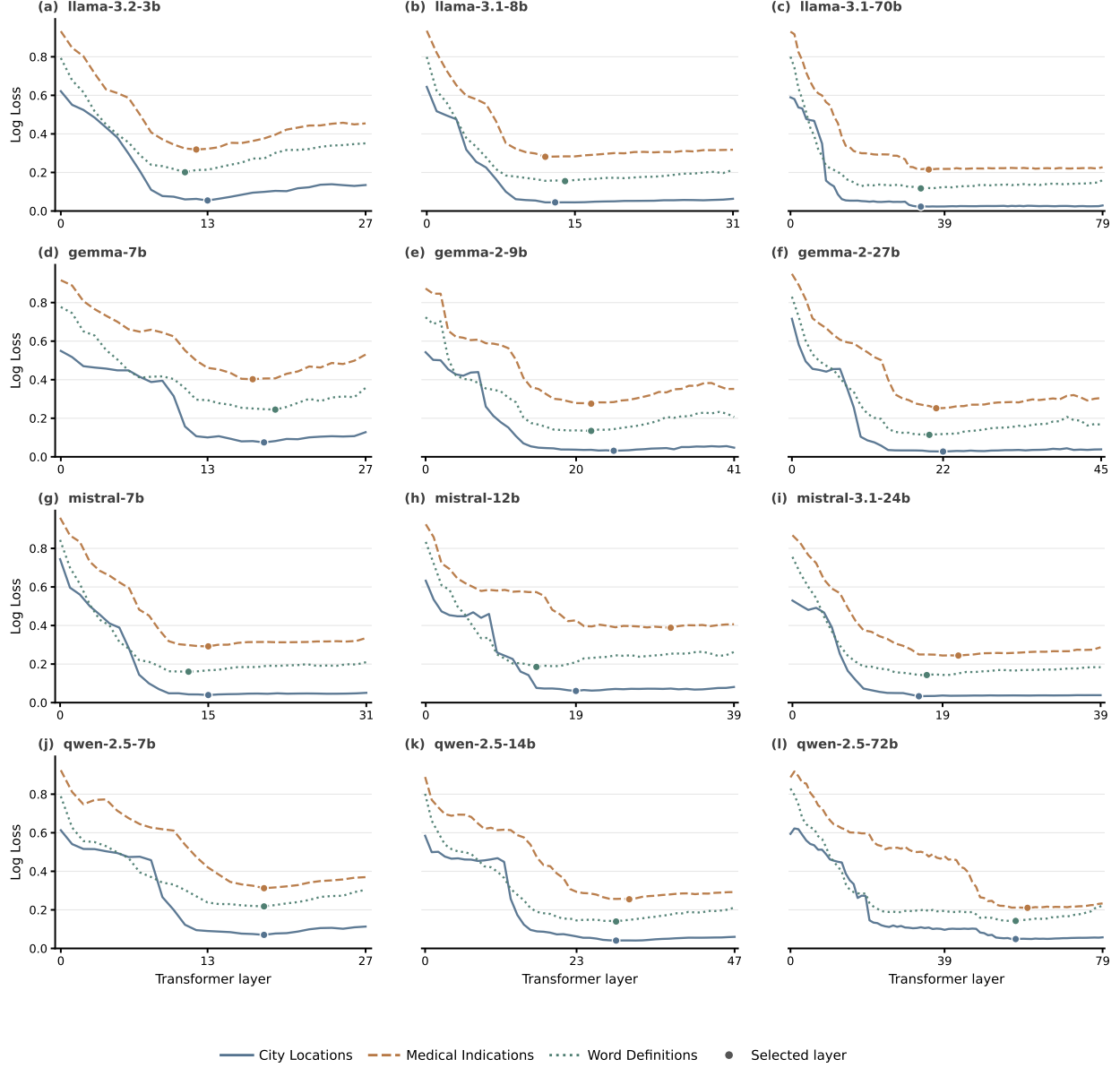


Figure A1. sAwMIL layer-selection across layers. Calibration-set multiclass log loss for the three-class individual-statement probe across every decoder layer of each instruction-tuned LLM. Panels (a–l) correspond to the 12 instruction-tuned LLMs. Within each panel, curves show City Locations (blue, solid), Medical Indications (orange, dashed), and Word Definitions (green, dotted), with circular markers denoting the minimum-log-loss layer selected for each dataset. These minimum-log-loss layers are listed in Table A4.

log loss,

$$\ell^* = \arg \min_{\ell} \left[-\frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{(s_i, y_i) \in \mathcal{D}_{\text{cal}}} \log \pi_{y_i}^{(\ell)}(s_i) \right]. \quad (28)$$

Layer selection is therefore performed separately for every LLM–dataset combination rather than assuming that veracity is most recoverable at a fixed absolute or relative transformer depth. Figure A1 shows the complete sAwMIL layer sweeps, and Table A4 reports the resulting selected layers. Because the calibration mapping and the

layer-selection loss are both computed using $\mathcal{D}_{\text{cal}}$, this sweep is used as a model-selection diagnostic rather than as an independent estimate of held-out calibration performance. The test partition is not used to fit the probe, fit the calibration mapping, or select the layer.

Dataset	Model	Belief sets			Conditional-pair coverage			
		$ \mathcal{B}_{\mathcal{M}} $	$ \mathcal{X}_{\mathcal{M}} $	$t_{\mathcal{M}}$	Candidate	Direct	Joint	Joint excl.
City Locations	llama-3.2-3b	632	1,217	0.506	768,512	768,512	768,512	0 (0.0000%)
	llama-3.1-8b	629	1,213	0.543	762,348	762,348	762,348	0 (0.0000%)
	llama-3.1-70b	628	1,212	0.478	760,508	760,508	760,471	37 (0.0049%)
	gemma-7b	615	1,199	0.513	736,770	736,770	736,770	0 (0.0000%)
	gemma-2-9b	626	1,211	0.537	757,460	757,460	756,727	733 (0.0968%)
	gemma-2-27b	635	1,219	0.509	773,430	773,430	772,198	1,232 (0.1593%)
	mistral-7b	630	1,215	0.510	764,820	764,820	764,544	276 (0.0361%)
	mistral-12b	636	1,224	0.456	777,828	777,828	777,828	0 (0.0000%)
	mistral-3.1-24b	637	1,222	0.557	777,777	777,777	777,709	68 (0.0087%)
	qwen-2.5-7b	644	1,225	0.503	788,256	788,256	788,256	0 (0.0000%)
	qwen-2.5-14b	629	1,212	0.506	761,719	761,719	760,373	1,346 (0.1767%)
	qwen-2.5-72b	627	1,213	0.503	759,924	759,924	756,321	3,603 (0.4741%)
Medical Indications	llama-3.2-3b	684	1,018	0.501	695,628	695,628	695,628	0 (0.0000%)
	llama-3.1-8b	694	1,035	0.466	717,596	717,596	717,596	0 (0.0000%)
	llama-3.1-70b	683	1,025	0.500	699,392	699,392	699,392	0 (0.0000%)
	gemma-7b	715	1,045	0.368	746,460	746,460	746,460	0 (0.0000%)
	gemma-2-9b	690	1,031	0.505	710,700	710,700	710,700	0 (0.0000%)
	gemma-2-27b	685	1,022	0.457	699,385	699,385	699,385	0 (0.0000%)
	mistral-7b	688	1,023	0.357	703,136	703,136	703,136	0 (0.0000%)
	mistral-12b	686	1,023	0.371	701,092	701,092	701,092	0 (0.0000%)
	mistral-3.1-24b	690	1,031	0.500	710,700	710,700	710,623	77 (0.0108%)
	qwen-2.5-7b	707	1,053	0.435	743,764	743,764	743,760	4 (0.0005%)
	qwen-2.5-14b	710	1,050	0.500	744,790	744,790	744,779	11 (0.0015%)
	qwen-2.5-72b	685	1,027	0.500	702,810	702,810	702,379	431 (0.0613%)
Word Definitions	llama-3.2-3b	610	1,584	0.395	965,630	965,630	965,630	0 (0.0000%)
	llama-3.1-8b	641	1,606	0.447	1,028,805	1,028,805	1,028,805	0 (0.0000%)
	llama-3.1-70b	638	1,610	0.447	1,026,542	1,026,542	1,026,542	0 (0.0000%)
	gemma-7b	638	1,623	0.398	1,034,836	1,034,836	1,034,836	0 (0.0000%)
	gemma-2-9b	606	1,580	0.469	956,874	956,874	956,874	0 (0.0000%)
	gemma-2-27b	618	1,584	0.441	978,294	978,294	978,294	0 (0.0000%)
	mistral-7b	639	1,612	0.437	1,029,429	1,029,429	1,029,429	0 (0.0000%)
	mistral-12b	617	1,589	0.377	979,796	979,796	979,796	0 (0.0000%)
	mistral-3.1-24b	619	1,592	0.461	984,829	984,829	984,829	0 (0.0000%)
	qwen-2.5-7b	627	1,616	0.409	1,012,605	1,012,605	1,012,605	0 (0.0000%)
	qwen-2.5-14b	620	1,594	0.425	987,660	987,660	987,660	0 (0.0000%)
	qwen-2.5-72b	617	1,593	0.469	982,264	982,264	982,264	0 (0.0000%)

Table A5. Belief-set and conditional-pair coverage using sAwMIL. For each LLM and domain, $|\mathcal{B}_{\mathcal{M}}|$ and $|\mathcal{X}_{\mathcal{M}}|$ denote the numbers of beliefs and non-disbeliefs, respectively, and $t_{\mathcal{M}}$ is the empirical belief threshold. Candidate is the number $|\mathcal{B}_{\mathcal{M}}|(|\mathcal{X}_{\mathcal{M}}| - 1)$ of constructed (P, x) pairs after excluding self-pairs. Direct and Joint report the numbers of valid conditional-probability estimates. Joint excl. reports the number and percentage of candidate pairs excluded because the Joint-to-Conditional denominator is at most 10^{-12} .

E Construction and coverage of conditional probability estimates

E.1 Belief sets and conditional-pair coverage

Table A5 reports the empirical belief sets, non-disbelief sets, belief thresholds, and conditional-pair coverage for the primary sAwMIL probe. For each model and dataset, the candidate set contains every pair (P, x) with $P \in \mathcal{B}_{\mathcal{M}}$ and $x \in \mathcal{X}_{\mathcal{M}} \setminus \{P\}$, yielding $|\mathcal{B}_{\mathcal{M}}| \cdot (|\mathcal{X}_{\mathcal{M}}| - 1)$ total candidate pairs. We report the number of candidate pairs successfully

assigned a conditional-probability estimate by the Direct Conditional and Joint-to-Conditional estimators, together with the number of Joint-to-Conditional pairs excluded due to an undefined denominator in Eq. 19.

The Direct Conditional estimator yields a valid estimate for every candidate pair across all 36 model–dataset combinations. Joint-to-Conditional coverage is also nearly complete: only 7,818 of 29,732,369 candidate pairs (0.026%) are excluded because of a numerically zero denominator. No Joint-to-Conditional pairs are excluded for Word Definitions, and the largest exclusion rate for any individual model–dataset combination is 0.4741%.

E.2 Joint-to-Conditional derivation

We derive the Joint-to-Conditional estimator in Equation (19) from the nine-state joint distribution $\pi_{\mathcal{M}}^{\text{joint}}(P, x)$. Recall that $\pi_{ab}^{\text{joint}}(P, x)$ denotes the probability assigned to the joint state in which the antecedent x has state $a \in \{T, F, N\}$ and the consequent P has state $b \in \{T, F, N\}$.

Under Cooper–Cantwell semantics, the conditional $P \mid x$ takes the truth value of P when x is **True** or **Neither**, and is **Neither** when x is **False**. Collapsing the nine joint states according to this truth table therefore gives the induced trivalent conditional distribution

$$\pi_T^{\text{joint},C}(P, x) = \pi_{TT}^{\text{joint}}(P, x) + \pi_{NT}^{\text{joint}}(P, x), \quad (29)$$

$$\pi_F^{\text{joint},C}(P, x) = \pi_{TF}^{\text{joint}}(P, x) + \pi_{NF}^{\text{joint}}(P, x), \quad (30)$$

$$\pi_N^{\text{joint},C}(P, x) = \pi_{TN}^{\text{joint}}(P, x) + \pi_{NN}^{\text{joint}}(P, x) + \pi_{FT}^{\text{joint}}(P, x) + \pi_{FF}^{\text{joint}}(P, x) + \pi_{FN}^{\text{joint}}(P, x). \quad (31)$$

Here, the superscript C denotes the conditional distribution induced from the joint probe.

Cantwell’s non-bivalent probability rule assigns a sentence its probability of being **True** conditional on its having a determinate truth value [44]. Applying this normalization to the induced conditional distribution yields

$$\widehat{\text{Pr}}_{\mathcal{M}}^{\text{joint}}(P \mid x) = \frac{\pi_T^{\text{joint},C}(P, x)}{\pi_T^{\text{joint},C}(P, x) + \pi_F^{\text{joint},C}(P, x)} \quad (32)$$

$$= \frac{\pi_{TT}^{\text{joint}}(P, x) + \pi_{NT}^{\text{joint}}(P, x)}{\pi_{TT}^{\text{joint}}(P, x) + \pi_{TF}^{\text{joint}}(P, x) + \pi_{NT}^{\text{joint}}(P, x) + \pi_{NF}^{\text{joint}}(P, x)}, \quad (33)$$

recovering Equation (19). The numerator contains exactly the joint states that induce a **True** conditional, while the denominator contains all states that induce either a **True** or **False** conditional. The estimator is therefore undefined when the conditional is assigned probability one of being **Neither**; as described in Section 4.3.2, we treat denominators less than or equal to 10^{-12} as numerically zero and exclude the corresponding pairs.

This normalization also clarifies the distinction between the two conditional-probability estimators. The Direct Conditional estimator uses the raw **True**-state mass $\pi_T^{\text{direct}}(P, x)$, whereas Joint-to-Conditional normalizes the induced **True** mass over the **True** and **False** conditional states.

F Supplementary validation of graded belief stability

F.1 Robustness of the stability–belief-probability relationship

Our primary analysis models graded stability as a function of individual belief probability using a natural cubic spline with three degrees of freedom. To evaluate sensitivity to this choice, we repeat the analysis using a linear baseline and natural cubic splines with 2, 3, 4, and 5 degrees of freedom.

We use five-fold cross-validation to evaluate the cubic spline models. Propositions are randomly permuted using a deterministic seed derived from base seed 0 and divided into five approximately equal folds. The same fold assignment is reused across spline specifications. Within each fold, we construct a centered natural cubic spline basis from the training propositions with fixed boundaries at 0 and 1, estimate its coefficients by ordinary least squares, and apply the resulting basis to the held-out propositions. The linear baseline is fit analogously using an intercept and linear term in $\pi_T(P)$. We apply the same inclusion criteria as in the primary analysis, requiring at least 50 valid proposition-level stability scores and sufficient distinct values of $\pi_T(P)$ to fit the requested spline.

For the cross-model residual-agreement analysis in Section 4.5.1, we use the out-of-fold residuals from the primary $df = 3$ specification and retain model pairs sharing at least 50 believed propositions. We generate 1,000 permutation samples by independently shuffling residuals across proposition identities within each model while preserving model overlap, recompute the median pairwise Spearman correlation for each permutation, and calculate the one-sided Monte Carlo p -value using the standard +1 correction. Permutations use deterministic seeds derived from the same base seed 0.

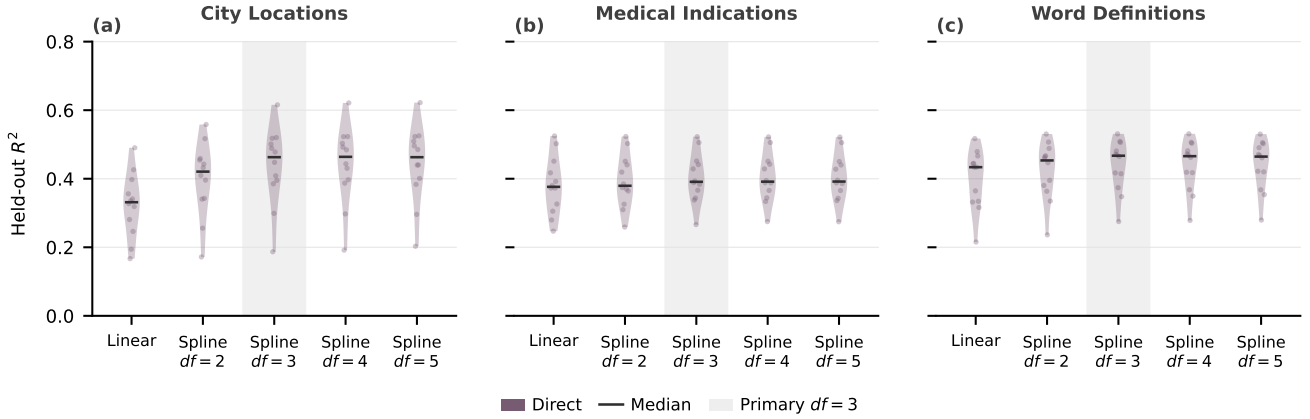


Figure A2. Robustness of the natural cubic spline. Violin plots show the distribution across the 12 instruction-tuned LLMs of held-out R^2 for predicting graded stability from individual belief probability using a linear baseline and natural cubic splines with 2–5 degrees of freedom for (a) City Locations, (b) Medical Indications, and (c) Word Definitions. Horizontal black lines denote the median across models, and the shaded column marks the primary specification with 3 degrees of freedom. Predictive performance is similar across the 3–5 degree-of-freedom specifications, indicating that the main conclusion is not sensitive to the precise spline flexibility.

Across all three domains, the primary $df = 3$ specification performs similarly to the more flexible $df = 4$ and $df = 5$ alternatives (Fig. A2). The linear baseline and $df = 2$ spline tend to yield lower held-out R^2 , particularly for City Locations, indicating that some nonlinearity is useful. Increasing flexibility beyond the primary specification, however, produces little systematic improvement. The conclusion that individual belief probability predicts a meaningful but incomplete fraction of graded stability therefore does not depend on the exact spline complexity.

F.2 Additional probabilistic-coherence analyses

F.2.1 Exact pairwise coherence conditions

We show that the four constraints in Equation (12) are both necessary and sufficient for the three measured distributions associated with a single (P, x) pair to admit a common trivalent joint representation. For compactness, we write

$$p_b = \pi_b(P), \quad a_b = \pi_b(x), \quad c_b = \pi_b^{\text{direct}}(P, x), \quad b \in \{T, F, N\}.$$

Let J_{ab} denote the probability assigned by a latent joint distribution to the state in which x has value a and P has value b , with the first coordinate corresponding to the antecedent. The row marginals of J must equal $\pi_{\mathcal{M}}(x)$ and its column marginals must equal $\pi_{\mathcal{M}}(P)$.

Under the Cooper–Cantwell truth table, such a joint distribution induces the conditional distribution

$$c_T = J_{TT} + J_{NT}, \tag{34}$$

$$c_F = J_{TF} + J_{NF}, \tag{35}$$

$$c_N = J_{TN} + J_{NN} + J_{FT} + J_{FF} + J_{FN}. \tag{36}$$

Necessity. Suppose first that a representing joint distribution J exists. Combining its marginals with the induced conditional distribution gives

$$p_T - c_T = J_{FT} \geq 0, \tag{37}$$

$$p_F - c_F = J_{FF} \geq 0, \tag{38}$$

$$c_N - a_F = J_{TN} + J_{NN} \geq 0, \tag{39}$$

$$p_N + a_F - c_N = J_{FN} \geq 0. \tag{40}$$

It follows that

$$c_T \leq p_T, \quad c_F \leq p_F, \quad a_F \leq c_N, \quad c_N \leq p_N + a_F,$$

which are exactly the four constraints in Equation (12).

Sufficiency. Conversely, suppose the four constraints hold. Define the vectors

$$\mathbf{r} = (c_T, c_F, c_N - a_F), \quad (41)$$

$$\mathbf{f} = (p_T - c_T, p_F - c_F, p_N + a_F - c_N), \quad (42)$$

with coordinates ordered as (T, F, N) . The four coherence constraints imply that every coordinate of $\mathbf{r}$ and $\mathbf{f}$ is nonnegative. Because $\mathbf{p}$ and $\mathbf{c}$ are each normalized probability distributions,

$$\sum_{b \in \{T, F, N\}} r_b = 1 - a_F, \quad \sum_{b \in \{T, F, N\}} f_b = a_F, \quad r_b + f_b = p_b.$$

The vector $\mathbf{f}$ can therefore be used as the **False**-antecedent row of the latent joint distribution, while $\mathbf{r}$ supplies the total mass of the **True**- and **Neither**-antecedent rows.

When $a_F < 1$, define

$$J_{Fb} = f_b, \quad J_{Tb} = \frac{a_T}{1 - a_F} r_b, \quad J_{Nb} = \frac{a_N}{1 - a_F} r_b, \quad b \in \{T, F, N\}. \quad (43)$$

All entries are nonnegative. Their row sums are a_F , a_T , and a_N , respectively, and because $a_T + a_N = 1 - a_F$, the column sum for each b is

$$J_{Tb} + J_{Nb} + J_{Fb} = r_b + f_b = p_b.$$

Thus, J has the required marginals. Moreover, the induced conditional has **True** and **False** masses $r_T = c_T$ and $r_F = c_F$, while its **Neither** mass is

$$a_F + r_N = a_F + (c_N - a_F) = c_N.$$

Hence, J reproduces all three measured distributions.

The remaining case is $a_F = 1$. Normalization then implies $a_T = a_N = 0$. Because $a_F \leq c_N$ and $\mathbf{c}$ is normalized, $c_N = 1$ and $c_T = c_F = 0$. Setting

$$J_{Fb} = p_b, \quad J_{Tb} = J_{Nb} = 0$$

for every $b \in \{T, F, N\}$ therefore provides a valid joint distribution with the required marginals and an everywhere-**Neither** conditional.

The four inequalities in Equation (12) are therefore necessary and sufficient for exact pairwise compatibility, provided that the three measured triples are themselves normalized nonnegative probability distributions.

Projection onto the coherent set. The pairwise coherent set $\mathcal{C}$ used to compute d_{CCK} is the intersection of three probability simplexes with the four closed linear half-spaces in Equation (12). It is therefore a nonempty, compact, and convex subset of $\mathbb{R}^9$. Consequently, every measured vector $\mathbf{v}_{P,x}$ has a unique Euclidean projection $\Pi_{\mathcal{C}}(\mathbf{v}_{P,x})$ onto this set, and Equation (13) is well defined.

Scope of the coherence analysis. These conditions characterize compatibility for a single (P, x) pair. Testing them separately across pairs does not establish that all measured probabilities for an LLM can be generated by one global joint distribution over its entire belief system. Such simultaneous compatibility is a stronger global feasibility problem requiring a shared probability distribution over admissible trivalent valuations of all propositions. Our coherence analysis tests the pairwise condition only.

F.2.2 Exact coherence

Table A6 reports exact CCK-coherence rates and individual constraint violations for the primary Direct Conditional estimates obtained using **sAwMIL**. As described in Section 4.5.2, a (P, x) pair is classified as exactly coherent when all four CCK inequalities are satisfied within the numerical tolerance of 10^{-8} . For each inequality, we additionally report how frequently it is violated and the mean signed excess among the pairs that violate it.

Exact coherence is rare for every model and domain (Tab. A6), with the largest observed rate equal to 0.206%. The two most frequently violated conditions are $\pi_F^{\text{direct}}(P, x) \leq \pi_F(P)$ and $\pi_N^{\text{direct}}(P, x) \leq \pi_N(P) + \pi_F(x)$, whereas violations of the remaining two inequalities occur substantially less often. Because exact coherence requires all four constraints to hold simultaneously, these frequent violations drive the near-zero exact-coherence rates reported in the main text and drive the distance-to-coherence analysis in Section 4.5.2.

One likely contributor to this pattern is the scale of the corresponding individual-statement probabilities. Because P is a probe-defined belief, $\pi_F(P)$ and $\pi_N(P)$ are often small. Similarly, because $\mathcal{X}_{\mathcal{M}}$ contains non-disbeliefs, its members x often have small $\pi_F(x)$. Even modest conditional probability mass assigned to **False** or **Neither** can

Dataset	Model	Exact	$\pi_T^{\text{direct}}(P, x) > \pi_T(P)$	$\pi_F^{\text{direct}}(P, x) > \pi_F(P)$	$\pi_F(x) > \pi_N^{\text{direct}}(P, x)$	$\pi_N^{\text{direct}}(P, x) > \pi_N(P) + \pi_F(x)$
City Locations	llama-3.2-3b	0.000%	2.7% (0.095)	90.4% (0.154)	3.4% (0.122)	96.6% (0.207)
	llama-3.1-8b	0.000%	5.2% (0.054)	88.0% (0.051)	5.8% (0.047)	94.2% (0.080)
	llama-3.1-70b	< 0.001%	7.9% (0.031)	80.5% (0.035)	11.4% (0.026)	88.5% (0.047)
	gemma-7b	< 0.001%	3.8% (0.107)	92.4% (0.159)	7.9% (0.091)	92.1% (0.102)
	gemma-2-9b	0.000%	3.1% (0.050)	93.9% (0.087)	11.0% (0.036)	89.0% (0.046)
	gemma-2-27b	0.000%	2.4% (0.067)	93.7% (0.104)	5.9% (0.073)	94.1% (0.109)
	mistral-7b	0.000%	3.5% (0.097)	91.7% (0.077)	8.8% (0.053)	91.2% (0.093)
	mistral-12b	0.004%	9.0% (0.055)	62.5% (0.082)	11.1% (0.075)	88.7% (0.112)
	mistral-3.1-24b	0.000%	8.8% (0.027)	86.0% (0.074)	21.2% (0.022)	78.8% (0.068)
	qwen-2.5-7b	< 0.001%	4.6% (0.052)	88.0% (0.125)	16.6% (0.052)	83.4% (0.112)
	qwen-2.5-14b	0.000%	8.8% (0.032)	78.8% (0.066)	18.5% (0.037)	81.5% (0.077)
	qwen-2.5-72b	0.000%	8.4% (0.031)	74.4% (0.056)	15.0% (0.039)	85.0% (0.077)
Medical Indications	llama-3.2-3b	0.000%	3.6% (0.082)	63.8% (0.153)	12.1% (0.138)	87.9% (0.276)
	llama-3.1-8b	0.008%	5.2% (0.092)	59.7% (0.111)	13.2% (0.132)	86.7% (0.252)
	llama-3.1-70b	0.000%	7.4% (0.088)	65.8% (0.111)	13.6% (0.127)	86.4% (0.179)
	gemma-7b	0.097%	2.1% (0.095)	60.6% (0.152)	20.8% (0.132)	79.0% (0.250)
	gemma-2-9b	0.000%	8.4% (0.094)	72.1% (0.119)	21.8% (0.129)	78.1% (0.148)
	gemma-2-27b	0.015%	9.4% (0.103)	68.9% (0.100)	20.6% (0.134)	79.3% (0.150)
	mistral-7b	< 0.001%	7.3% (0.085)	63.8% (0.090)	21.6% (0.134)	78.4% (0.183)
	mistral-12b	0.200%	13.4% (0.099)	45.4% (0.107)	31.7% (0.132)	67.9% (0.184)
	mistral-3.1-24b	0.015%	11.9% (0.096)	58.2% (0.054)	19.8% (0.134)	80.1% (0.128)
	qwen-2.5-7b	0.035%	3.5% (0.087)	75.8% (0.184)	23.4% (0.132)	76.2% (0.204)
	qwen-2.5-14b	0.000%	6.3% (0.088)	72.9% (0.125)	25.3% (0.127)	74.7% (0.178)
	qwen-2.5-72b	0.000%	7.1% (0.086)	74.7% (0.079)	22.6% (0.114)	77.4% (0.144)
Word Definitions	llama-3.2-3b	0.073%	6.6% (0.091)	64.2% (0.100)	5.7% (0.106)	93.9% (0.223)
	llama-3.1-8b	0.020%	4.6% (0.086)	77.5% (0.103)	6.8% (0.125)	93.0% (0.177)
	llama-3.1-70b	0.018%	6.5% (0.085)	73.7% (0.103)	4.7% (0.121)	95.1% (0.185)
	gemma-7b	0.206%	2.7% (0.088)	64.1% (0.140)	8.1% (0.118)	91.4% (0.259)
	gemma-2-9b	0.012%	8.2% (0.109)	82.3% (0.135)	10.1% (0.123)	88.8% (0.120)
	gemma-2-27b	0.003%	3.6% (0.109)	87.2% (0.145)	5.1% (0.110)	94.7% (0.169)
	mistral-7b	0.030%	4.7% (0.094)	72.5% (0.097)	6.4% (0.124)	93.2% (0.207)
	mistral-12b	0.054%	2.7% (0.092)	73.1% (0.132)	5.6% (0.126)	94.3% (0.323)
	mistral-3.1-24b	0.026%	7.0% (0.101)	76.0% (0.103)	7.8% (0.125)	91.8% (0.173)
	qwen-2.5-7b	0.080%	7.7% (0.108)	76.2% (0.135)	10.9% (0.120)	87.6% (0.189)
	qwen-2.5-14b	0.008%	6.2% (0.093)	79.7% (0.132)	9.0% (0.113)	90.7% (0.162)
	qwen-2.5-72b	0.021%	3.8% (0.082)	82.0% (0.164)	8.2% (0.117)	91.6% (0.189)

Table A6. Exact CCK coherence and constraint violations for Direct Conditional estimates using sAwMIL. For each LLM and domain, we report the percentage of (P, x) pairs satisfying all four CCK constraints simultaneously within tolerance 10^{-8} . Here, 0.000% signifies exactly zero fully coherent pairs, while < 0.001% means a small number of exactly coherent pairs exists. Each remaining column reports the percentage of pairs violating each constraint, with the mean excess across violating pairs shown in parentheses. Exact coherence is rare across all models and domains, with no model–domain combination exceeding 0.206%. Violations occur most frequently for the constraints on conditional **False** and **Neither** probability mass.

therefore violate the second or fourth constraint. The fourth constraint may additionally be sensitive to the Direct Conditional representation itself: in “Given x , P ,” the conditioning proposition precedes P in the autoregressive sequence, allowing the representation of the conditional to incorporate x and potentially shift probability mass toward **Neither** relative to P in isolation.

G Additional behavioral results

G.1 Matched-pairs analysis

Table A7 reports diagnostics for the probability-matched pairs used in our primary behavioral analysis. As described in Section 4.6, matching is performed separately within each challenge-sequence block so that both propositions in a pair receive the same conversational challenges in the same order. The reported sample additionally requires both propositions’ initial behavioral judgments to agree with their individual-statement probe classifications, as in the primary analysis. For each model, n denotes the number of pairs, while $|\Delta\pi_T|$ and $|\Delta\gamma|$ quantify pairwise differences in individual belief probability and graded stability, respectively.

Across all three domains, matching yields small differences in individual belief probability while retaining substantially larger differences in graded stability. Mean $|\Delta\pi_T|$ is at most 0.00553 across the reported model–domain combinations, whereas mean $|\Delta\gamma|$ ranges from 0.066 to 0.300. Thus, the matched pairs are closely aligned in individual belief probability while preserving meaningful variation in graded stability.

Dataset	Model	n	$ \Delta\pi_T $	$ \Delta\gamma $
City Locations	llama-3.2-3b	315	0.00435 (0.01055)	0.235 (0.173)
	llama-3.1-8b	305	0.00181 (0.00556)	0.090 (0.122)
	llama-3.1-70b	287	0.00483 (0.02893)	0.066 (0.116)
	gemma-7b	304	0.00377 (0.01042)	0.175 (0.175)
	gemma-2-9b	291	0.00370 (0.01686)	0.082 (0.141)
	gemma-2-27b	316	0.00362 (0.01372)	0.133 (0.143)
	mistral-7b	303	0.00456 (0.01924)	0.113 (0.157)
	mistral-12b	278	0.00495 (0.01547)	0.146 (0.189)
	mistral-3.1-24b	307	0.00362 (0.01821)	0.100 (0.151)
	qwen-2.5-7b	312	0.00350 (0.01001)	0.136 (0.157)
	qwen-2.5-14b	290	0.00314 (0.01289)	0.103 (0.159)
	qwen-2.5-72b	276	0.00282 (0.01189)	0.096 (0.178)
Medical Indications	llama-3.2-3b	342	0.00442 (0.00530)	0.234 (0.185)
	llama-3.1-8b	344	0.00389 (0.00518)	0.248 (0.204)
	llama-3.1-70b	340	0.00404 (0.00763)	0.198 (0.181)
	gemma-7b	356	0.00430 (0.00639)	0.266 (0.217)
	gemma-2-9b	344	0.00381 (0.00535)	0.210 (0.206)
	gemma-2-27b	338	0.00408 (0.00648)	0.182 (0.204)
	mistral-7b	329	0.00456 (0.00993)	0.179 (0.223)
	mistral-12b	339	0.00515 (0.00856)	0.200 (0.204)
	mistral-3.1-24b	337	0.00379 (0.00564)	0.138 (0.176)
	qwen-2.5-7b	353	0.00441 (0.00579)	0.263 (0.206)
	qwen-2.5-14b	354	0.00406 (0.00494)	0.232 (0.206)
	qwen-2.5-72b	337	0.00401 (0.00719)	0.155 (0.191)
Word Definitions	llama-3.2-3b	303	0.00510 (0.00960)	0.221 (0.233)
	llama-3.1-8b	316	0.00408 (0.00787)	0.169 (0.191)
	llama-3.1-70b	316	0.00429 (0.00772)	0.199 (0.205)
	gemma-7b	317	0.00538 (0.00766)	0.262 (0.230)
	gemma-2-9b	300	0.00488 (0.01071)	0.179 (0.190)
	gemma-2-27b	307	0.00460 (0.01048)	0.204 (0.200)
	mistral-7b	315	0.00475 (0.00910)	0.201 (0.218)
	mistral-12b	308	0.00553 (0.01139)	0.300 (0.254)
	mistral-3.1-24b	306	0.00462 (0.00839)	0.201 (0.218)
	qwen-2.5-7b	311	0.00447 (0.00834)	0.189 (0.199)
	qwen-2.5-14b	308	0.00459 (0.00962)	0.169 (0.187)
	qwen-2.5-72b	305	0.00474 (0.00925)	0.253 (0.226)

Table A7. Behavioral matching quality. For each LLM and domain, n is the number of matched proposition pairs, $|\Delta\pi_T|$ is the absolute difference in belief probability between the two propositions in each pair, and $|\Delta\gamma|$ is their absolute difference in graded stability. Values for $|\Delta\pi_T|$ and $|\Delta\gamma|$ report the mean with standard deviation in parentheses. Across domains and models, matching produces small differences in individual belief probability while retaining substantially larger differences in graded stability.

G.2 Uncertainty in behavioral resilience effects

Figure 5 reports ± 1 bootstrap standard error to visualize the precision of each model–domain effect. Table A8 reports the corresponding 95% percentile bootstrap confidence intervals for mean ΔM . Mean ΔM is positive in 83.3% of model–domain settings. The 95% confidence interval is entirely positive in 11 of 36 (30.6%) settings, entirely negative in 1 of 36 (2.8%), and includes zero in the remaining 24 of 36 (66.7%).

Dataset	Model	Mean ΔM	Bootstrap SE	95% CI
City Locations	llama-3.2-3b	-0.0256	0.0118	[-0.0488, -0.0018]
	llama-3.1-8b	0.0220	0.0308	[-0.0402, 0.0815]
	llama-3.1-70b	0.0001	0.0049	[-0.0098, 0.0095]
	gemma-7b	0.0101	0.0080	[-0.0058, 0.0256]
	gemma-2-9b	0.0180	0.0077	[0.0030, 0.0329]
	gemma-2-27b	0.0784	0.0144	[0.0497, 0.1069]
	mistral-7b	0.0426	0.0147	[0.0156, 0.0730]
	mistral-12b	0.0938	0.0179	[0.0582, 0.1289]
	mistral-3.1-24b	-0.0012	0.0076	[-0.0160, 0.0137]
	qwen-2.5-7b	0.0107	0.0095	[-0.0080, 0.0293]
	qwen-2.5-14b	0.0089	0.0068	[-0.0040, 0.0227]
	qwen-2.5-72b	0.0107	0.0064	[-0.0015, 0.0238]
Medical Indications	llama-3.2-3b	-0.0070	0.0153	[-0.0368, 0.0231]
	llama-3.1-8b	0.0255	0.0307	[-0.0357, 0.0860]
	llama-3.1-70b	0.0097	0.0115	[-0.0131, 0.0323]
	gemma-7b	0.0764	0.0169	[0.0438, 0.1096]
	gemma-2-9b	0.0522	0.0131	[0.0265, 0.0778]
	gemma-2-27b	0.0289	0.0167	[-0.0040, 0.0611]
	mistral-7b	0.0256	0.0213	[-0.0160, 0.0680]
	mistral-12b	0.0344	0.0173	[0.0003, 0.0687]
	mistral-3.1-24b	0.0005	0.0076	[-0.0140, 0.0156]
	qwen-2.5-7b	0.1025	0.0147	[0.0740, 0.1313]
	qwen-2.5-14b	0.0081	0.0125	[-0.0162, 0.0328]
	qwen-2.5-72b	-0.0058	0.0112	[-0.0273, 0.0160]
Word Definitions	llama-3.2-3b	0.0050	0.0252	[-0.0418, 0.0561]
	llama-3.1-8b	0.0607	0.0558	[-0.0515, 0.1659]
	llama-3.1-70b	0.0211	0.0154	[-0.0088, 0.0510]
	gemma-7b	0.0408	0.0139	[0.0135, 0.0681]
	gemma-2-9b	0.0346	0.0164	[0.0034, 0.0672]
	gemma-2-27b	0.0145	0.0228	[-0.0302, 0.0584]
	mistral-7b	0.0568	0.0281	[0.0009, 0.1102]
	mistral-12b	0.0257	0.0373	[-0.0471, 0.0995]
	mistral-3.1-24b	0.0206	0.0148	[-0.0088, 0.0491]
	qwen-2.5-7b	0.0226	0.0171	[-0.0116, 0.0555]
	qwen-2.5-14b	-0.0039	0.0114	[-0.0266, 0.0179]
	qwen-2.5-72b	-0.0096	0.0191	[-0.0476, 0.0266]

Table A8. Uncertainty in behavioral resilience effects. We report the mean behavioral movement difference ΔM , bootstrap standard error, and 95% percentile bootstrap confidence interval using the **sAwMIL** probe and Direct Conditional graded stability. Uncertainty is estimated from 10,000 sequence-stratified matched-pair bootstrap resamples. Mean ΔM is positive in 83.3% of model-domain settings; the 95% confidence interval is entirely positive in 30.6% of settings, entirely negative in 2.8%, and includes zero in the remaining 66.7%.

Parameter count

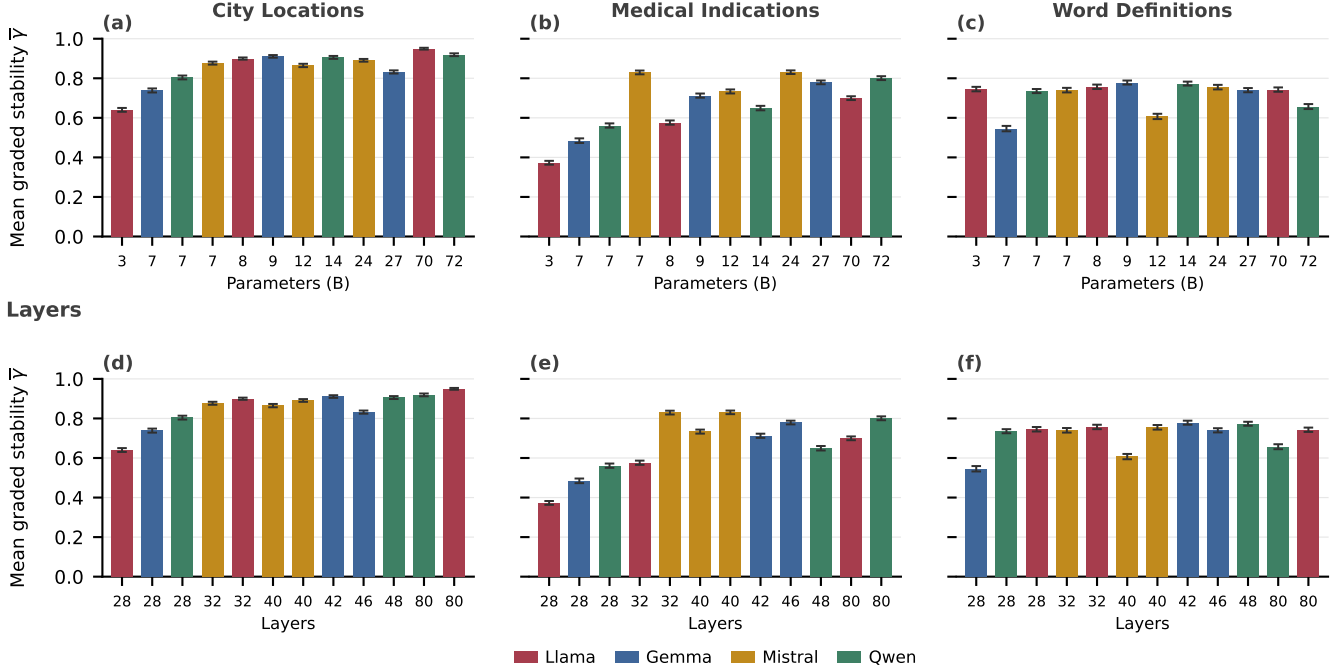


Figure A3. Association of model scale with graded stability. Bars show mean graded stability $\bar{\gamma}_{\mathcal{M}}$ for the 12 instruction-tuned LLMs, ordered by nominal parameter count in (a) City Locations, (b) Medical Indications, and (c) Word Definitions, and by number of decoder layers in (d) City Locations, (e) Medical Indications, and (f) Word Definitions. Colors denote Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) model families. Error bars denote ± 1 bootstrap standard error. Neither parameter count nor model depth shows a consistent relationship with graded stability.

H Model scaling

The between-model variation in graded stability raises the possibility that $\gamma_{\mathcal{M}}(P)$ increases systematically with model scale. We examine this relationship using two measures of scale: parameter count and number of layers. For each domain, we compute the mean graded stability $\bar{\gamma}$ across each model’s beliefs and order the 12 instruction-tuned LLMs by nominal parameter count (Fig. A3(a–c)) and model depth (Fig. A3(d–f)).

We do not observe a consistent scaling relationship under either measure. Mean stability generally increases across some of the smaller models in City Locations and Medical Indications, but these trends are neither monotonic nor consistent across domains. Word Definitions shows little evidence of increasing stability with either parameter count or model depth. Moreover, models with the same nominal parameter count or the same number of layers can exhibit substantially different mean stability. Thus, although model scale may contribute to some of the within-family differences observed in Section 2.4, neither parameter count nor model depth provides a systematic explanation for variation in graded belief stability across models.

I Robustness checks

I.1 Base-model robustness

To test whether our primary findings depend on instruction tuning, we repeat the main-text analyses using the pretrained base counterparts of the 12 instruction-tuned LLMs. The models are matched by family and parameter scale, with exact specifications reported in Section C.

I.1.1 Replication of primary analyses in base models

The principal representation-based findings are qualitatively similar in the base models. Individual belief probability continues to explain a meaningful but incomplete fraction of graded stability (Fig. A4), and the residual stability

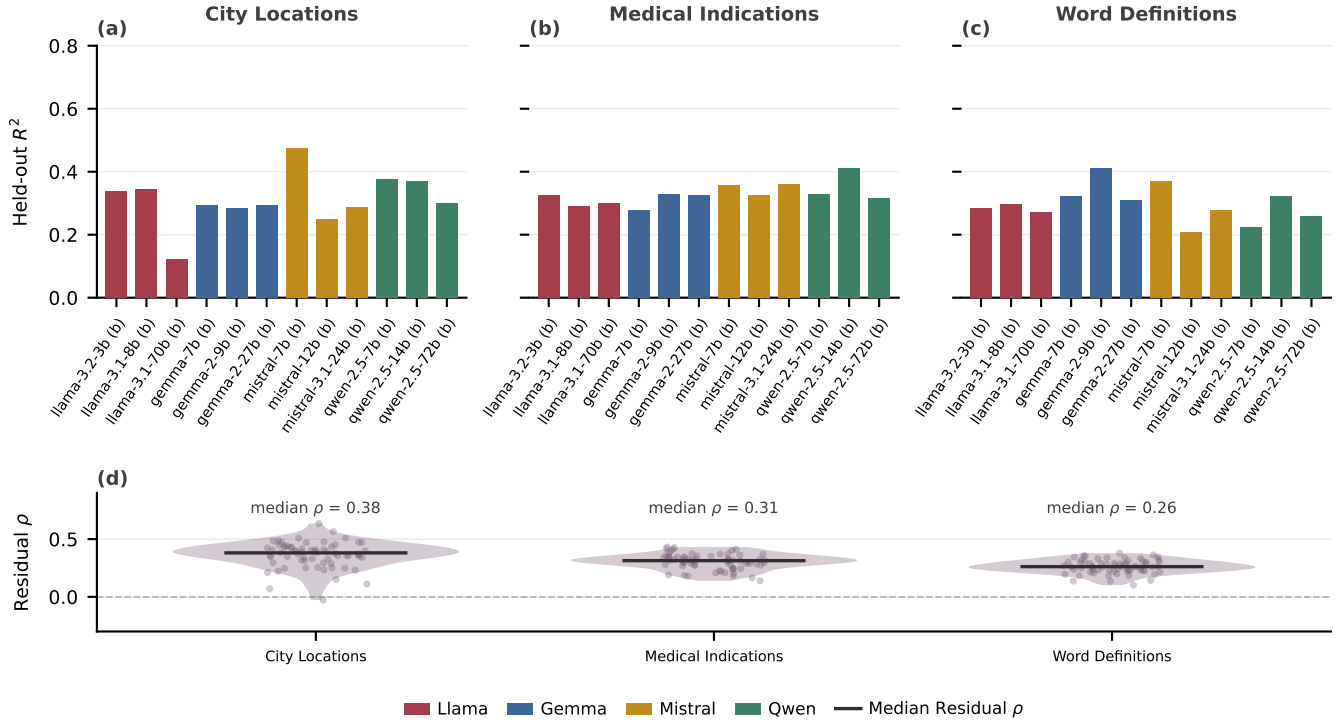


Figure A4. Relationship between individual belief probability and graded stability in base models. We display the held-out R^2 for predicting graded stability from individual belief probability in (a) City Locations, (b) Medical Indications, and (c) Word Definitions for the 12 matched pretrained base models from the Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) families. Panel (d) shows the distribution of pairwise Spearman correlations between models' residual graded-stability scores after removing the fitted belief-probability relationship, with black lines denoting median correlations. As with the instruction-tuned models, belief probability explains only part of graded stability and the remaining proposition-level structure is positively shared across models.

remaining after accounting for belief probability exhibits positive cross-model structure. Median residual correlations are $\rho = 0.38$ for City Locations, $\rho = 0.31$ for Medical Indications, and $\rho = 0.26$ for Word Definitions.

The approximate-coherence result also persists in the base models (Fig. A5). Most measured probability systems remain relatively close to the CCK-coherent set, with City Locations generally requiring the smallest adjustments and Medical Indications and Word Definitions showing larger distances.

Domain-level variation in graded stability is likewise broadly preserved (Fig. A6). City Locations remains concentrated near high stability for most base models, whereas Medical Indications and Word Definitions generally exhibit broader distributions and more mass at intermediate stability values. The base models typically show greater between-model heterogeneity than the instruction-tuned models, particularly for the Medical Indications and Word Definitions domains. There are, however, several notable exceptions to the overall pattern. In City Locations, for example, Llama-3.2-3B is substantially more stable than in the instruction-tuned setting, whereas Mistral-12B is markedly less stable.

The behavioral replication is weaker (Fig. A7). Unlike the instruction-tuned models, for which lower-stability beliefs exhibit greater mean movement in most model-domain settings, the base-model effects are generally smaller and cluster more closely around zero. We interpret this difference cautiously because the behavioral challenge paradigm is itself an instruction-following, multi-turn interaction. The pretrained base models have not undergone the instruction-following post-training of their matched instruction-tuned counterparts, and in our implementation the conversational history is therefore supplied to base models as an explicit plain-text transcript rather than through a model-specific chat template. The weaker base-model behavioral association may therefore reflect the mismatch between the challenge task and the models' training rather than a failure of the underlying representational stability measure.

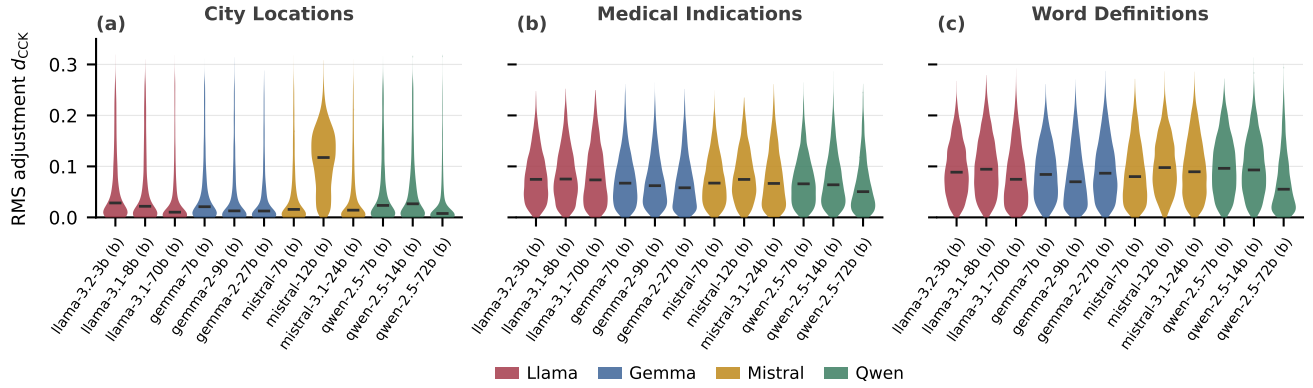


Figure A5. Distance from probabilistic coherence in base models. Violin plots show the distribution across (P, x) pairs of the RMS adjustment d_{CCK} required to project the measured probability distributions for individual statements and conditionals onto the nearest CCK-coherent probability system for (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Horizontal black lines denote within-model medians. Base models remain comparatively close to the coherent set, although distances and between-model heterogeneity are slightly larger than in the instruction-tuned models.

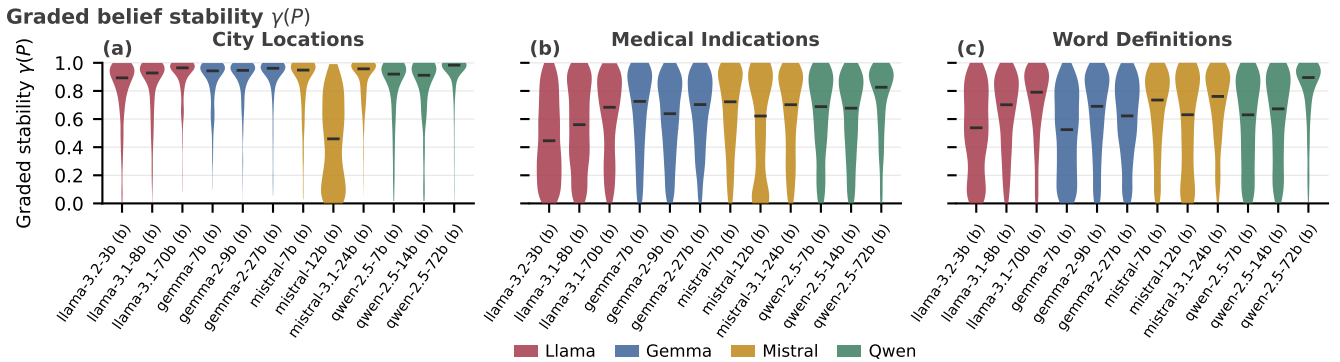


Figure A6. Variation in graded belief stability across domains in base models. Violin plots show the proposition-level distribution of Direct Conditional graded stability for each of the pretrained base models in (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Horizontal black lines denote within-model medians. City Locations is generally concentrated at higher stability values, while Medical Indications and Word Definitions show broader and more model-dependent distributions, reproducing the main domain-level pattern with greater heterogeneity.

1.1.2 Association with instruction tuning

We next compare each base model directly with its instruction-tuned counterpart while holding belief identity fixed. For each matched model pair, we restrict the comparison to propositions believed by both models and define the average association with instruction tuning as

$$\Delta_{\text{inst}} := \frac{1}{|\mathcal{B}_{\text{base}} \cap \mathcal{B}_{\text{inst}}|} \sum_{P \in \mathcal{B}_{\text{base}} \cap \mathcal{B}_{\text{inst}}} [\gamma_{\text{inst}}(P) - \gamma_{\text{base}}(P)]. \quad (44)$$

Thus, $\Delta_{\text{inst}} > 0$ indicates that instruction tuning is associated with greater graded stability for the same set of shared beliefs. Across the 36 matched model-domain comparisons, 28 (77.8%) have $\Delta_{\text{inst}} > 0$ (Fig. A8). The effect is therefore positive more often than not, but its magnitude varies substantially across models and domains and several comparisons are negative. We interpret instruction tuning as exhibiting a positive directional tendency rather than a uniform stabilizing effect.

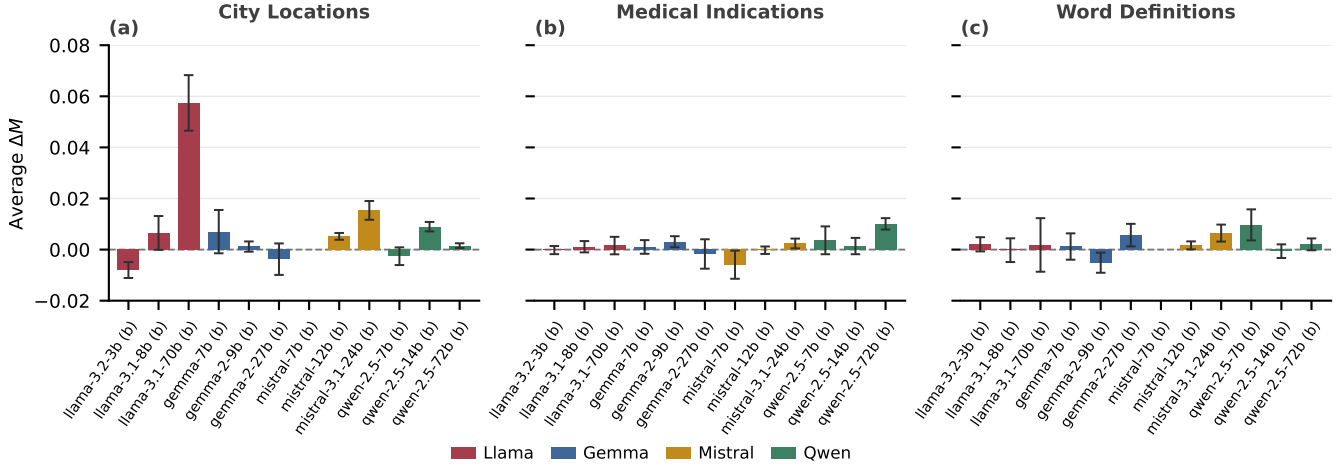


Figure A7. Behavioral resilience among probability-matched beliefs in base models. Bars show the mean difference in behavioral movement ΔM between matched lower- and higher-stability beliefs for the pretrained base models in (a) City Locations, (b) Medical Indications, and (c) Word Definitions, with error bars denoting ± 1 bootstrap standard error, for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Positive values indicate greater movement for the lower-stability belief. In contrast to the clearer directional association in instruction-tuned models, base-model effects are smaller and more heterogeneous, suggesting that the behavioral relationship is sensitive to instruction-following post-training.

Effect of instruction tuning on graded stability

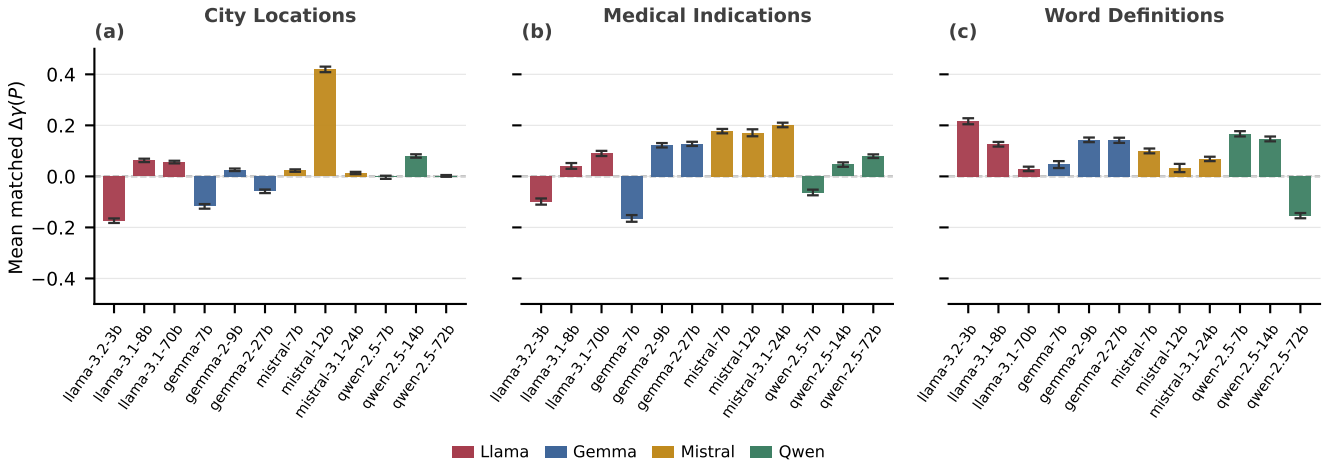


Figure A8. Association between instruction tuning and graded belief stability. Bars show the mean matched difference Δ_{inst} in graded stability between each instruction-tuned model and its pretrained base counterpart over propositions believed by both models for (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Positive values indicate greater graded stability after instruction tuning, and error bars capture ± 1 standard error. Instruction-tuned models have higher measured stability in 77.8% of matched comparisons, but the magnitude and direction of the effect remain heterogeneous.

1.2 Conditional-estimator robustness

Our primary analyses use the Direct Conditional estimator, which estimates conditional support from representations of statements of the form “Given x , P .” To test whether our results depend on this operationalization of $\Pr_{\mathcal{M}}(P | x)$, we repeat the principal analyses using the Joint-to-Conditional estimator described in Section 4.3.2. While Direct Conditional probes the conditional statement itself, Joint-to-Conditional represents x and P jointly and derives the corresponding conditional probability from the resulting nine-state distribution. We first examine an additional

representational choice introduced by the Joint estimator, the ordering of x and P in the conjunction, before replicating the main analyses and directly comparing the two graded-stability estimates.

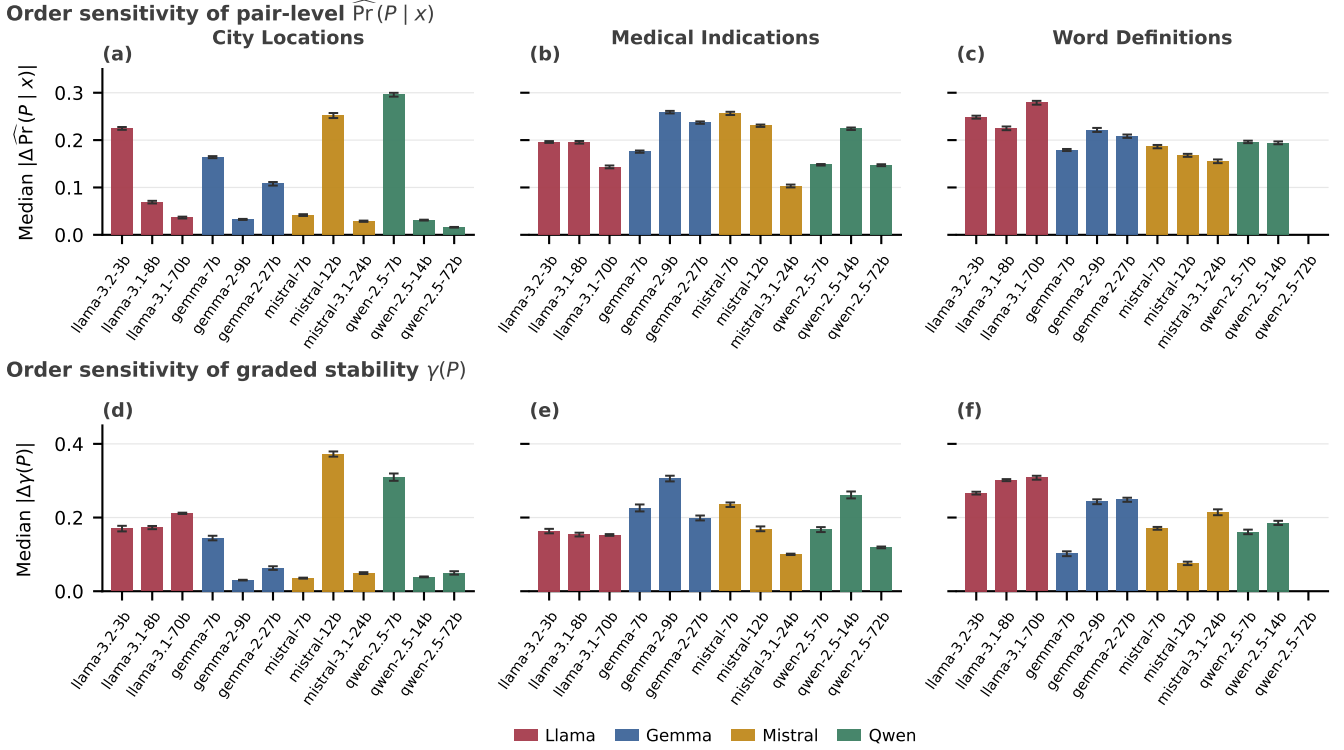


Figure A9. Sensitivity of the Joint-to-Conditional estimator to conjunction ordering. We compare Joint estimates obtained using the templates “ x and P ” and “ P and x ”. For each model, bars show the median absolute change in $\widehat{\Pr}_{\mathcal{M}}^{\text{joint}}(P | x)$ across evaluated (P, x) combinations when the conjunction order is reversed, for (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). We additionally display (d–f) the median absolute change in $\gamma_{\mathcal{M}}^{\text{joint}}(P)$ across target beliefs after graded stability is recomputed using the reversed-order estimates. Error bars denote ± 1 bootstrap standard error. Reversing the conjunction order changes both the estimated conditional probabilities and the resulting graded-stability scores, revealing an order sensitivity specific to the Joint-to-Conditional operationalization.

1.2.1 Sensitivity to conjunction ordering

The Joint-to-Conditional estimator requires a natural-language representation of the two statements being jointly probed. In our primary implementation, we use “ x and P ”, placing the conditioning proposition x first and the target belief P second. The calibrated Joint probe produces the nine-state distribution described in Section 4.3.2, from which we derive $\widehat{\Pr}_{\mathcal{M}}^{\text{joint}}(P | x)$. To test whether this estimate depends on the conjunction order, we independently train and calibrate a second Joint probe using the reversed template “ P and x ”.

Figure A9 shows that this choice is consequential. Reversing the conjunction changes the estimated conditional probabilities, with the size of the change varying substantially across models and domains (Fig. A9(a–c)). These differences also affect the downstream graded stability measure (Fig. A9(d–f)). We treat this sensitivity as an additional measurement choice introduced by the Joint-to-Conditional estimator. An autoregressive LLM produces different hidden representations when the same two propositions appear in different token orders, and its training objective does not require the probe-relevant representations of “ x and P ” and “ P and x ” to be invariant. As such, the observed differences should not necessarily be interpreted as evidence about an underlying order-sensitive joint belief distribution. Rather, they show that the Joint operationalization inherits sensitivity to the linguistic representation used to elicit that distribution. We fix the order to “ x and P ” throughout the analyses below, and Direct Conditional remains our primary estimator because it does not introduce this additional conjunction-ordering choice.

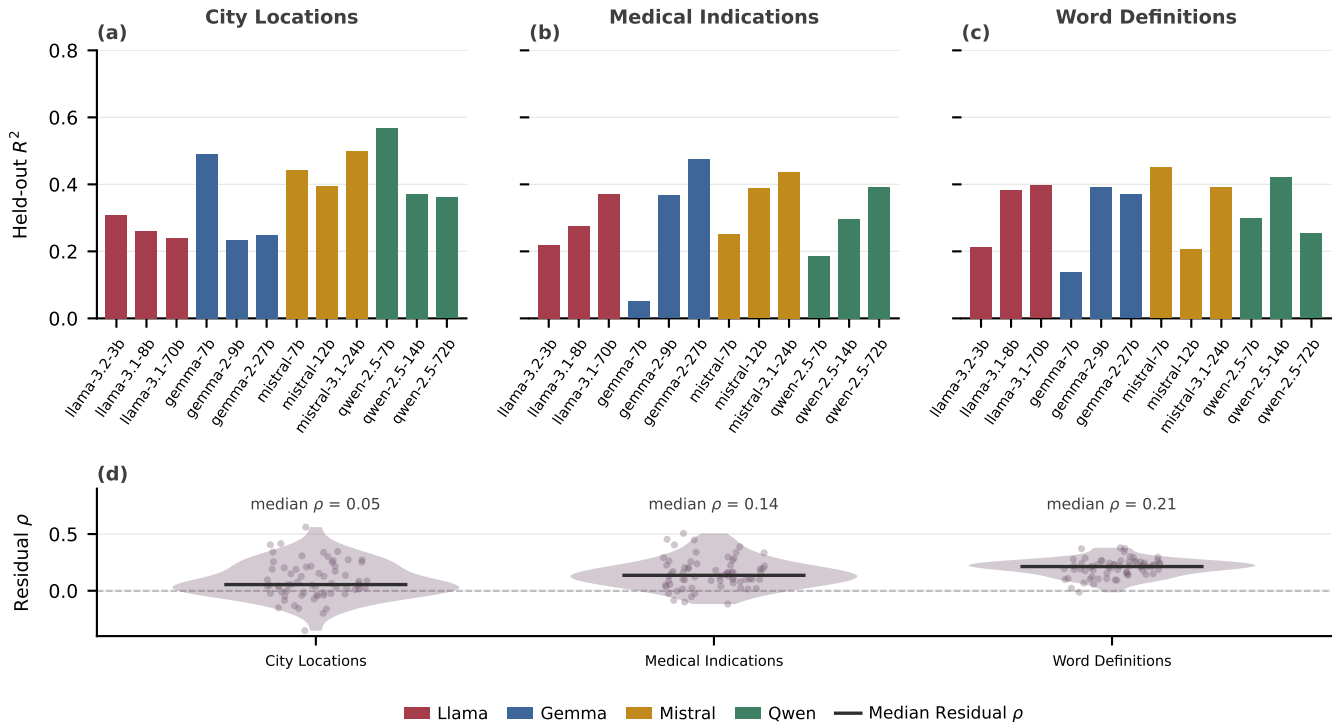


Figure A10. Relationship between individual belief probability and Joint-to-Conditional graded stability. We display the held-out R^2 for predicting Joint-to-Conditional graded stability from individual belief probability in (a) City Locations, (b) Medical Indications, and (c) Word Definitions for the 12 instruction-tuned LLMs for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Panel (d) shows the distribution of pairwise Spearman correlations between models’ residual graded-stability scores after removing the fitted belief-probability relationship, with black lines denoting median correlations. As with Direct Conditional, individual belief probability explains only part of graded stability, and residual stability remains positively correlated across most model pairs.

1.2.2 Replication with Joint-to-Conditional

We repeat the main-text analyses using Joint-to-Conditional graded stability for the same 12 instruction-tuned LLMs (Figs. A10–A13). The principal representation-based findings are broadly preserved, while the absolute scale of graded stability and its behavioral association show greater sensitivity to the conditional estimator.

Individual belief probability remains an incomplete predictor of Joint-to-Conditional graded stability (Fig. A10). Held-out R^2 remains well below one across every model and domain. Cross-model residual correlations are also predominantly positive, although less uniformly than under Direct Conditional: 71.2% are positive for City Locations, 89.4% for Medical Indications, and 98.5% for Word Definitions. The corresponding median residual Spearman correlations are $\rho = 0.05$, $\rho = 0.14$, and $\rho = 0.21$, respectively. Thus, the Joint estimator preserves the central result that graded stability contains systematic variation not captured by individual belief probability, while the strength of the shared residual structure is estimator-dependent.

Using the same CCK projection analysis as in Section 4.5.2, but substituting the Joint-derived conditional estimates, we find that the measured probability systems again lie relatively close to the coherent set (Fig. A11). Importantly, coherence is not guaranteed by the Joint construction: the individual-statement distributions for P and x are estimated independently rather than obtained as marginals of the probed joint distribution. City Locations is typically closest to the coherent set, while Medical Indications and Word Definitions show larger adjustments.

The clearest estimator-dependent effect appears in the absolute distribution of graded stability (Fig. A12). Joint-to-Conditional scores are strongly concentrated near $\gamma = 1$ across all three domains. City Locations remains highly stable, but Medical Indications and Word Definitions are also shifted substantially toward the upper end of the scale, reducing the domain separation observed under Direct Conditional. This difference is consistent with the construction of the estimators: Joint-to-Conditional renormalizes over determinate conditional states, whereas Direct Conditional allows Neither probability mass to reduce the estimated support for $P | x$. The pronounced

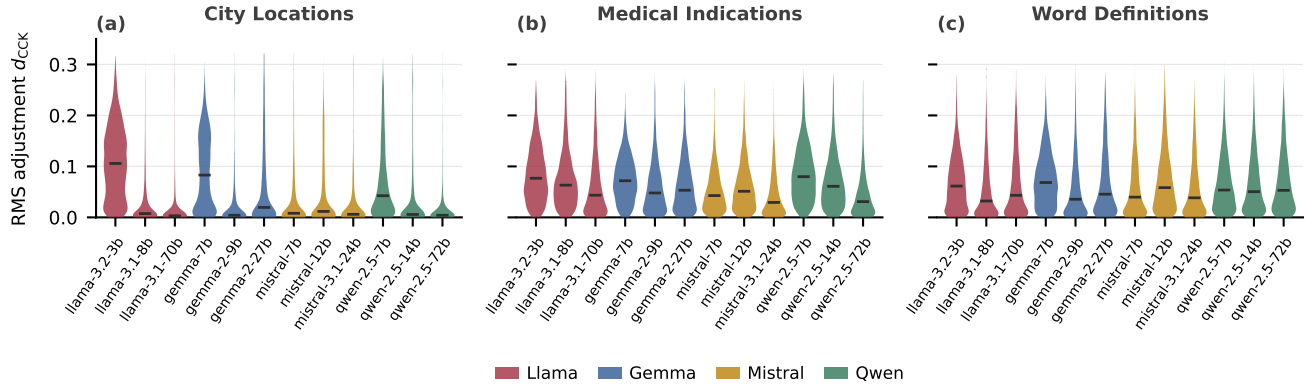


Figure A11. Distance of Joint-to-Conditional probability estimates from probabilistic coherence. Violin plots show the distribution across (P, x) combinations of the RMS adjustment d_{CCK} required to project measured probability distributions for individual statements and conditionals onto the nearest CCK-coherent probability system for (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Horizontal black lines denote within-model medians. As under Direct Conditional, the measured systems generally lie close to the CCK-coherent set, with City Locations showing the smallest distances for most models.

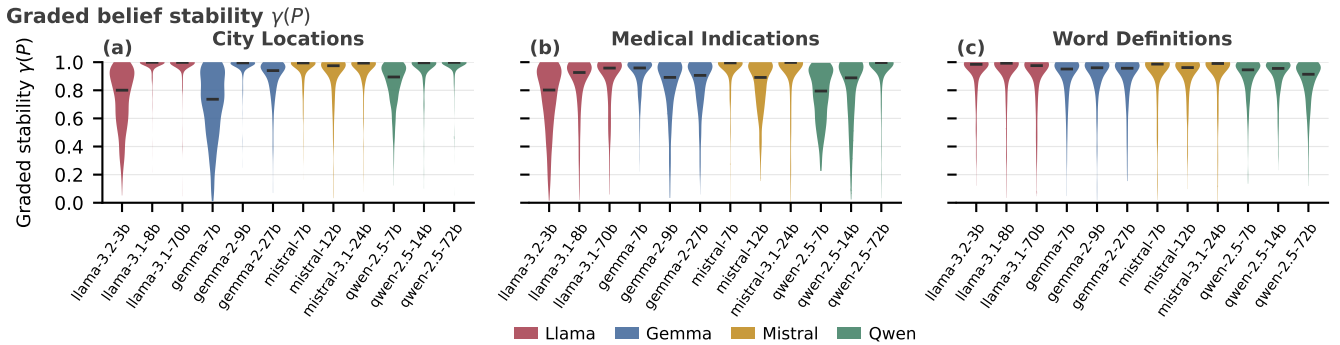


Figure A12. Variation in Joint-to-Conditional graded belief stability across domains. Violin plots show the proposition-level distribution of Joint-to-Conditional graded stability for each of the 12 instruction-tuned LLMs in (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Horizontal black lines denote within-model medians. Joint-to-Conditional stability is strongly concentrated near $\gamma = 1$ across all three domains, producing weaker domain separation than under Direct Conditional.

domain ordering in the primary analysis should therefore be interpreted as partly dependent on the numerical operationalization of conditional support, even though Direct and Joint stability remain strongly related within models (Section 1.2.3).

The behavioral association is weaker but remains more often positive than negative under Joint-to-Conditional (Fig. A13). Mean ΔM is positive in 26/36 (72.2%) model–domain settings: 10/12 for City Locations, 9/12 for Medical Indications, and 7/12 for Word Definitions. Thus, probability-matched lower-stability beliefs still tend to move more under conversational challenge, but the relationship is less consistent than under Direct Conditional.

1.2.3 Direct–Joint agreement

Finally, we ask whether the two estimators preserve the relative ordering of beliefs even where their absolute stability values differ. For each model and domain, we compute the Spearman correlation between $\gamma_{\mathcal{M}}^{\text{direct}}(P)$ and $\gamma_{\mathcal{M}}^{\text{joint}}(P)$ over propositions with valid stability estimates under both approaches (Fig. A14).

The two operationalizations produce positively correlated stability rankings in every model–domain setting. Median Spearman correlations across models are $\rho = 0.78$ for City Locations, $\rho = 0.81$ for Medical Indications, and

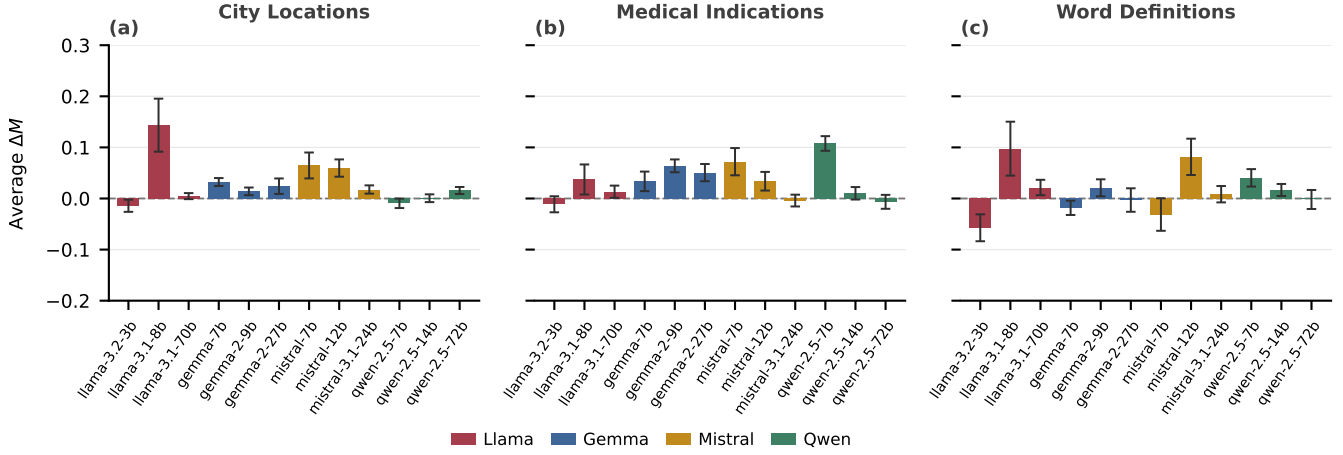


Figure A13. Behavioral resilience using Joint-to-Conditional graded stability among probability-matched beliefs. Bars show the mean difference in behavioral movement ΔM between matched lower- and higher-stability beliefs for the 12 instruction-tuned LLMs in (a) City Locations, (b) Medical Indications, and (c) Word Definitions, with error bars denoting ± 1 bootstrap standard error, for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Positive values indicate greater movement for the lower-stability belief. Mean ΔM is positive in 26/36 (72.2%) model-domain settings, although the association is less consistent than under Direct Conditional.

Agreement between graded stability operationalizations

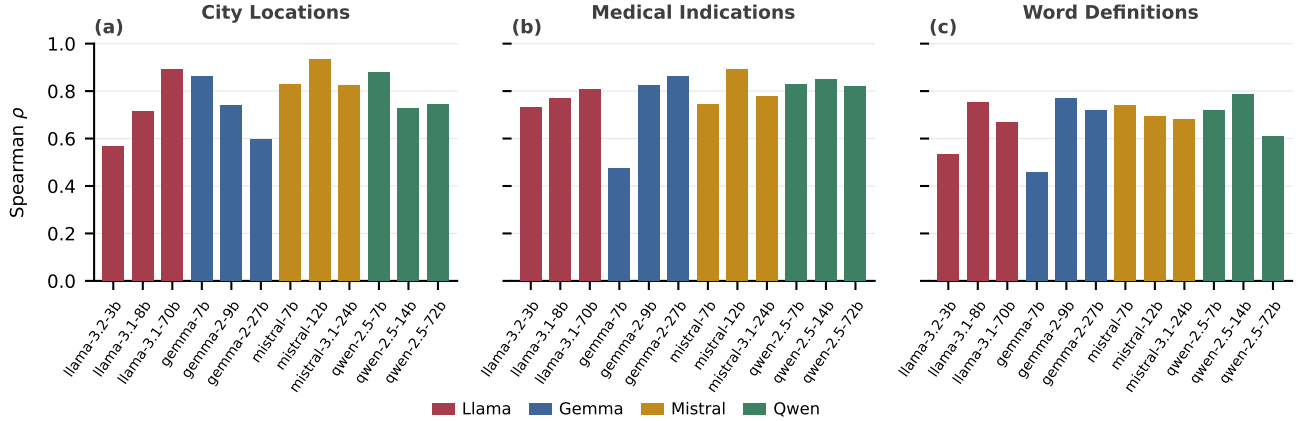


Figure A14. Agreement between Direct Conditional and Joint-to-Conditional graded stability. Bars show the Spearman correlation between $\gamma_{\mathcal{M}}^{\text{direct}}(P)$ and $\gamma_{\mathcal{M}}^{\text{joint}}(P)$ for each instruction-tuned LLM in (a) City Locations, (b) Medical Indications, and (c) Word Definitions for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Correlations are positive for every model and domain, indicating substantial agreement in the relative ranking of beliefs by stability despite differences in the estimators' absolute scales.

$\rho = 0.71$ for Word Definitions. Thus, despite their differences in representation and absolute numerical scale, Direct Conditional and Joint-to-Conditional generally agree on which beliefs are relatively more or less stable within a model.

1.3 Alternative-probe robustness

We repeat the principal analyses using two single-instance linear probes, an SVM [30] and Mass Mean [17], to evaluate whether the graded-stability results depend on the sAwMIL probe used in the main text. Both alternative probes operate on the hidden representation of the final non-padding token. For each probe, model, and dataset, we independently select the transformer layer that minimizes three-class calibration log loss using the procedure in

Model	City Locations		Medical Indications		Word Definitions	
	SVM	Mass Mean	SVM	Mass Mean	SVM	Mass Mean
llama-3.2-3b	11	7	12	11	11	11
llama-3.1-8b	25	17	14	13	13	14
llama-3.1-70b	76	36	65	37	29	30
gemma-7b	17	19	17	17	16	16
gemma-2-9b	22	23	20	21	20	19
gemma-2-27b	22	22	21	43	20	22
mistral-7b	14	16	14	15	16	15
mistral-12b	17	39	16	20	15	19
mistral-3.1-24b	19	21	19	17	18	20
qwen-2.5-7b	20	19	22	18	18	17
qwen-2.5-14b	27	30	31	30	29	23
qwen-2.5-72b	55	63	56	60	56	57

Table A9. Selected layers for the alternative probes. Zero-indexed layers selected independently for the SVM and Mass Mean probes for each instruction-tuned LLM. The selected probe-specific layer is subsequently used for the corresponding individual belief probability and Direct Conditional analyses.

Section D.3. The resulting layers are reported in Table A9 and are subsequently used for both the individual-statement and Direct Conditional analyses.

For the SVM probe, we standardize each activation dimension to zero mean and unit variance using a **StandardScaler** fit only on the training representations. We then fit one **LinearSVC** per class in a one-versus-all construction, with $C = 1.0$, ℓ_2 regularization, squared-hinge loss, a maximum of 10,000 iterations, convergence tolerance 10^{-4} , no class weighting, and random seed 0.

For Mass Mean, we construct one class-versus-rest direction for each class $j \in \{0, \dots, K-1\}$. Let μ_j denote the mean training representation for class j and μ_{-j} the mean training representation over all remaining classes. We compute the difference vector

$$\Delta\mu_j = \mu_j - \mu_{-j}, \quad (45)$$

and normalize $\Delta\mu_j$ to the unit ℓ_2 norm before scoring. The decision boundary is placed halfway between μ_j and μ_{-j} , and statements are scored by their signed projection along the normalized $\Delta\mu_j$ direction relative to this midpoint.

For both probes, the resulting class scores are converted to probability distributions using the multinomial logistic-regression calibration procedure described in Section D.1. The calibrator is fit only on the calibration split with $C = 1.0$, the L-BFGS solver, a maximum of 5,000 iterations, and tolerance 10^{-6} . Direct Conditional probes are trained separately from the corresponding individual-statement probes on the same conditional training and calibration examples used for **sAwMIL**, while retaining the probe-specific layer selected from the sweep.

Across these analyses, the SVM largely reproduces the principal representation-based results obtained with **sAwMIL**, whereas Mass Mean is less consistent across several downstream analyses. Both alternative probes nevertheless preserve the central result that graded stability is not fully explained by the fitted relationship with individual belief probability (Fig. A15). For the SVM, median cross-model residual correlations are $\rho = 0.25, 0.20$, and 0.19 for City Locations, Medical Indications, and Word Definitions, respectively. The corresponding Mass Mean medians are $\rho = 0.17, 0.14$, and 0.22 . Thus, under both alternative probes, substantial proposition-level variation remains after accounting for individual belief probability, and models continue to show positively correlated residual stability.

The probabilistic-coherence results show a clearer difference between the two alternative probes (Fig. A16). The SVM yields distance-to-coherence distributions broadly similar to those obtained with **sAwMIL**. By contrast, Mass Mean generally requires larger adjustments to reach the CCK-coherent set, particularly for City Locations and Word Definitions.

The domain-level characterization shows the same pattern (Fig. A17). Under the SVM, City Locations remains concentrated near high graded stability, while Medical Indications and Word Definitions exhibit broader distributions extending further into intermediate and low stability values. This separation is substantially weaker under Mass Mean, with several City Locations models in particular receiving much lower graded-stability estimates than under either **sAwMIL** or the SVM.

Finally, the behavioral association is more robust under the SVM than under Mass Mean (Fig. A18). SVM-based

SVM

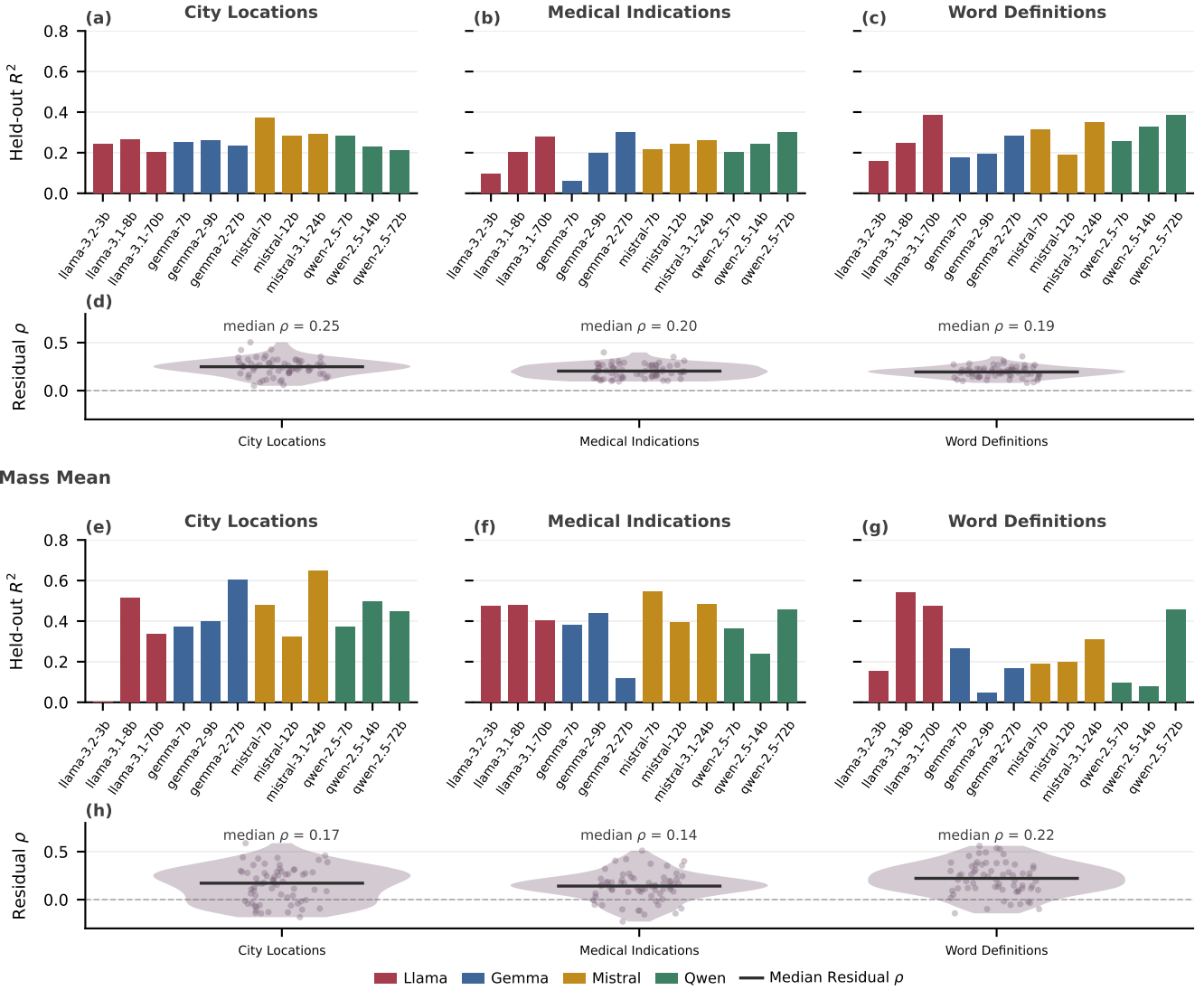


Figure A15. Relationship between individual belief probability and graded stability under alternative probes. For the SVM, we show held-out R^2 for predicting graded stability from individual belief probability in (a) City Locations, (b) Medical Indications, and (c) Word Definitions, respectively, for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green). Panel (d) shows pairwise cross-model Spearman correlations between residual graded-stability scores. Panels (e–h) report the corresponding Mass Mean results. Black lines in (d) and (h) denote median residual correlations. Under both probes, individual belief probability leaves substantial proposition-level variation unexplained and residual stability remains largely positively shared across models.

graded stability remains predominantly associated with greater behavioral resilience among probability-matched beliefs, although the effects are more heterogeneous than in the primary sAwMIL analysis. Under Mass Mean, the estimated effects are generally smaller and less consistent in direction.

The weaker Mass Mean results are consistent with known sensitivities of centroid-based truth directions. Mass Mean was introduced as a binary method that estimates a truth direction from the difference between the centroids of True and False representations [17]. Previous work applying this probe in a three-class setting found that it can become unstable when Neither examples are incorporated, since changes in the composition of the representation sets directly alter the estimated class centroids [9]. Our trivalent implementation similarly constructs three class-versus-rest centroid directions for True, False, and Neither and calibrates the resulting scores into a shared multiclass probability distribution.

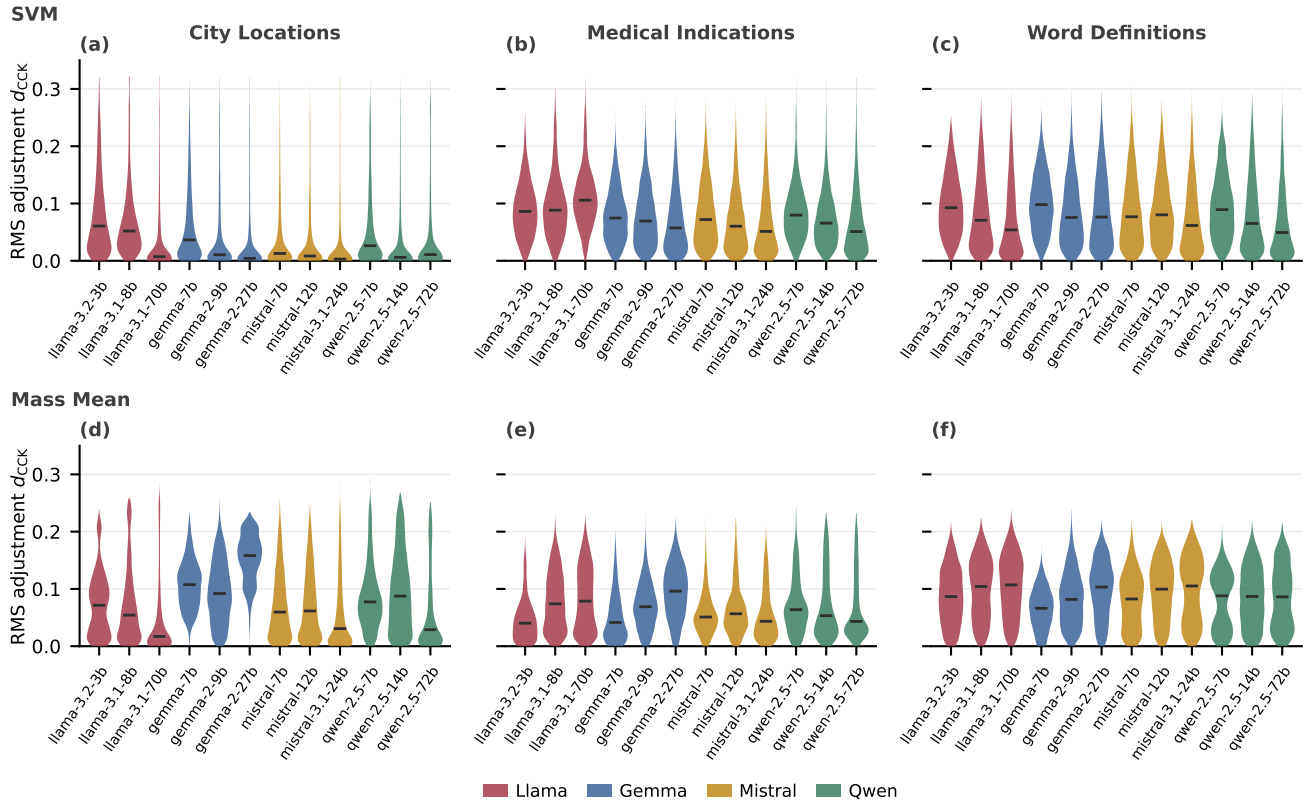


Figure A16. Distance from probabilistic coherence under alternative probes. Violin plots show the RMS adjustment d_{CCK} required to project measured probability distributions for individual statements and conditionals onto the nearest CCK-coherent system. Panels (a–c) show results for the SVM for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green), and panels (d–f) for Mass Mean, with columns corresponding to City Locations, Medical Indications, and Word Definitions. Horizontal black lines denote within-model medians. The SVM broadly reproduces the approximate-coherence pattern obtained with sAwMIL, whereas Mass Mean yields larger and more heterogeneous distances from the coherent set.

This sensitivity is especially important for graded stability because our downstream quantities depend on the absolute allocation of probability mass across all three classes, not merely on whether a binary truth direction ranks **True** statements above **False** statements. **Mass Mean** can therefore remain informative as a veracity probe while being less well suited to the trivalent probability estimation required here. We consequently view the strong SVM replication as evidence that the principal results are not specific to the multi-instance structure of sAwMIL, while the **Mass Mean** results identify a meaningful sensitivity to probes that are not naturally designed for multiclass probability estimation.

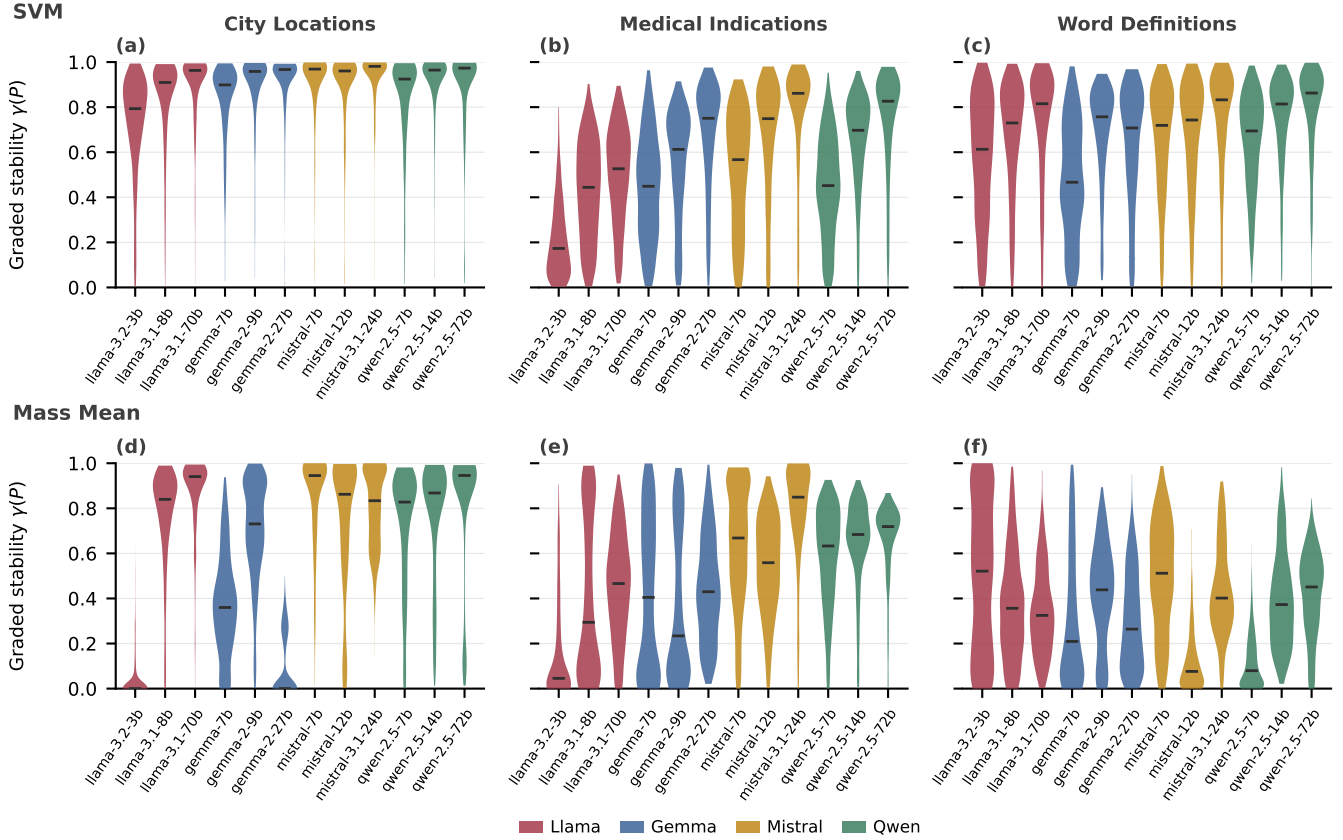


Figure A17. Variation in graded belief stability under alternative probes. Violin plots show proposition-level Direct Conditional graded-stability distributions. Panels (a–c) show results for the SVM for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) and panels (d–f) for Mass Mean, with columns corresponding to City Locations, Medical Indications, and Word Definitions. Horizontal black lines denote within-model medians. The SVM preserves the primary domain-level pattern, whereas Mass Mean produces more heterogeneous distributions and substantially weaker separation across domains.

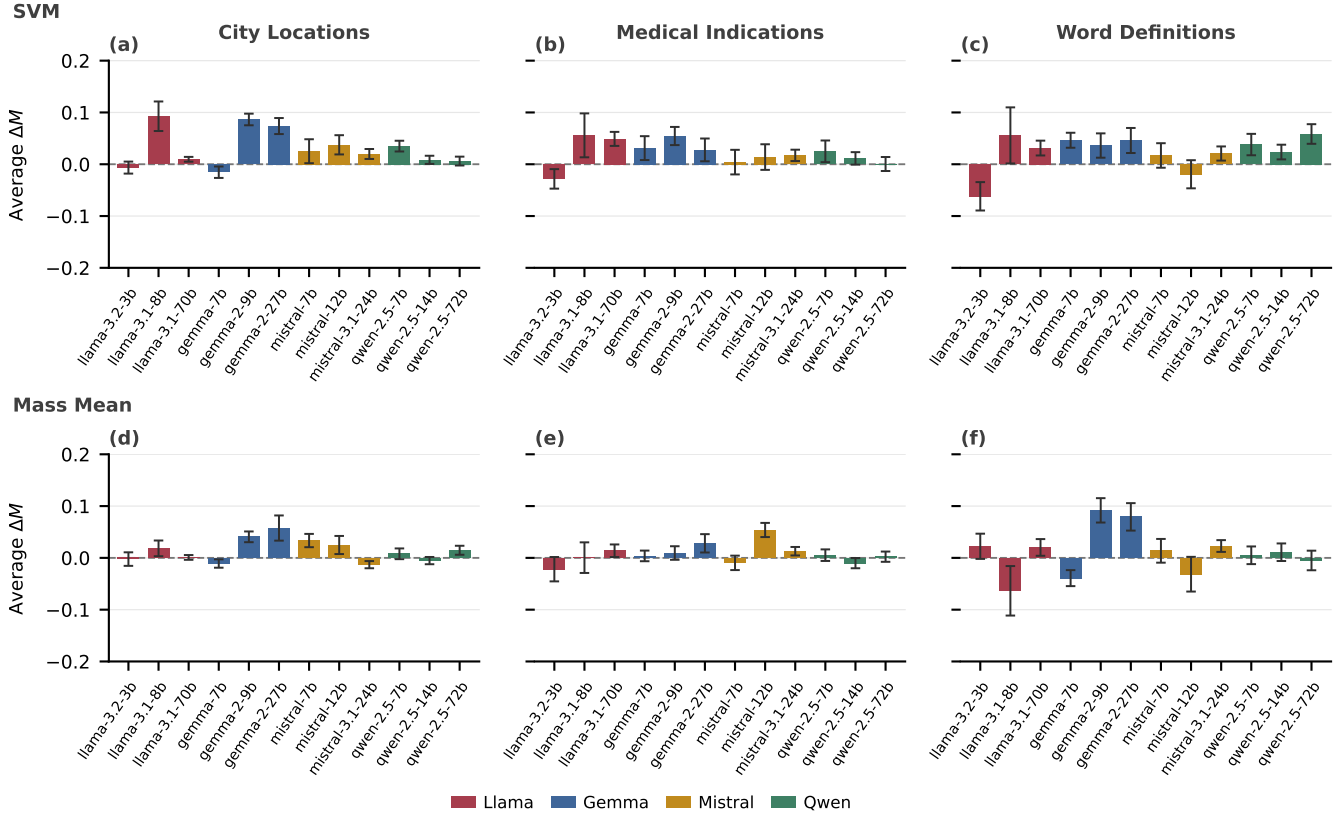


Figure A18. Behavioral resilience under alternative probes. Bars show the mean behavioral movement difference ΔM between probability-matched lower- and higher-stability beliefs. Panels (a–c) show results for the SVM for Llama (red), Gemma (blue), Mistral (yellow), and Qwen (green) and panels (d–f) for Mass Mean, with columns corresponding to City Locations, Medical Indications, and Word Definitions. Error bars denote ± 1 bootstrap standard error, and positive values indicate greater movement for the lower-stability belief. The directional association remains visible across many SVM model–domain combinations but is weaker and less consistent under Mass Mean.