Continuous Interpretive Steering for Scalar Diversity

Ye-eun Cho
Sungkyunkwan University
Seoul, South Korea
joyenn@skku.edu

Abstract

Pragmatic inference is inherently graded. Different lexical items give rise to pragmatic enrichment to different degrees. Scalar implicature exemplifies this property through scalar diversity, where implicature strength varies across scalar items. However, evaluations of pragmatic inference in large language models (LLMs) often rely on prompt-based manipulations. Beyond prompt-level effects, this study introduces Continuous Interpretive Steering (CIS), a method that probes graded pragmatic interpretation by treating activation-level steering strength as a continuous experimental variable. To support this analysis, this study introduces a new dataset, GraSD, which encodes graded scalar diversity. Experiments on four LLMs show that uniform activation steering increases pragmatic interpretations globally but collapses item-level variation, whereas graded activation steering yields differentiated interpretive shifts aligned with scalar diversity grades. It indicates that graded sensitivity is encoded in the representation space and can be systematically recovered through controlled intervention. Together, CIS and GraSD provide a principled framework for evaluating graded pragmatic sensitivity in LLMs.

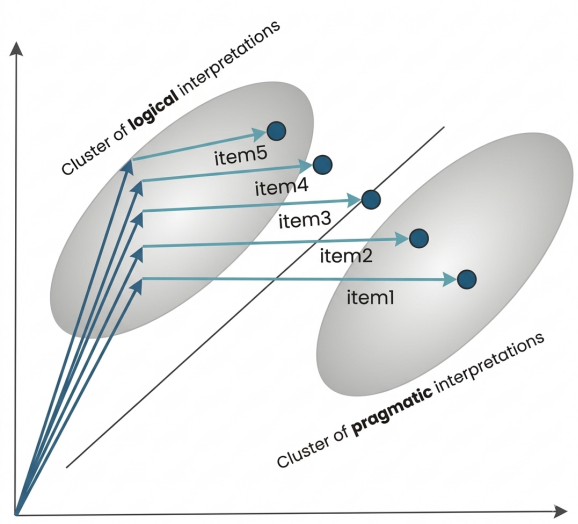

Figure 1: Continuous interpretive steering in representation space

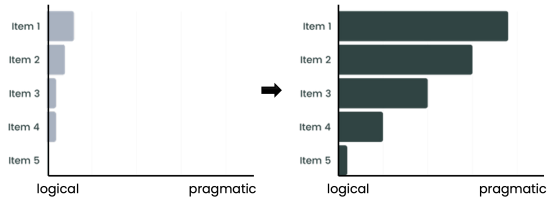

Figure 2: Graded changes in interpretive outcomes before and after steering

1 Introduction

Pragmatic inference enables comprehenders to derive meanings that go beyond literal truth conditions. A canonical case is scalar implicature, where an utterance containing a weaker scalar term (e.g., *some*) is often interpreted as negating stronger alternatives (e.g., *not all*) when those alternatives are relevant (Horn, 1972). Scalar implicature has therefore been widely used to evaluate whether large language models (LLMs) exhibit human-like pragmatic behavior, and recent work has reported both successes and failures depending on experimental framing and task format (Hu et al., 2023; Cho and Kim, 2024; Wu et al., 2024; Cho and Maeng, 2025; Tsvilodub et al., 2025).

However, most evaluations of pragmatic inference in LLMs rely primarily on prompt manipulations, varying instructions, exemplars, or question phrasing to elicit a pragmatic interpretation (Webson and Pavlick, 2022; Mielke et al., 2022; Turpin et al., 2023; Park et al., 2024; Cho, 2025; Cho and Maeng, 2025). While practical, prompt-based approaches are inherently ephemeral and can be brittle; measured pragmatic reasoning abilities may

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12661–12678

reflect sensitivity to surface form rather than stable internal representations. This concern is echoed by work arguing that pragmatic benchmarks can be highly prompt-sensitive, complicating mechanistic interpretation and cross-model comparison (Ruis et al., 2023).

Therefore, instead of inducing pragmatic behavior only through prompts, this work intervenes directly in the model's internal activations at inference time. Inference-time steering methods modify hidden states during a forward pass without updating model weights, offering a lightweight alternative to fine-tuning and a more direct probe of model-internal computations (Turner et al., 2023; Zou et al., 2023).

Importantly, pragmatic phenomena in human language are inherently graded. Specifically, in terms of scalar implicature, different scalar items trigger pragmatic enrichment at different degrees, which is known as scalar diversity (van Tiel et al., 2016). Consequently, if LLMs are to approximate human-like pragmatic behavior, their pragmatic interpretations should likewise be modulated in an item-sensitive manner.

For this purpose, this study introduces Continuous Interpretive Steering (CIS) for scalar diversity. This approach constructs a steering direction in activation space and varies its strength continuously. Unlike prior steering work, which primarily relies on fixed intervention vectors to optimize downstream behaviors, the present approach treats steering strength as an experimental variable for analysis. Figure 1 illustrates the core idea of this approach in representation space and Figure 2 depicts the expected differences in interpretive outcomes before and after steering. To evaluate graded interpretive sensitivity across scalar items, a new dataset, GraSD (Graded Scalar Diversity), is constructed and used in this study.

2 Background

2.1 Scalar Diversity

Scalar implicature indicates that utterances containing weaker scalar terms (e.g., *some*) are often interpreted as negating stronger alternatives (e.g., *not all*) when the stronger alternatives are relevant (Horn, 1972). However, growing empirical evidence suggests that different scalar items give rise to pragmatic interpretations to different degrees (van Tiel et al., 2016; Ronai and Xiang, 2021, 2024; Pankratz and Van Tiel, 2021; see also Degen and Tanenhaus, 2015). This phenomenon is known as scalar diversity.

For example, van Tiel et al. (2016) showed that scalar items differ in how strongly they license implicatures: utterances with *some* robustly invite *not all*, whereas terms such as *warm* rarely give rise to *not hot*. Crucially, this variability persists even when contextual factors are carefully controlled. This graded nature of implicature strength across scalar items adopted from van Tiel et al. (2016) is represented in Appendix A. Subsequent psycholinguistic studies have reinforced this tendency by showing that scalar implicature judgments often exhibit graded response patterns across various scalar items (Ronai and Xiang, 2021, 2024; Pankratz and Van Tiel, 2021). These findings suggest that scalar diversity is a robust empirical property of human pragmatic behavior. Scalar implicature, in other words, is not simply present or absent, but instantiated to different degrees depending on the lexical items under consideration. These properties have direct implications for modeling pragmatic inference.

Recent work attempting to elicit human-like pragmatic behavior often aim to shift models away from literal interpretations toward pragmatic ones (Hu et al., 2023; Cho and Kim, 2024; Wu et al., 2024; Cho and Maeng, 2025; Tsvilodub et al., 2025). However, such approaches risk collapsing scalar diversity by treating pragmatic enrichment as a uniform target. If a model is pushed indiscriminately toward pragmatic interpretation, it may overgenerate enriched readings in cases where humans exhibit only weak or marginal implicatures.

From the perspective of scalar diversity, human-like pragmatic reasoning requires the capacity to modulate the strength of pragmatic enrichment at the level of individual items, reflecting the graded patterns observed in human judgments. This motivates an evaluation framework that goes beyond binary success or failure and instead examines whether models can reproduce item-level variation in pragmatic interpretation.

2.2 Internal Activation Interventions

While LLMs effectively elicit certain behaviors with input-level manipulations, such as prompt engineering or task rephrasing, these approaches operate indirectly on the model's internal computations and are often sensitive to surface form. An alternative approach intervenes directly on a model's internal activations at inference time

(Turner et al., 2023; Zou et al., 2023; Wang et al., 2024). Inference-time steering methods modify hidden states during the forward pass of a pre-trained model, without updating its parameters. This paradigm has been explored under various labels, including activation engineering and representation engineering, and has been shown to enable lightweight control over model outputs with minimal computational cost (Turner et al., 2023; Zou et al., 2023).

Most existing steering approaches construct a fixed intervention vector, often derived from contrastive examples, and apply it uniformly to bias model behavior toward a desired outcome (Turner et al., 2023; Wang et al., 2024; Bayat et al., 2025; Højer et al., 2025; Neplenbroek et al., 2025; Soo et al., 2025; Suri et al., 2025). These methods have primarily been evaluated on downstream objectives such as sentiment control, truthfulness, or alignment, where the goal is to optimize performance rather than to analyze internal structure. Consequently, steering is typically treated as a one-shot control mechanism rather than as an experimental variable.

Recent work has begun to incorporate gradient-based signals, such as attribution scores, to construct more targeted steering directions (Nguyen et al., 2025). However, even in these settings, steering strength is rarely varied systematically, and evaluation focuses on end-task improvements rather than on graded shifts in interpretation.

In contrast, the present study treats steering as a probing tool. By varying steering strength continuously, internal intervention is used to trace how interpretations change across scalar items. This perspective aligns naturally with scalar diversity. If pragmatic inference is internally represented as a graded state, then small changes in internal activations should produce smooth, item-sensitive shifts. Steering thus provides a means of probing the geometry of pragmatic inference within LLMs, complementing traditional prompt-based evaluations.

3 Dataset

To evaluate graded pragmatic behavior in LLMs, this study constructs a dataset, called GraSD. GraSD consolidates <weak, strong> item pairs from four prior studies that investigate scalar diversity under controlled experimental settings as follows:

Type	Sentence
Anchor	Some pets prefer to sleep in the sun.
Logical	All pets prefer to sleep in the sun.
Pragmatic	Not all pets prefer to sleep in the sun.

Table 1: Example of experimental sentences for the pair <some, all>.

- 43 pairs (van Tiel et al., 2016)
- 43 pairs (Ronai and Xiang, 2021)
- 50 pairs (Pankratz and Van Tiel, 2021)
- 60 pairs (Ronai and Xiang, 2024)

After removing duplicate entries across sources, a total of 121 <weak, strong> pairs were retained. These pairs serve as the basis for constructing experimental sentences consisting of anchor, logical, and pragmatic variants as in Table 1. For each scalar pair, the anchor sentence contains the weak scalar term and represents the baseline utterance from which interpretations are derived. The logical variant replaces the weak term with its stronger alternative, thereby expressing a semantically stronger proposition within the same scale. The pragmatic variant encodes the scalar implicature associated with the anchor, typically realized as the negation of the stronger alternative.[1]

To enable large-scale evaluation, the 121 base item pairs were augmented into a range of contextualized sentence instances. A theory-driven, constraint-based augmentation strategy was employed, providing the GPT-4o model (OpenAI, 2024) with linguistically motivated constraints derived from scalar implicature theory, which is inspired by Su et al. (2025). Based on these theoretical constraints, the model generated diverse contextual instances while consistently preserving the intended forms among the anchor, logical, and pragmatic sentences. Compared to rule-based methods, this approach produces richer and more varied contexts; and compared to unconstrained generation, it offers greater control over the pragmatic conditions under which scalar implicatures may arise. Through this approach, by producing 100 instances per scalar item pair, a total of 121,000 sentence instances were created. The scalar item pairs included in GraSD and the prompt used for the data

[1] In formal semantics, weak terms (e.g., *some*) are compatible with stronger cases (e.g., *all*), as they express existential quantification (Horn, 1972). The logical variant therefore represents a stronger proposition within the scale, while the pragmatic variant corresponds to the negation of this stronger alternative.

augmentation are listed in Appendix C. The complete dataset, experimental code and results are publicly available.[2]

4 Methods

4.1 Models

Experiments are conducted on four open-weight transformer-based language models: LLaMA3 (Grattafiori et al., 2024), Qwen2 (Team et al., 2024b), Gemma2 (Team et al., 2024a), and OLMo (Groeneveld et al., 2024). These models were selected to cover a range of architectures and training regimes while maintaining full access to internal representations required for activation-level steering.

4.2 Activation-level Steering

This study implements activation-level steering as an intervention defined in the model's activation space at inference time, without updating model parameters. Rather than modifying hidden states during the forward pass, the approach operates on extracted internal representations, enabling controlled manipulation of activation geometry while preserving the model's original computation.

Let $h_l(x) \in \mathbb{R}^d$ denote the hidden activation produced by layer l for an input sentence x. For each input, hidden representations are extracted from a fixed set of k transformer layers and concatenated to form a multi-layer representation vector $h(x) \in \mathbb{R}^{kd}$. Steering is defined by an intervention vector $v \in \mathbb{R}^{kd}$, which represents a direction in this aggregated activation space. Given a scalar coefficient $\alpha \in \mathbb{R}$, the steered representation is defined as:

$$h'(x) = h(x) + \alpha v \tag{1}$$

Here, v specifies the direction of representational change, while α controls the magnitude of the intervention. The intervention vector is computed once and shared across all inputs, ensuring that variation in model behavior arises from differences in steering strength rather than from item-specific directions.

This approach is closely related to prior work on activation and representation engineering (Turner et al., 2023; Wang et al., 2024; Bayat et al., 2025; Højer et al., 2025; Neplenbroek et al., 2025; Soo et al., 2025; Suri et al., 2025), but differs in its use of steering as an analytical probe rather than as a control mechanism for downstream task optimization.

4.3 Continuous Interpretive Steering

Building on the activation-level steering framework described above, Continuous Interpretive Steering (CIS) is introduced as a method that treats steering strength as a continuous experimental variable for probing graded pragmatic interpretation. Rather than using steering to enforce a fixed behavioral outcome, controlled variation in steering magnitude is exploited to examine how internal representations shift across interpretive alternatives.

Each dataset item consists of an anchor sentence and its corresponding logical and pragmatic variants. For a given item, multi-layer representations are extracted for all three sentence types under fixed model parameters. Using these representations, a pragmatic direction in activation space is defined as the difference between the pragmatic and logical representations.

The anchor representation is then steered along this pragmatic direction by varying the steering coefficient α over a continuous range. This yields a family of steered anchor representations corresponding to different degrees of internal displacement toward the pragmatic interpretation. Crucially, the intervention direction is held constant across items, and only the magnitude of steering is varied, allowing item-level sensitivity to be assessed independently of direction-specific effects.

In practice, the pragmatic direction is estimated from a limited subset of instances per item (three instances per item pair), indicating that a stable interpretive direction can be extracted from a relatively small sample of data. This direction is subsequently held constant across all items during analysis.

5 Experiments

5.1 Uniform Activation Steering

Uniform activation steering applies a single steering direction with a fixed magnitude to all items, regardless of their lexical properties or interpretive characteristics. This condition serves as a control condition that probes the effect of applying activation steering without differentiation across items.

Concretely, a steering vector is added to the model's internal representations at a fixed layer, scaled by a constant coefficient α. The same steering coefficient is used for every item in the dataset,

[2] https://github.com/joyennn/CIS

ensuring that all representations are shifted by an equal amount along the same direction in representation space.

The purpose of this setup is to examine the effects of a uniform representational shift on interpretive outcomes, while preserving a simple and controlled intervention scheme. By holding steering strength constant, this condition provides a reference point for assessing whether activation steering alone can induce systematic changes in interpretation.

5.2 Graded Activation Steering

Graded activation steering extends the uniform steering setup by allowing steering strength to vary across items, while keeping the steering direction identical to that used in uniform activation steering. This condition introduces controlled variation in steering magnitude as an experimental variable.

Specifically, items are assigned to discrete grade conditions (A–E), each of which is associated with a different steering coefficient α. For example, grade A corresponds to the strongest steering magnitude and grade E to the weakest. Importantly, grades are used solely as a mechanism for modulating steering strength and do not encode semantic or interpretive categories in themselves. Items in different grades therefore experience different magnitudes of representational shift along the same steering direction.

5.3 Analysis Strategies

Interpretive preference is operationalized in representational terms. For each steered anchor representation, cosine similarity to the pragmatic and logical representations of the same item is computed. An item is considered to favor the pragmatic interpretation when the steered anchor becomes more similar to the pragmatic variant than to the logical variant. By tracking how this preference changes as a function of α, a graded profile of interpretive sensitivity is obtained for each item.

Differences between steering conditions are evaluated using two complementary statistical tests. First, Wilcoxon signed-rank tests (Wilcoxon, 1945) are used to assess whether activation steering induces a significant shift in interpretive preference relative to the baseline condition. Second, Spearman rank correlations (Spearman, 1987) are employed to examine whether item-level interpretive preferences observed in the baseline condition are preserved under steering, thereby assessing the degree to which steering maintains graded sensitivity across items.

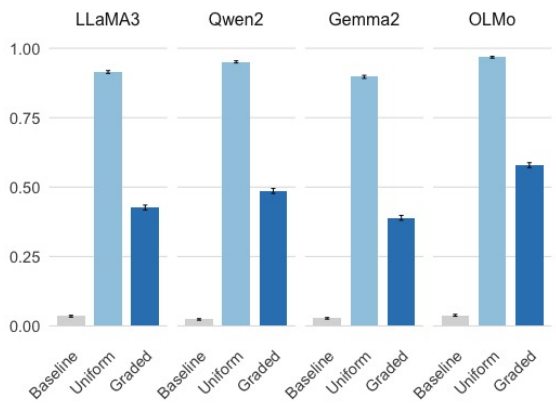

Figure 3: Proportion of pragmatic interpretations for baseline, uniform activation steering, and graded activation steering conditions across LLMs

6 Results

6.1 Overview of Interpretive Patterns

Across all four models, clear differences emerge in interpretive behavior across steering conditions as shown in Figure 3, which summarizes the proportion of instances favoring the pragmatic interpretation across baseline, uniform and graded activation steering without distinguishing individual items.

Specifically, across the four models, pragmatic interpretations are rare in the baseline condition, indicating that, in the absence of steering, model preferences are not biased toward pragmatic reasoning. Introducing activation steering leads to a substantial shift in overall interpretive behavior. Under uniform activation steering, the proportion of pragmatic interpretations increases markedly relative to baseline. Under graded activation steering, pragmatic interpretations remain substantially more frequent than in the baseline condition, although the overall proportion is reduced relative to uniform steering.

The increase from baseline to uniform conditions suggests that activation steering exerts a robust global effect on interpretive preferences across models. The lower proportion under graded activation steering compared to uniform activation steering implies that modulation of steering strength redistributes interpretive sensitivity. These patterns are more strongly captured in item-level proportions of pragmatic interpretations under each steering condition, which are reported in Appendix D.

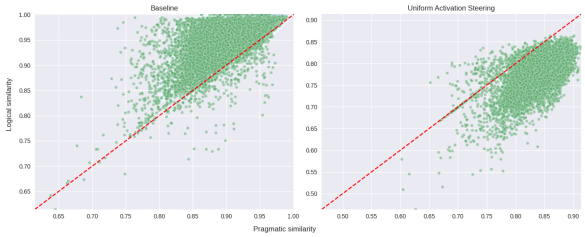

Figure 4: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (left) and uniform activation steering (right) for OLMo

	Uniform		Graded	
Model	W	p-value	W	p-value
LLaMA3	5.0	$< .001$	29.0	$< .001$
Qwen2	1.0	$< .001$	27.0	$< .001$
Gemma2	6.0	$< .001$	28.5	$< .001$
OLMo	2.0	$< .001$	32.0	$< .001$

Table 2: Wilcoxon signed-rank test results comparing uniform and graded activation steering across LLMs

	Uniform		Graded	
Model	ρ	p-value	ρ	p-value
LLaMA3	0.02	0.78	0.38	$< .001$
Qwen2	-0.01	0.91	0.27	$< .01$
Gemma2	0.05	0.60	0.39	$< .001$
OLMo	-0.02	0.81	0.38	$< .001$

Table 3: Spearman rank correlations comparing uniform and graded activation steering across LLMs

6.2 Effects of Uniform Activation Steering

Uniform activation steering, implemented with a fixed steering magnitude, induces systematic changes in interpretive preference across models. Results from the Wilcoxon signed-rank tests in Table 2 show that uniform activation steering leads to a robust and statistically significant shift in interpretive behavior for all four models (ps < .001). This indicates that uniform activation steering consistently increases the overall proportion of pragmatic interpretations relative to the unsteered baseline.

However, this global shift is not accompanied by sensitivity to item-level variation. As shown by the Spearman rank correlations in Table 3, correlations between baseline interpretive preferences and those observed under uniform steering are weak and non-significant across all models (ps = .78, .91, .60, .81). This suggests that uniform activation steering does not preserve the relative ordering of items, but increases pragmatic responses in a relatively homogeneous manner across items.

Figure 4 illustrates this pattern for a representative model, showing a uniform shift in pragmatic similarity under uniform activation steering relative to the baseline. Corresponding plots for the remaining models are provided in Appendix B.

These results characterize uniform activation steering as a coarse-grained intervention. While it reliably induces a global shift toward pragmatic interpretations, it provides limited insight into graded interpretive sensitivity or differential item effects. This limitation motivates the use of graded activation steering.

6.3 Effects of Graded Activation Steering

Unlike uniform activation steering, graded activation steering modulates steering strength continuously across items. Results from the Wilcoxon signed-rank tests in Table 2 show that graded activation steering also produces a statistically significant shift in interpretive behavior relative to the baseline condition across all four models (ps < .001). Compared to uniform steering, graded steering produces larger Wilcoxon statistics (W), reflecting more differentiated item-level responses rather than uniformly strong shifts across items.

This indication is reinforced by Spearman rank correlations in Table 3. Correlations between baseline interpretive preferences and those observed under graded steering are moderate and statistically significant for all models (ps < .01, .001). This suggests that graded activation steering preserves item-level sensitivity in interpretive preferences.

Figure 5 illustrates this pattern for a representative model, showing that graded steering yields differentiated responses across graded items, reflecting graded interpretive sensitivity. Corresponding plots for the remaining models are provided in Appendix E.

These results characterize graded activation steering as a fine-grained intervention. While it induces a reliable global shift toward pragmatic interpretations, it does so in a manner that preserves item-level structure and relative interpretive differences. This property makes graded activation steering particularly well-suited for probing pragmatic interpretive behavior.

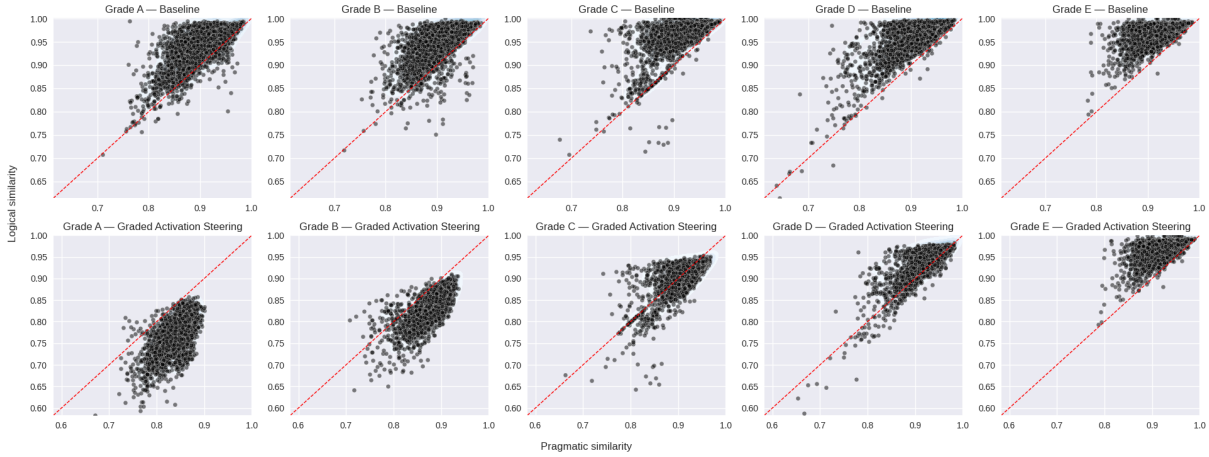

Figure 5: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (left) and uniform activation steering (right) for OLMo

6.4 Direct Comparison: Uniform vs Graded Steering

A direct contrast between uniform and graded activation steering reveals qualitatively distinct interpretive patterns. Although both steering regimes produce statistically significant deviations from the baseline condition, their effects differ fundamentally in terms of distributional structure and item-level organization.

At the aggregate level, both uniform and graded steering yield robust Wilcoxon signed-rank test results, indicating reliable shifts toward pragmatic interpretations relative to baseline. However, as shown in Sections 6.2 and 6.3, statistical significance alone does not distinguish between homogeneous perturbation and structurally organized change. Additional descriptive statistics for grade-wise deviations under uniform and graded steering are reported in Appendix F.

A direct comparison therefore requires examining how steering effects are distributed across individual items. Figure 6 visualizes the distribution of item-level changes in pragmatic interpretation rate (Δ relative to baseline) under the two steering regimes. Under uniform activation steering, the distribution is sharply skewed toward large positive values, with most items exhibiting similarly strong increases in pragmatic interpretation. This pattern indicates a broad and relatively homogeneous displacement of model behavior, consistent with a coarse-grained intervention that affects items uniformly regardless of their baseline interpretive tendencies. Corresponding histograms for the remaining models are provided in Appendix G.

In contrast, graded activation steering produces a substantially broader distribution of item-level changes. While many items still show increased pragmatic interpretation, the magnitude of change varies widely, ranging from negligible shifts to large positive deviations. This heterogeneous distribution reflects differentiated item-level responses.

Importantly, this variability is not random. As indicated by the significant Spearman rank correlations reported in Table 3, the magnitude of item-level responses under graded steering increases systematically along the pre-defined A–E grade hierarchy, with higher-grade items exhibiting larger deviations from baseline than lower-grade items.

In addition, these grade-wise results confirm that graded activation steering induces interpretive change in a manner that respects and exploits the underlying A–E grade structure. By aligning steering effects with a theoretically motivated hierarchy of pragmatic strength, graded steering enables the emergence of fine-grained interpretive organization that cannot be achieved through fixed-magnitude interventions.

7 Conclusion

This study proposed Continuous Interpretive Steering (CIS) as a method for probing graded pragmatic interpretation in LLMs, with a particular focus on scalar diversity. Specifically, CIS conceptualizes pragmatic interpretation as a graded internal phenomenon that can be examined by systematically varying steering strength in activation space.

Across four LLMs, a clear contrast was observed between uniform and graded activation steering. Uniform steering reliably increased the overall rate of pragmatic interpretations, but did so in a

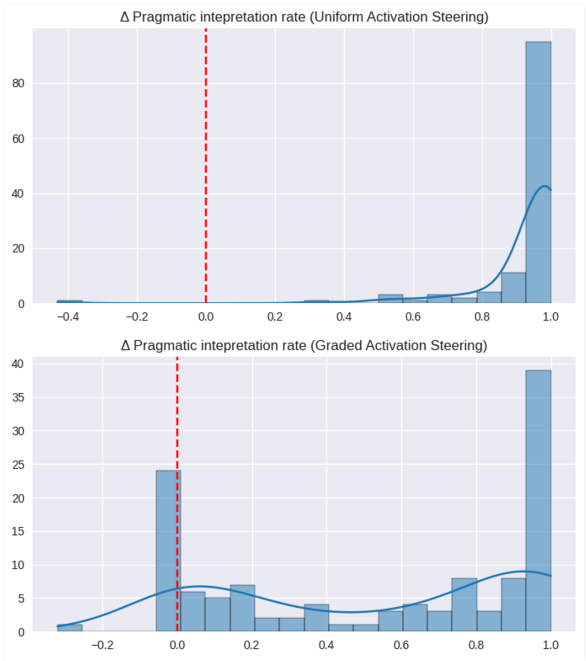

Figure 6: Histograms showing item-level changes in pragmatic interpretation rate (Δ relative to baseline) under uniform (top) and graded (bottom) activation steering for OLMo.

largely homogeneous manner that obscured item-level differences. In contrast, graded activation steering yielded differentiated interpretive shifts that aligned with independently established scalar diversity grades and preserved the sensitivity to item-specific variation.

These findings highlight a limitation of aggregate evaluations of pragmatic behavior. Global increases in pragmatic interpretations alone do not capture structured pragmatic competence. Instead, preserving graded, item-level variation provides a more informative signal of how pragmatic distinctions are organized internally. By treating steering strength as a continuous experimental variable, CIS enables a fine-grained analysis of how internal representations shift between logical and pragmatic interpretations. It also indicates that the representational space encodes graded sensitivity that can be systematically recovered through controlled intervention.

In support of this analysis, the study introduced GraSD, a new dataset that encodes graded pragmatic strength across diverse scalar item pairs. By pairing GraSD with graded activation steering, this work provides both a methodological framework and a publicly available resource for evaluating graded pragmatic sensitivity in LLMs.

While the present study focused on scalar implicature, the proposed approach is not limited to this domain. Many pragmatic phenomena exhibit graded interpretive profiles across lexical items and contexts. CIS, therefore, offers a general tool for investigating the internal organization of pragmatic inference in LLMs.

Limitations

This study has several limitations. First, the analysis focuses exclusively on scalar implicature, and the extent to which the proposed approach generalizes to other pragmatic phenomena remains an open question. Second, the construction of graded activation steering relies on a fixed steering direction estimated from a limited subset of instances, which may not capture the full variability of pragmatic representations across contexts. Finally, interpretive preference is operationalized through representational similarity rather than explicit behavioral judgments, which provides a controlled internal measure but may not fully reflect end-user interpretations. Future work may address these limitations by extending the approach to a broader range of pragmatic phenomena, model architectures, and evaluation measures.

References

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. 2025. Steering large language model activations in sparse spaces. *arXiv preprint arXiv:2503.00177*.

Ye-eun Cho. 2025. Prompting strategies of generative ai for korean pragmatic inference. *Korean Journal of Linguistics*, 50(2):423–455.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Ye-eun Cho and Yunho Maeng. 2025. Can vision-language models infer speaker's ignorance? the role of visual and linguistic cues. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 298–308, Suzhou, China. Association for Computational Linguistics.

Judith Degen and Michael K Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science*, 39(4):667–710.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 15789–15809.

Bertram Højer, Oliver Jarvis, and Stefan Heinrich. 2025. Improving reasoning performance in large language models via representation engineering. In *13th International Conference on Learning Representations (ICLR 2025)*, pages 1–18. International Conference on Learning Representations (ICLR).

Laurence Robert Horn. 1972. *On the semantic properties of logical operators in English*. University of California, Los Angeles.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2025. Reading between the prompts: How stereotypes shape LLM's implicit personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20378–20411. Association for Computational Linguistics.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Grains: Gradient-based attribution for inference-time steering of llms and vlms. *arXiv preprint arXiv:2507.18043*.

OpenAI. 2024. *Hello GPT-4o*. OpenAI.

Elizabeth Pankratz and Bob Van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4):562–594.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024. Pragmatic competence evaluation of large language models for the korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266.

Eszter Ronai and Ming Xiang. 2021. Pragmatic inferences are qud-sensitive: An experimental study. *Journal of Linguistics*, 57(4):841–870.

Eszter Ronai and Ming Xiang. 2024. What could have been said? alternatives and variability in pragmatic inferences. *Journal of Memory and Language*, 136:104507.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36:20827–20905.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint arXiv:2501.09929*.

Charles Spearman. 1987. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Fang-Yi Su, Gia-Han Ngo, Ben Phan, and Jung-Hsien Chiang. 2025. Cas: enhancing implicit constrained data augmentation with semantic enrichment for biomedical relation extraction and beyond. *Database*, 2025:baaf025.

Manan Suri, Nishit Anand, and Amisha Bhaskar. 2025. Mitigating memorization in llms using activation steering. *arXiv preprint arXiv:2503.06040*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024a. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team et al. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Polina Tsvilodub, Kanishk Gandhi, Haoran Zhao, Jan-Philipp Fränken, Michael Franke, and Noah D Goodman. 2025. Non-literal understanding of number words by language models. *arXiv preprint arXiv:2502.06204*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Bob van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of semantics*, 33(1):137–175.

Weixuan Wang, Jingyuan Yang, and Wei Peng. 2024. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2300–2344.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Scalar Diversity in Human Judgments

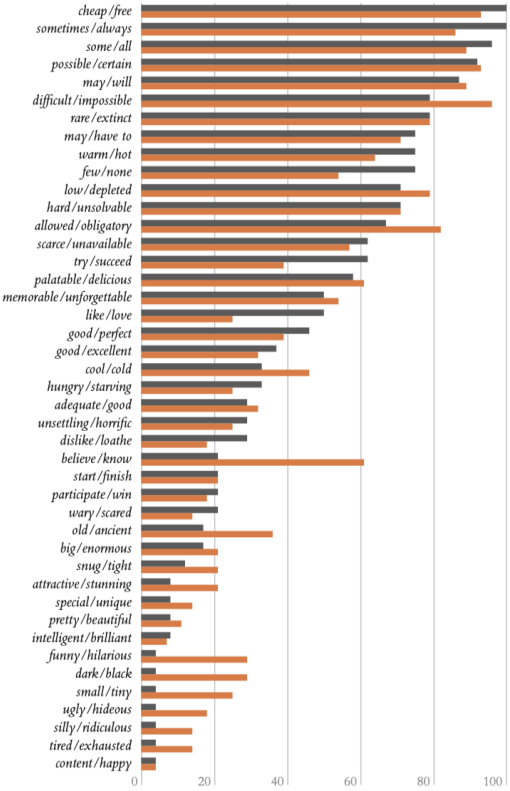

Figure 7: Graded scalar diversity observed in human judgments, adapted from van Tiel et al. (2016). Different scalar items exhibit systematically varying strengths of pragmatic interpretation

B Scatter Plots for Uniform Activation Steering

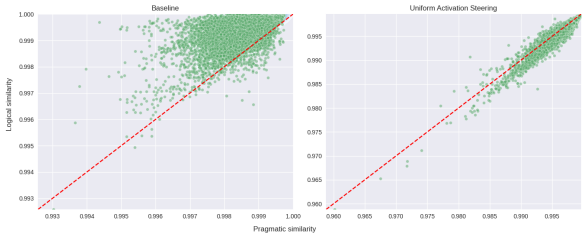

Figure 8: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (left) and uniform activation steering (right) for LLaMA3

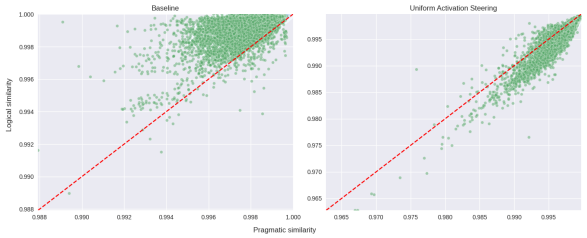

Figure 9: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (left) and uniform activation steering (right) for Qwen2

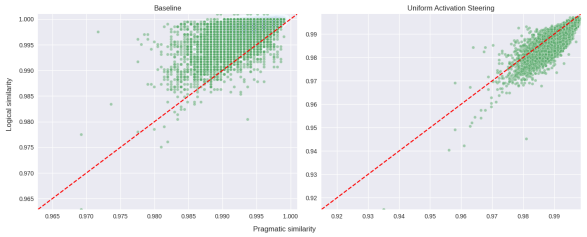

Figure 10: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (left) and uniform activation steering (right) for Gemma2

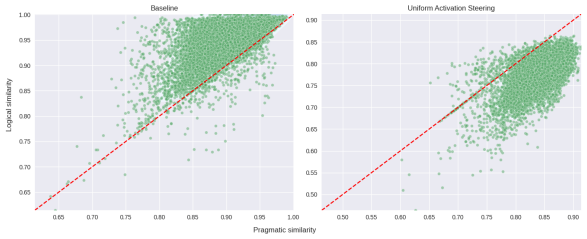

Figure 11: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (left) and uniform activation steering (right) for OLMo

C Items in GraSD dataset

C.1 Scalar Item Pairs from Source Studies

van Tiel et al. (2016)	Ronai & Xiang (2021)	Pankratz & van Tiel (2021)	Ronai & Xiang (2024)
cheap-free	sometimes-always	grey-black	partially-completely
sometimes-always	rare-extinct	thin-invisible	some-all
some-all	cheap-free	tough-impossible	mostly-entirely
possible-certain	some-all	painful-deadly	match-exceed
may-will	difficult-impossible	okay-great	slow-stop
difficult-impossible	possible-certain	tiny-imperceptible	possible-certain
rare-extinct	allowed-obligatory	hot-boiling	allowed-obligatory
may-have to	may-have to	elegant-ostentatious	reduce-eliminate
warm-hot	low-depleted	casual-sloppy	try-succeed
few-none	hard-unsolvable	mysterious-magical	primarily-exclusively
low-depleted	warm-hot	rough-unfriendly	similar-identical
hard-unsolvable	scarce-unavailable	decent-great	once-twice
allowed-obligatory	few-none	comfortable-luxurious	largely-totally
scarce-unavailable	believe-know	pale-white	difficult-impossible
try-succeed	try-succeed	thick-impenetrable	equally-more
palatable-delicious	palatable-delicious	warm-hot	usually-always
memorable-unforgettable	good-perfect	quiet-silent	tolerate-encourage
like-love	memorable-unforgettable	amazing-miraculous	permit-require
good-perfect	participate-win	cool-cold	believe-know
good-excellent	good-excellent	wrong-evil	overwhelmingly-unanimously
cool-cold	cool-cold	wet-sopping	hard-unsolvable
hungry-starving	may-will	soft-mushy	probable-necessarily
adequate-good	adequate-good	sad-tragic	double-triple
unsettling-horrific	start-finish	bizarre-surreal	here-everywhere
dislike-loathe	unsettling-horrific	uncomfortable-painful	palatable-delicious
believe-know	old-ancient	evil-satanic	cool-cold
start-finish	like-love	happy-ecstatic	warm-hot
participate-win	hungry-starving	cold-frosty	damage-destroy
wary-scared	silly-ridiculous	red-scarlet	old-ancient
old-ancient	dislike-loathe	funny-hilarious	good-excellent
big-enormous	big-enormous	dark-black	start-finish
snug-tight	wary-scared	big-huge	dark-black
attractive-stunning	tired-exhausted	nice-great	survive-thrive
special-unique	attractive-stunning	odd-bizarre	or-and
pretty-beautiful	small-tiny	busy-full	dislike-loathe
intelligent-brilliant	dark-black	honest-blunt	hungry-starving
funny-hilarious	content-happy	angry-violent	harmful-deadly
dark-black	special-unique	polite-friendly	begin-complete
small-tiny	snug-tight	dry-arid	small-tiny
ugly-hideous	pretty-beautiful	wonderful-perfect	intimidating-terrifying
silly-ridiculous	funny-hilarious	enjoyable-great	want-need
tired-exhausted	ugly-hideous	cute-beautiful	well-superbly
content-happy	intelligent-brilliant	weird-alien	understandable-articulate
		bright-brilliant	snug-tight
		damp-wet	overweight-obese
		fat-obese	like-love
		pretty-beautiful	willing-eager
		smart-brilliant	serious-life-threatening
		emotional-sentimental	pretty-beautiful
		calm-meditative	dirty-filthy
			unpleasant-disgusting
			ugly-hideous
			polished-impeccable
			intelligent-brilliant
			happy-ecstatic
			funny-hilarious
			big-enormous
			attractive-stunning
			tired-exhausted
			scared-petrified

C.2 Graded Scalar Item Pairs (A-E) in GraSD dataset

A	B	C	D	E
cheap-free	few-none	cool-cold	start-finish	intelligent-brilliant
sometimes-always	hard-unsolvable	good-excellent	wary-scared	ugly-hideous
some-all	low-depleted	wrong-evil	big-enormous	pretty-beautiful
possible-certain	scarce-unavailable	adequate-good	snug-tight	small-tiny
rare-extinct	believe-know	unsettling-horrific	silly-ridiculous	content-happy
may-have to	mysterious-magical	memorable-unforgettable	tired-exhausted	special-unique
grey-black	largely-totally	wet-sopping	dark-black	attractive-stunning
thin-invisible	double-triple	soft-mushy	funny-hilarious	enjoyable-great
tough-impossible	once-twice	sad-tragic	happy-ecstatic	cute-beautiful
painful-deadly	may-will	bizarre-surreal	big-huge	weird-alien
okay-great	warm-hot	uncomfortable-painful	nice-great	bright-brilliant
tiny-imperceptible	rough-unfriendly	evil-satanic	odd-bizarre	damp-wet
hot-boiling	decent-great	cold-frosty	busy-full	fat-obese
elegant-ostentatious	comfortable-luxurious	red-scarlet	honest-blunt	smart-brilliant
casual-sloppy	pale-white	good-perfect	angry-violent	emotional-sentimental
partially-completely	thick-impenetrable	like-love	polite-friendly	calm-meditative
mostly-entirely	quiet-silent	damage-destroy	harmful-deadly	dirty-filthy
match-exceed	amazing-miraculous	survive-thrive	begin-complete	unpleasant-disgusting
slow-stop	equally-more	or-and	intimidating-terrifying	polished-impeccable
reduce-eliminate	usually-always	palatable-delicious	want-need	wonderful-perfect
try-succeed	tolerate-encourage	dislike-loathe	well-superbly	serious-life-threatening
primarily-exclusively	permit-require	participate in-win	understandable-articulate	scared-petrified
similar-identical	overwhelmingly-unanimously	old-ancient	overweight-obese	willing-eager
allowed-obligatory	probable-necessary	here-everywhere	hungry-starving	dry-arid
difficult-impossible				

C.3 Prompt for Data Augmentation

```
Generate three English sentences based on the scalar pair <{weak}, {strong}>.

Requirements:

All three sentences must describe the same underlying proposition.

The anchor sentence must contain {weak}.

The logical sentence must replace {weak} with the strong term {strong}.

The pragmatic sentence must contain the negation of {strong} and express the
scalar implicature of the anchor.

Ensure the sentences are natural and plausible.



<Output format>

Anchor:

Logical:

Pragmatic:
```

D Item-level pragmatic interpretation rates

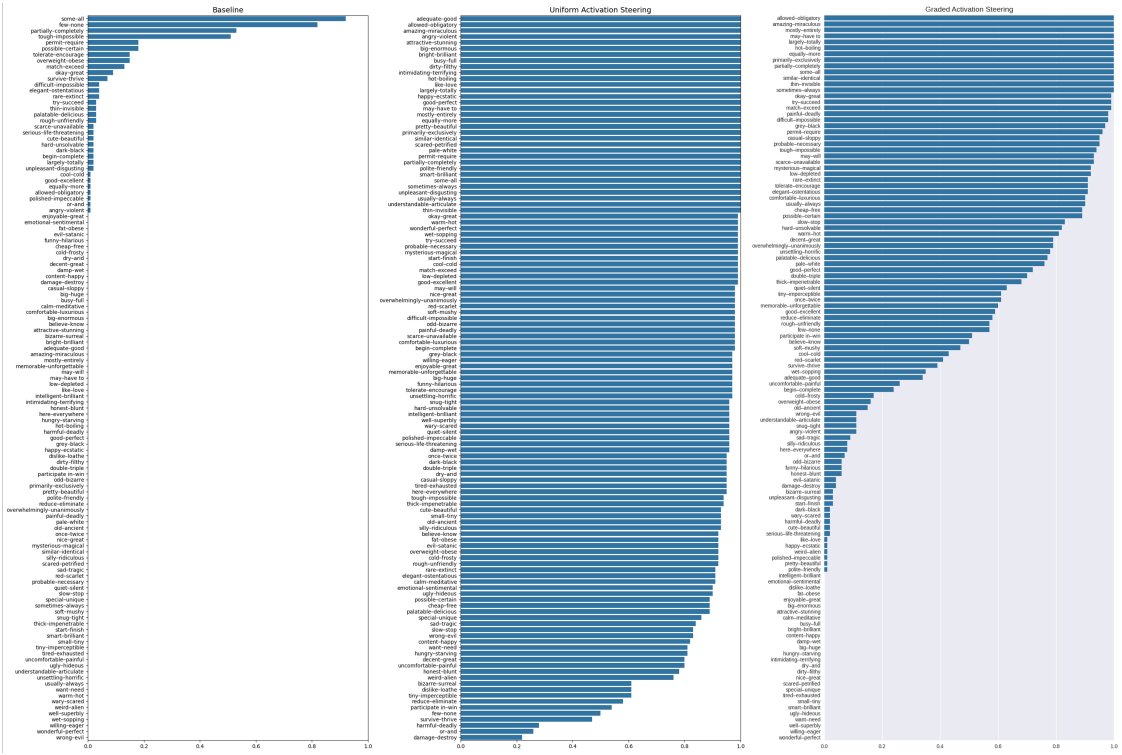

Figure 12: Item-level proportions of pragmatic interpretations for LLaMA3 across baseline, uniform activation steering, and graded activation steering conditions

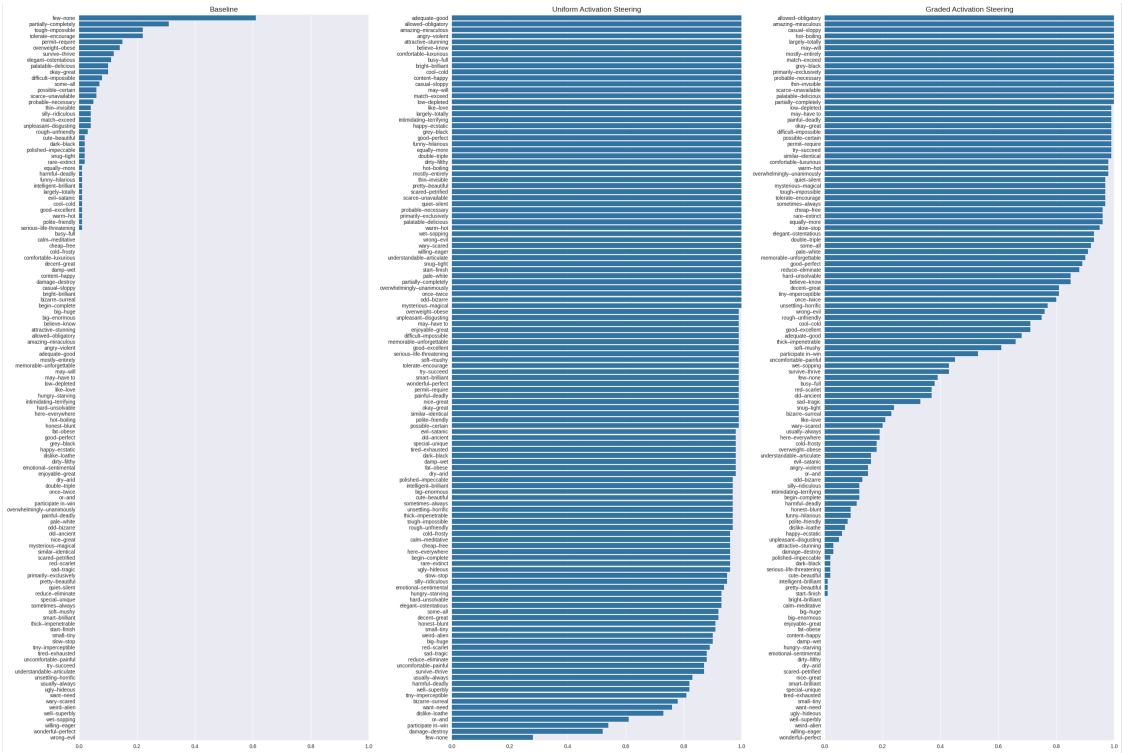

Figure 13: Item-level proportions of pragmatic interpretations for Qwen2 across baseline, uniform activation steering, and graded activation steering conditions

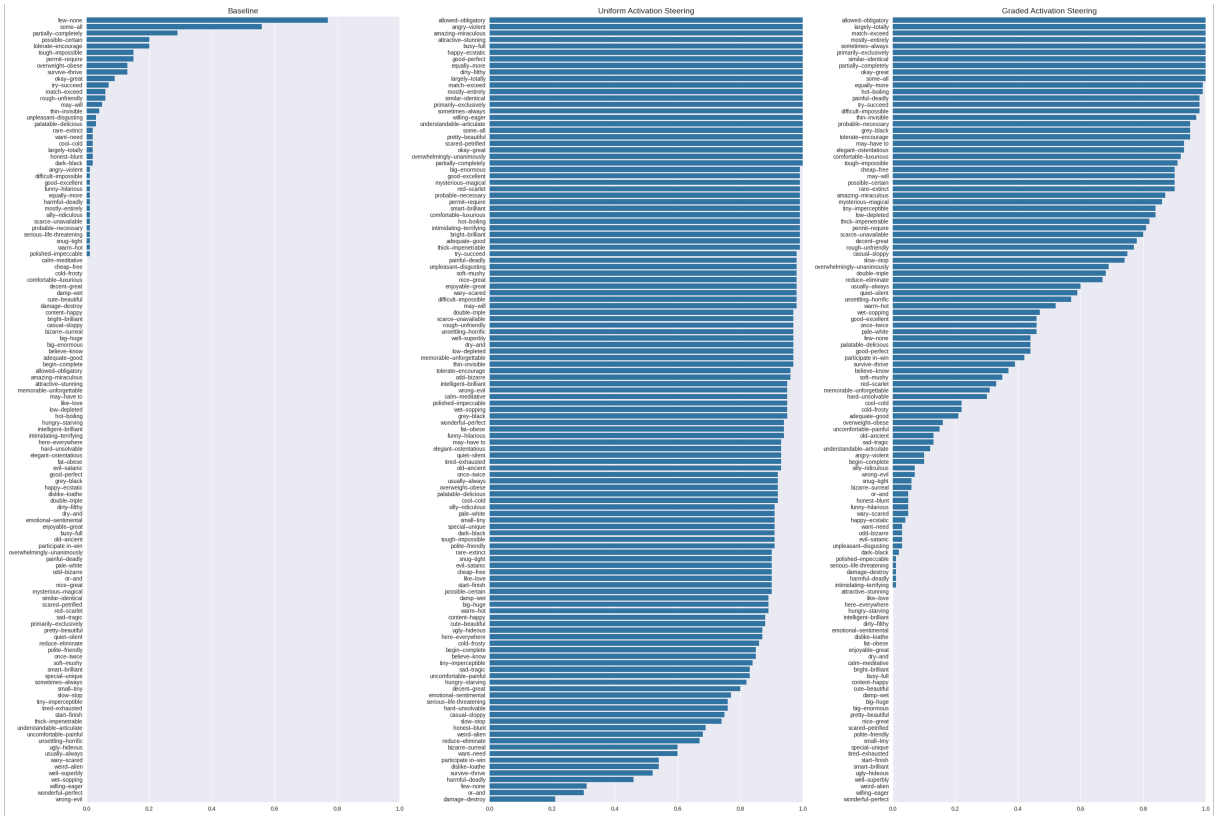

Figure 14: Item-level proportions of pragmatic interpretations for Gemma2 across baseline, uniform activation steering, and graded activation steering conditions

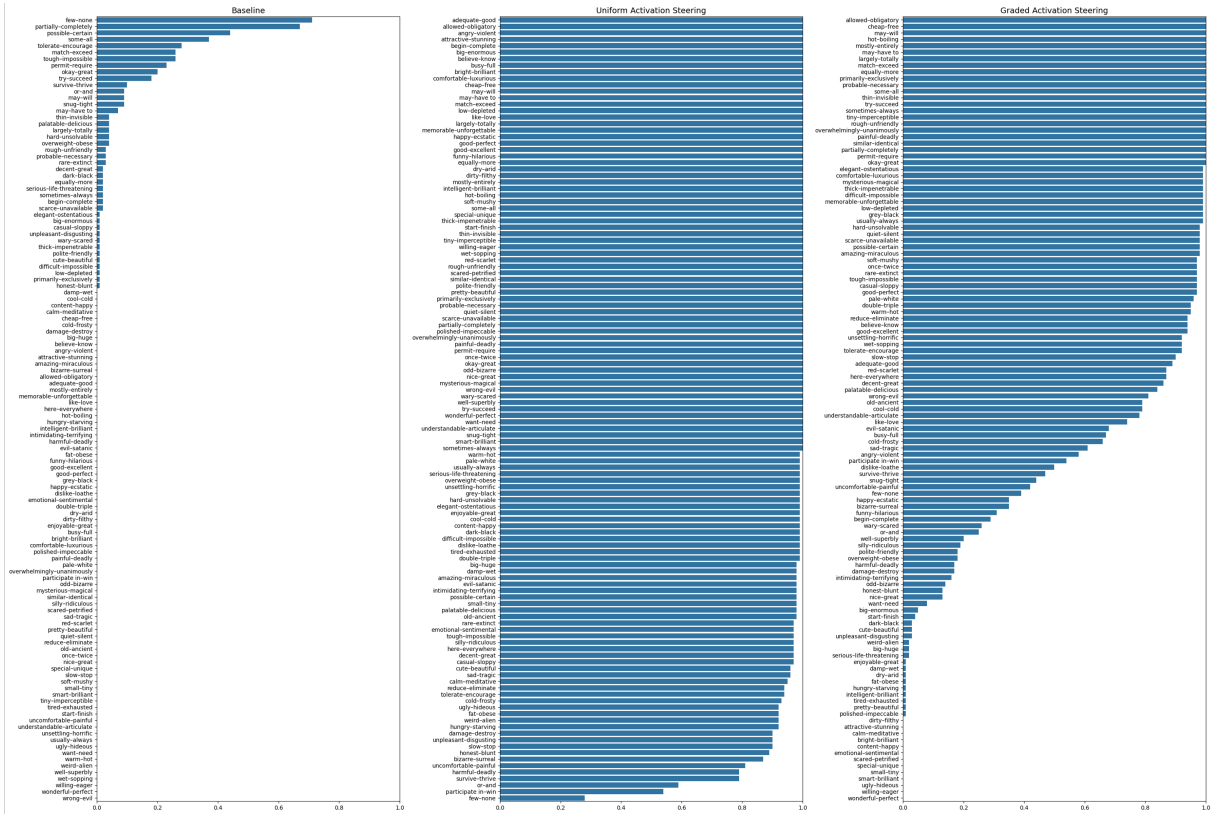

Figure 15: Item-level proportions of pragmatic interpretations for OLMo across baseline, uniform activation steering, and graded activation steering conditions

E Scatter Plots for Graded Activation Steering

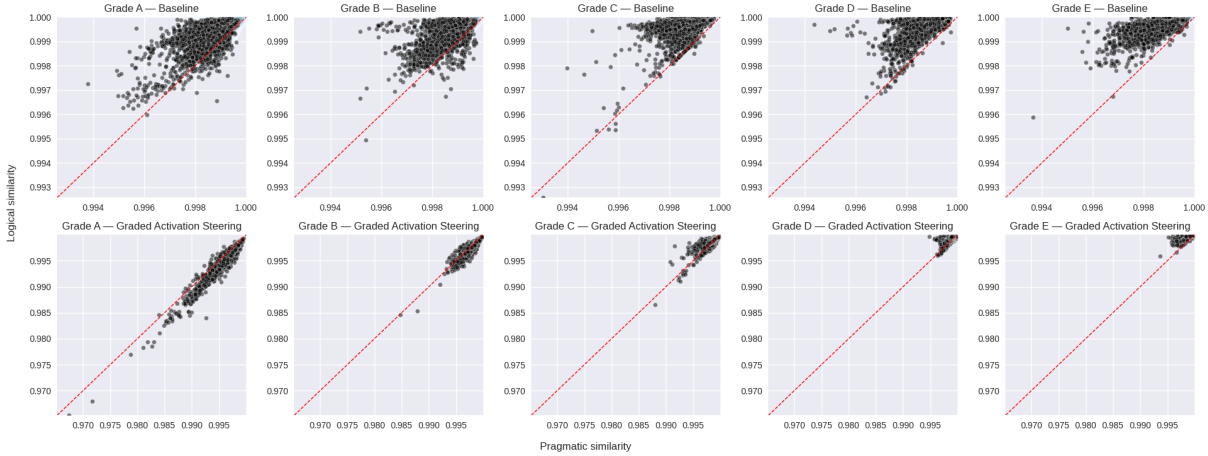

Figure 16: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (upper) and graded activation steering (lower) for LLaMA3, based on the graded item set

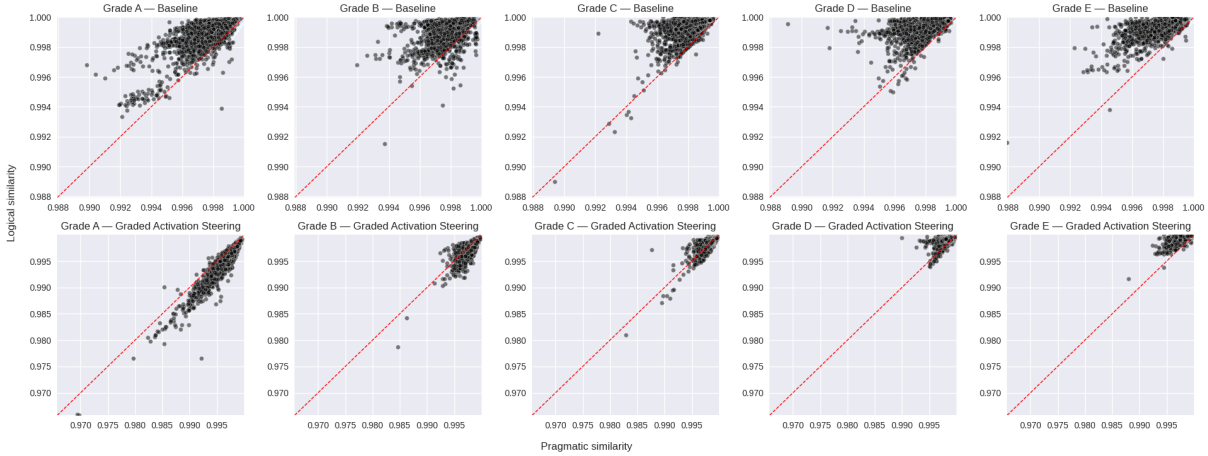

Figure 17: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (upper) and graded activation steering (lower) for Qwen2, based on the graded item set

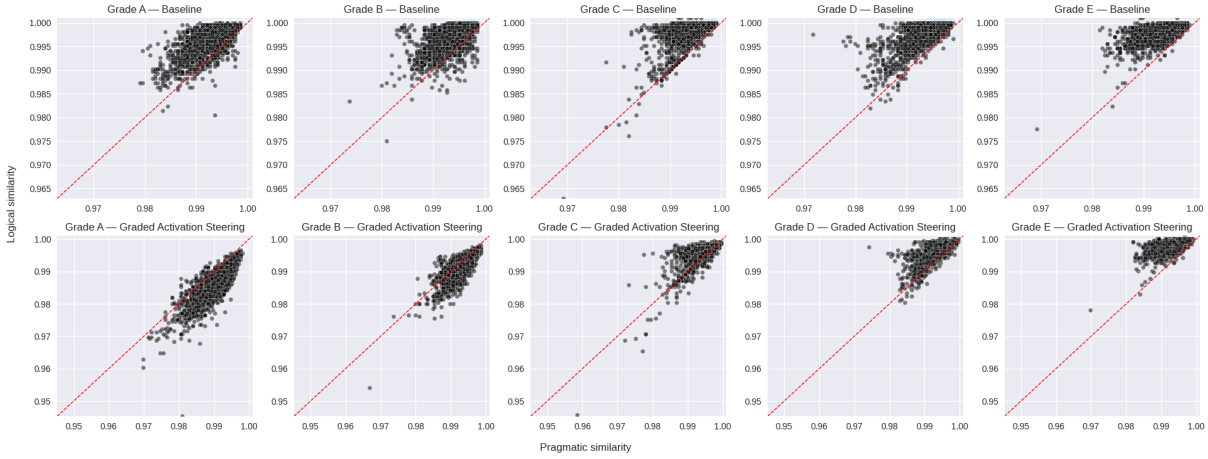

Figure 18: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (upper) and graded activation steering (lower) for Gemma2, based on the graded item set

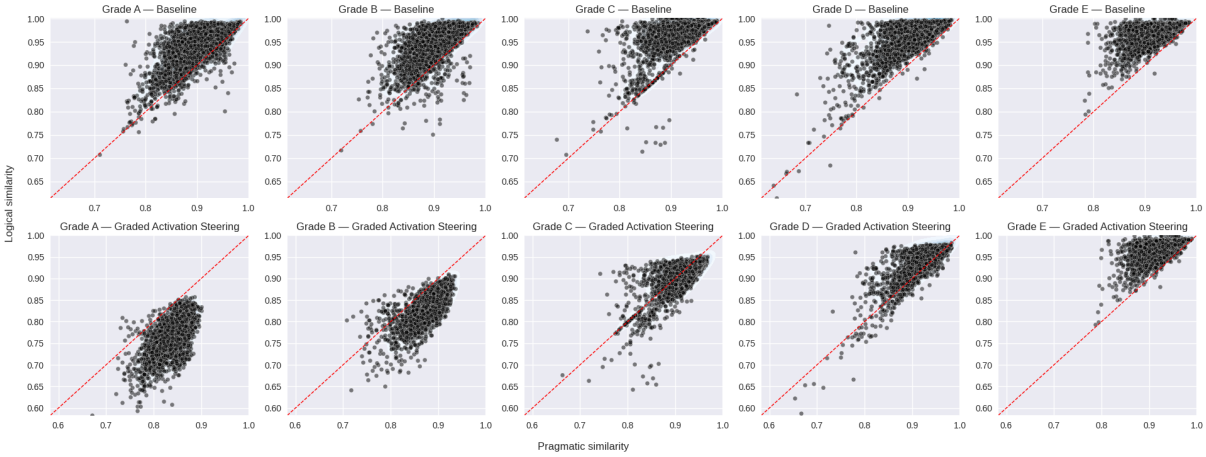

Figure 19: Scatter plots illustrating the relationship between pragmatic similarity (x-axis) and logical similarity (y-axis) under the baseline condition (upper) and graded activation steering (lower) for OLMo, based on the graded item set

F Descriptive statistics of item-level deviations under uniform and graded activation steering

	LLaMA3				Qwen2				Gemma2				OLMo			
	Uniform		Graded		Uniform		Graded		Uniform		Graded		Uniform		Graded	
Grade	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
A	0.83	0.37	0.83	0.37	0.93	0.25	0.93	0.25	0.87	0.33	0.87	0.33	0.88	0.32	0.88	0.32
B	0.92	0.26	0.79	0.41	0.93	0.24	0.84	0.36	0.91	0.29	0.70	0.46	0.94	0.24	0.92	0.28
C	0.81	0.39	0.30	0.46	0.88	0.32	0.45	0.50	0.80	0.40	0.22	0.41	0.92	0.27	0.69	0.46
D	0.91	0.28	0.04	0.19	0.95	0.22	0.09	0.29	0.88	0.32	0.03	0.16	0.97	0.17	0.22	0.41
E	0.94	0.23	0.001	0.035	0.97	0.16	0.00	0.00	0.92	0.27	0.00	0.00	0.98	0.15	0.005	0.073

G Distributions of item-level changes under uniform and graded activation steering

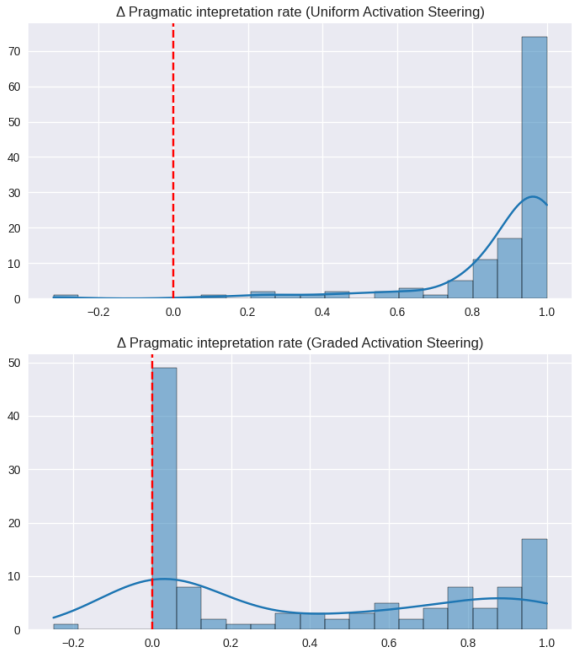

Figure 20: Histograms showing item-level changes in pragmatic interpretation rate (Δ relative to baseline) under uniform (top) and graded (bottom) activation steering for LLaMA3

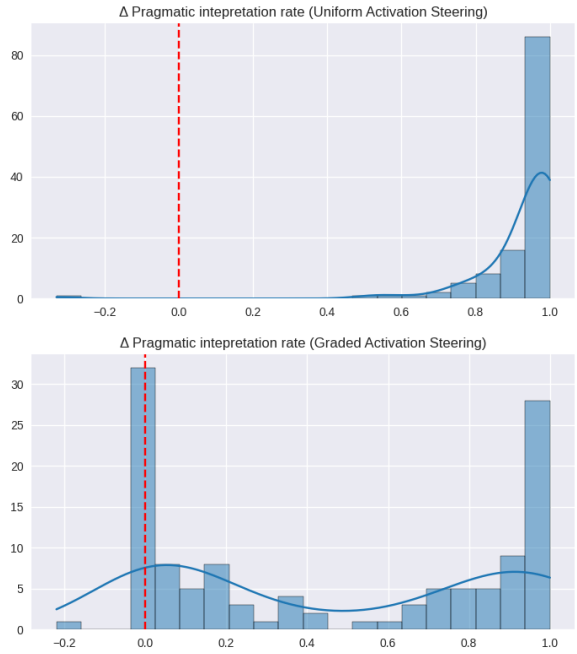

Figure 21: Histograms showing item-level changes in pragmatic interpretation rate (Δ relative to baseline) under uniform (top) and graded (bottom) activation steering for Qwen2

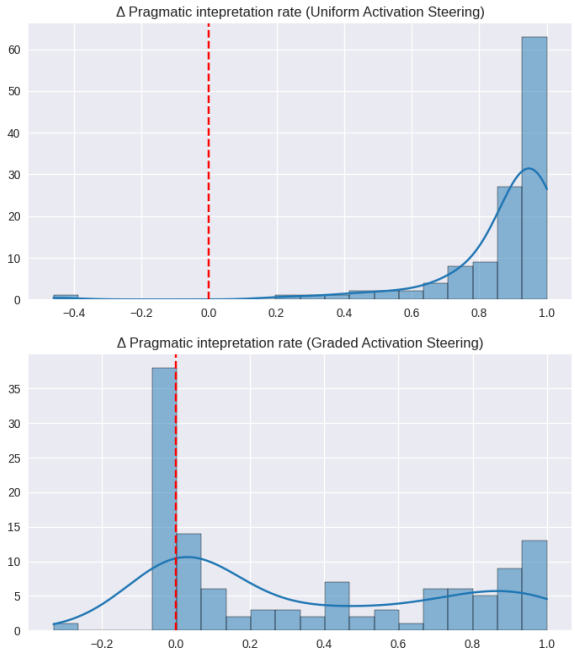

Figure 22: Histograms showing item-level changes in pragmatic interpretation rate (Δ relative to baseline) under uniform (top) and graded (bottom) activation steering for Gemma2

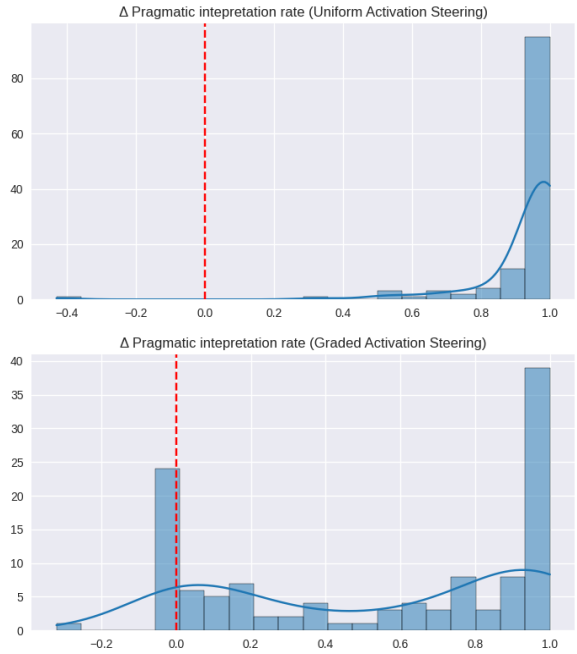

Figure 23: Histograms showing item-level changes in pragmatic interpretation rate (Δ relative to baseline) under uniform (top) and graded (bottom) activation steering for OLMo