$&!#* vector: Probing sentence embeddings for linguistic properties. In Gurevych, I. & Miyao, Y. (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2126–2136. https://doi.org/10.18653/v1/P18-1198 (Association for Computational Linguistics, Melbourne, Australia, 2018).

18. Hewitt, J. & Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419 (2019).

19. Tenney, I., Das, D. & Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, 4593–4601. https://doi.org/10.48550/arXiv.1905.05950 (2019).

20. Sharma, M. *et al.* Towards understanding sycophancy in language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)*. https://doi.org/10.48550/arXiv.2310.13548 (2024).

21. Wei, J. *et al.* Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846* https://doi.org/10.48550/arXiv.2303.03846 (2023).

22. Bigelow, E. *et al.* Belief dynamics reveal the dual nature of in-context learning and activation steering. *arXiv preprint arXiv:2511.00617* https://doi.org/10.48550/arXiv.2511.00617 (2025).

23. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).

24. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

25. 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*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6 (2020).

26. Wikipedia contributors. List of fictional settlements (2025). https://en.wikipedia.org/wiki/List_of_fictional_settlements.

27. Wikipedia contributors. List of fictional city-states in literature (2025). https://en.wikipedia.org/wiki/List_of_fictional_city-states_in_literature.

28. Fandom NeoEncyclopedia. List of fictional diseases (2025). https://neoencyclopedia.fandom.com/wiki/List_of_fictional_diseases.

29. Fandom NeoEncyclopedia. List of fictional toxins (2025). https://neoencyclopedia.fandom.com/wiki/List_of_fictional_toxins.

30. Chemeurope Encyclopedia. List of fictional medicines and drugs (2025). https://www.chemeurope.com/en/encyclopedia/List_of_fictional_medicines_and_drugs.html.

31. Tomasula, S. The Thackery T. Lambshead pocket guide to eccentric & discredited diseases. *The Rev. Contemp. Fiction* **24** (2004).

32. Almaden, S. A. Dahl dictionary: A list of 103 words made-up by Roald Dahl (2023). https://beelinguapp.com/blog/Dahl%20Dictionary:%20A%20List%20of%20103%20Words%20Made-up%20By%20Roald%20Dahl.

33. Schleitwiler, P. & Shuflin, G. Dothraki initial text (2025). https://conlang.org/language-creation-conference/lcc5/1-dothraki-initial-text/.

34. Dict-Na'vi.com Online Dictionary. wordlist "substantive (noun)" (2025). https://dict-navi.com/en/dictionary/list/?type=classification&ID=1.

A Notation

We summarize the mathematical notation used throughout the manuscript in Table A1.

Symbol	Description
$\mathcal{M}$	A fixed large language model (LLM).
l	Layer index used for activation extraction.
$\mathcal{S} = \{s_i\}_{i=1}^{N}$	Set of N natural-language statements.
y_i	Ground-truth veracity label of s_i, $y_i \in \{\texttt{True}, \texttt{False}, \texttt{Neither}\}$.
$\mathcal{N}$	Set of all `Neither` statements: $\mathcal{N} = \{s_i \mid y_i = \texttt{Neither}\}$.
$\mathcal{N}_{\text{fam}}$, $\mathcal{N}_{\text{unf}}$	Partition of $\mathcal{N}$ into epistemically familiar (`Fictional`) and unfamiliar (`Synthetic`) subsets.
$\mathcal{N}_{\text{fam}}^{(T)}$, $\mathcal{N}_{\text{fam}}^{(F)}$	Canonically true and false subsets of familiar fictional statements (used in `Fictional(T)`).
$\phi_{\mathcal{M},l}$	Representation map from a statement to its layer-l hidden representations under $\mathcal{M}$.
$\mathbf{z}_i^{(l)}$	Layer-l representation of statement s_i, $\mathbf{z}_i^{(l)} = \phi_{\mathcal{M},l}(s_i)$.
$\mathcal{D}$	Dataset of statements, representations, and labels: $\mathcal{D} = \{(s_i, \mathbf{z}_i^{(l)}, y_i)\}_{i=1}^{N}$.
$\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$	Training and test splits of $\mathcal{D}$.
Θ	Semantic interpretation specifying which `Neither` statements are treated as compatible with truth.
$\mathcal{N}_\Theta \subseteq \mathcal{N}$	Subset of `Neither` statements treated as compatible with truth under Θ.
$g(\cdot, \Theta)$	Evaluation function mapping a statement and interpretation to a binary judgment: $g(s, \Theta) \in \{\texttt{True}, \texttt{Not True}\}$.
g_{p}	Representational instantiation of g via a probe over $\phi_{\mathcal{M},l}(s)$.
g_{zs}	Behavioral instantiation of g via zero-shot prompting with belief context C_Θ.
C_Θ	The belief context inserted into a zero-shot prompt to instantiate perturbation Θ.
Θ_0	Baseline semantic interpretation (no `Neither` treated as true).
$\mathcal{B}_{\text{true}}^{\Theta_0}$	Baseline belief set: $\{s_i \mid y_i = \texttt{True},\ g(s_i, \Theta_0) = \texttt{True}\}$.
$\mathcal{B}_{\text{true}}^{\Theta}$	Belief set under perturbed interpretation Θ.
$\mathcal{R}$	Epistemic retractions: $\mathcal{R} = \mathcal{B}_{\text{true}}^{\Theta_0} \setminus \mathcal{B}_{\text{true}}^{\Theta}$.
$\mathcal{E}$	Epistemic expansions: $\mathcal{E} = \mathcal{B}_{\text{true}}^{\Theta} \setminus \mathcal{B}_{\text{true}}^{\Theta_0}$.

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

B Data

B.1 Data Generation

We use statements from the City Locations, Medical Indications, and Word Definitions datasets introduced in [9]. City statements take the form *"The city of [city] is (not) located in [country],"* (omitting *"The city of"* when redundant). Medical statements follow *"[drug] is (not) indicated for the treatment of [disease/condition]."* Word Definition statements draw from three templates: *"[word] is (not) a [instanceOf],"* *"[word] is (not) a type of [typeOf],"* and *"[word] is (not) a synonym of [synonym]."* No personal data, identifying information, or user-generated content is included.

B.1.1 True, False, and Synthetic Statements

We take the `True`, `False`, and `Synthetic` statements from the datasets introduced in [9]. All statements are constructed with both affirmative and negated forms. `Synthetic` entities are generated using a Markov-chain–based name generator (`namemaker`[5]) and undergo multi-stage filtering, including database checks, model tagging, and web-search validation, to ensure no accidental overlap with real entities. Validated names are then paired to form grammatically coherent but semantically meaningless statements that follow each template. Because `Synthetic` entities do not exist and cannot have appeared in training corpora, LLMs have no basis for assigning them a truth value. Accordingly, these statements function as `Neither` cases: unknown claims for which belief should be suspended rather than confidently classified as true or false.

B.1.2 Fictional Statements

In addition to `Synthetic` statements, which represent *unseen and unknown* claims, we construct new sets of `Fictional` statements for all three domains. `Fictional` statements also function as `Neither` statements in our experiments as they reference entities that do not exist in the real world and therefore lack real-world truth value. However, unlike `Synthetic` statements, many `Fictional` entities are likely to have appeared in LLM training corpora.[6] As such, they represent a complementary form of `Neither`: claims that an LLM may recognize, but that

[5] https://github.com/Rickmsd/namemaker.

[6] For later analyses, we additionally annotate fictional statements with their within-universe factual status (`Fictional (T)` or `Fictional (F)`), but this labeling is not used in the primary `True` vs. `Not True` classification tasks.

still lie outside the true–false axis relevant to factual grounding.

To ensure that `Fictional` statements remain genuinely non-factual, all terms were validated to exclude any real-world overlap, and fictional lexical items appearing in any natural language were excluded to prevent misinterpretation by multilingual LLMs. `Fictional` statements were then constructed using the same templates as the `True`, `False`, and `Synthetic` statements, including both affirmative and negated forms.

Fictional City Locations. Fictional cities and countries span literature, film, radio, television, comics, animation, and games [26, 27]. Each ⟨city, location⟩ pair is included only when an identifiable enclosing region exists. When multiple spatial resolutions are available, we select the most specific (e.g., ⟨Quahog, Rhode Island⟩ rather than ⟨Quahog, United States⟩).

Fictional Medical Indications. Fictional drug and disease statements are drawn from (1) *NeoEncyclopedia Wiki* [28, 29]; (2) ChemEurope's *List of Fictional Medicines and Drugs* [30]; and (3) *The Thackery T. Lambshead Pocket Guide to Eccentric & Discredited Diseases* [31]. Drug–disease pairs are included when a treatment relationship exists according to the fictional source.

Fictional Word Definitions. Fictional lexical items are compiled from (1) *Gobblefunk* [32]; (2) *Dothraki* [33]; and (3) *Na'vi* [34]. Dothraki and Na'vi have formal linguistic structure, whereas Gobblefunk is a playful neologistic extension of English.

B.1.3 Noise

The `Noise` statements contain no linguistic content. We generate $n_{\text{noise}} = 0.10 \cdot |\mathcal{D}|$ random activation sequences by sampling from a multivariate Gaussian with per-feature mean, standard deviation, and sequence-length distribution matched to the LLM activations. These distributionally consistent but non-semantic sequences serve as a control, allowing us to test whether observed representational differences arise from semantic content or from statistical variation in activation space.

B.2 Data Splits for Probing Experiments

Dataset	Train	Calibration	Test	Total
City Locations	4746 (0.54)	1772 (0.20)	2229 (0.25)	8747 (1.00)
Medical Indications	4636 (0.55)	1721 (0.20)	2121 (0.25)	8478 (1.00)
Word Definitions	6488 (0.54)	2514 (0.21)	3041 (0.25)	12043 (1.00)

Table A2. Dataset splits. The number of statements used in training, calibration, and testing of the probe. The proportion of total statements is reported in parentheses.

Table A2 summarizes the partitions used for all experiments. Each dataset is split exclusively into training, calibration, and test sets to prevent data leakage. Approximately 55% of statements are used for training, 20% for calibration, and 25% for testing. We use identical splits in all conditions.

C LLMs

Table A3 lists the sixteen open-source LLMs used in our experiments. The set spans four major model families, Gemma, Llama, Mistral, and Qwen, with approximately 3 billion to 15 billion parameters and release dates between February and September 2024. For each family, we include both base (pre-trained) and chat-tuned variants. Together, these LLMs provide a representative cross-section of current decoder-only architectures varying in scale, origin, and training objectives.

D LLM-level Results

The main text reports LLM-agnostic stability patterns aggregated across sixteen LLMs. Here we provide the corresponding LLM-level results.

D.1 Linguistic and Representational Structure of `Neither` Statements (by LLM)

Figures A1-A16 show the pairwise activation distance matrices for all sixteen LLMs. Three general representational patterns emerge. The first, observed in `_gemma-2-9b` (Fig. A1) and `gemma-2-9b` (Fig. A10), shows `Fictional` and `Synthetic` statements clustering near `True` and `False` statements, with `Noise` forming a distinct outlier. The