

"But Isn't It All Just Semantics?" The Importance of Evaluating Uncertainty, Reliability, and Random Effects

Katarina Pejcinovic¹, Marianne Casilio², Robert C. Gale¹,
Gerasimos Fergadiotis², Steven Bedrick¹,

¹Division of Informatics, Clinical Epidemiology and Translational Data Science,
Oregon Health & Science University

² Department of Speech and Hearing Sciences, Portland State University

Correspondence: pejcinovic@ohsu.edu

Abstract

Picture naming tests, which involve showing images and asking patients to name them, are widely used clinical and research tools to evaluate a patient's ability to retrieve and produce words. One essential aspect of such tests is categorizing the type of semantic similarity between the image and the patient's response, as classifying errors is important for treatment planning and prognosis. These tests are tedious to score and prone to inter-rater variability, both of which could be improved with automation via LLMs. This task highlights the clinical potential of LLMs, but also the need for robust statistical testing of LLM performance. In this work, we show that traditional evaluation methods can be misleading for complex experimental designs. Using generalized linear mixed effects models and conformal prediction for uncertainty quantification, we identify sources of variation in model performance and estimate the uncertainty of model predictions. We show that seemingly large differences in model performance from traditional evaluations are misleading, driven instead by item-level and category difficulty. Additionally, uncertainty in predictions varies substantially between semantic categories. These results suggest that for NLP evaluations with nested, clustered, or uneven category distributions, uncertainty quantification can be necessary to avoid misleading conclusions for high-stakes clinical tasks.

1 Introduction

Confrontation picture naming tests (PNTs) such as the Boston Naming Test (Kaplan et al., 2001) are widely used in both clinical and research settings to assess an individual's abilities at word retrieval and production. During the test, a patient is presented with a series of images ("targets") and asked to name what is depicted in each image ("responses" or "productions"). This ability to retrieve and produce the correct word can be impaired in individuals who have suffered a stroke or other traumatic

brain injury (TBI), and manifests as *anomia*, or an inability to recall and produce words for objects (Goodglass and Wingfield, 1997). Anomia is one of the key markers of the speech and language disorder known as aphasia, and correctly classifying anomic errors, known as paraphasias, is important for treatment planning (Kendall et al., 2008) and prognosis (Schwartz and Brecher, 2000). Beyond their use with TBI and stroke patients, PNTs are used in a variety of clinical and research scenarios in neurology, psychology, and speech-language pathology; however, their use is not without difficulties.

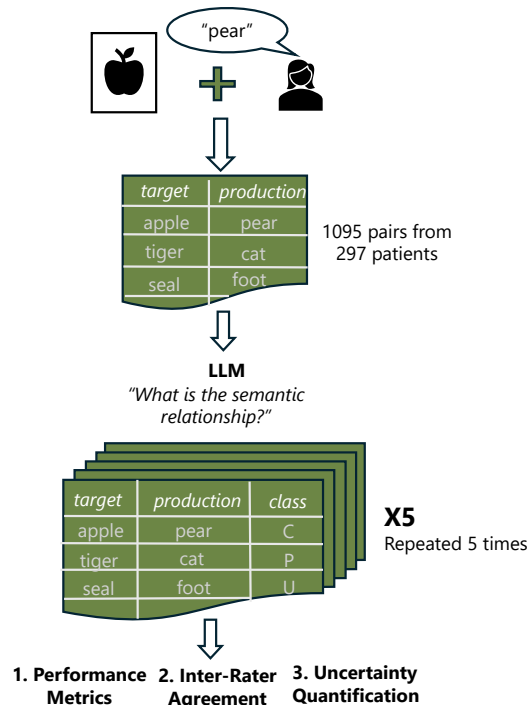


Figure 1: Experimental Pipeline.

Currently, highly trained speech-language pathologists administer tests, which can limit access. Additionally, administering such tests can be

a lengthy and tedious affair for both the clinician and the patient. Perhaps more importantly, naming test scoring can be highly subjective on the part of the clinician administering the test, particularly pertaining to the determination of the semantic similarity between the image and the patient’s response, an essential aspect of classifying paraphasias (Roach et al., 1996).

This aspect of naming test scoring has clear overlap with the analysis and quantification of semantic similarity, something which has been of interest within NLP for many decades, from the foundational Rubenstein and Goodenough (Rubenstein and Goodenough, 1965), to more recent benchmarks (Hill et al., 2015) (see section 2). However, the *clinical* dimensions of semantic similarity, in particular the definitions of similarity that are relevant to the context of confrontation naming tests, remain underexplored and lack strong benchmarking.

For many purposes, a simple binary classification of semantic similarity (i.e., related vs. not) between image and response is sufficient; other times, however, a finer-grained analysis is warranted. In both clinical and research settings, it is sometimes helpful to understand the ways in which a patient’s response is (or is not) related to the target image: for example, if a subject is systematically producing hyponyms, that could mean something clinically different than if they were producing co-hyponyms. Naming tests represent a high-impact opportunity for clinical automation, which would increase the accuracy and availability of these tests by reducing the time needed to classify semantic similarity both from a binary and fine-grained perspective as well as possibly decreasing between-provider variability in test administration (Salem et al., 2023; Casilio et al., 2023).

When considering the use of automation in clinical scenarios, the question of evaluation is of paramount importance. One must consider not only correctness (i.e., the system’s output must be accurate) but also validity: the metrics and evaluation methodologies that one chooses must capture clinically relevant aspects of the system’s behavior (Alaa et al., 2025); LLMs, with their inherent stochasticity, also introduce significant concerns around consistency. From a methodological standpoint, many clinical scenarios (including the naming task described above) bring with them complex nested study designs, which often render traditional descriptive approaches to machine learning eval-

uation (e.g., simple comparison of classification metrics such as accuracy, etc.) insufficient at best and misleading at worst.

Category	Example
Synonym	toilet → commode
Category coordinate (Co-hyponymy)	banana → apple
Superordinate (Hypernymy)	apple → fruit
Subordinate (Hyponymy)	flower → rose
Associated	bench → park
Diminutives	dog → doggy
Unrelated	sock → mock

Table 1: Examples of semantic similarity types.

Therefore, in this work we aim to rigorously evaluate LLMs on this particular task of classifying clinically relevant categories of semantic similarity. Our first aim is to quantify variability between model runs at this task with inter-run variability metrics. The second aim is to determine whether LLMs can accurately classify different types of semantic similarity and derive robust statistical estimation of performance across tasks using generalized linear mixed models, which help disentangle effects of interest (e.g., differences between classifiers, hyperparameter settings, etc.) from other effects (artifacts introduced by nested structure in one’s data or experimental design, etc.) (Ferro and Silvello, 2016).

The third and final aim is to quantify the uncertainty in predictions utilizing conformal prediction (CP), in particular comparing uncertainty among the different semantic similarity classes. CP is a rigorous uncertainty quantification method (Vovk et al., 2005; Angelopoulos and Bates, 2023), and has had success in classifying multiple-choice style outputs from LLMs (Kumar et al., 2023). All of these aims were assessed for a closed, proprietary model, GPT-4o-mini (GPT) (OpenAI, 2024), and two locally hosted 27B parameter models: LLaMA-3.3-Instruct (Llama) (Touvron et al., 2023), and MedGemma (Sjellergren et al., 2025).

2 Related Work

There has been previous work on automating the classification of picture-naming errors. Salem et al. introduced the ParAlg algorithm, which uti-

LLM Classification Accuracy by Semantic Group						
Macro binary accuracy and F1 followed by exact category match by type. Mean and 95% CI across 5 runs.						
Semantic Group	MedGemma-27B		GPT-4o-mini		Llama-3.3-Instruct	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Binary (Related/Unrelated)	78.6%	(78.0, 79.3)	76.5%	(75.8, 77.1)	83.2%	(82.7, 83.7)
F1 Score	83.0%	(82.9–83.2)	66.9%	(65.8–67.9)	83.1%	(82.9–83.3)
associated	62.7%	(61.5, 64.0)	48.9%	(46.6, 51.1)	64.1%	(63.8, 64.4)
category coordinate	35.7%	(34.6, 36.8)	16.4%	(14.5, 18.2)	45.8%	(43.6, 48.0)
diminutives	100.0%	(100.0, 100.0)	100.0%	(100.0, 100.0)	100.0%	(100.0, 100.0)
subordinate	38.3%	(32.7, 44.0)	41.7%	(31.3, 52.0)	36.7%	(31.0, 42.3)
superordinate	17.5%	(12.3, 22.7)	13.8%	(10.3, 17.2)	58.8%	(51.3, 66.2)
synonym	80.0%	(78.1, 81.9)	25.2%	(21.6, 28.8)	48.3%	(48.3, 48.3)
unrelated	82.4%	(81.7, 83.1)	94.6%	(93.7, 95.5)	85.1%	(84.6, 85.7)

Table 2: Descriptive accuracy and 95% confidence intervals across 5 independent runs and stratified by semantic category. The binary row represents the accuracy when collapsing all the semantic categories into related vs. unrelated.

lized and compared the effectiveness of word2vec and BERT in classifying binary semantic similarity (Salem et al., 2023). ParAlg was later expanded by Casilio et al. in 2023 to include multinomial classification (Casilio et al., 2023).

Additionally, quantifying semantic similarity has long been of interest to the NLP community, starting with the seminal work of Rubenstein and Goodenough (Rubenstein and Goodenough, 1965). Semantic similarity quantification benchmarks, such as WordSim-353 (Agirre et al., 2009) and SimLex-999 (Hill et al., 2015), served as critical tools in the development and evaluation (Camacho-Collados and Pilehvar, 2018; Apidianaki, 2023), as well as the empirical validation (Caliskan et al., 2017; Mandera et al., 2017), of modern methods for lexical representation and analysis in NLP. In more recent years, there has been interest in using LLMs to quantify semantic similarity (Moskvoretskii et al., 2024; Diera and Scherp, 2026), though to our knowledge no studies have focused specifically on clinically relevant semantic categories, nor have they assessed this specific clinical context.

This work extends prior work on automatic PNT classification on three fronts. First, this work builds upon previous automation that used word2vec and BERT to now assessing using LLMs. Second, we expand previous classification tasks from binary

to multiple semantic similarity categories. Finally, we also incorporate uncertainty quantification (UQ) methods, specifically GLMMs and conformal prediction. For clinical applications of NLP, UQ is especially important because model errors may vary systematically by item, patient, or semantic category; this variation, left uncharacterized, can severely limit the clinical usability of NLP-powered tools. Our use of GLMMs and conformal prediction therefore complements prior work on PNT automation by explicitly modeling variation in performance and estimating uncertainty around model predictions. Finally, this work builds upon previous work on evaluating LLMs at semantic similarity quantification by assessing clinically-meaningful definitions of semantic similarity.

3 Methods

3.1 Data

Data were obtained from the publicly available Moss Aphasia Psycholinguistics Project Database (MAPPD) dataset (Mirman et al., 2010). This dataset contains representative coverage of seven well-characterized categories of semantic error seen in aphasia assessment, and is widely used in aphasiology research, thus facilitating comparison and reuse of our results. Its class distribution is unbalanced, reflecting a realistic distribution of

how frequently different forms of semantic errors occur in the aphasic patient population. The dataset is composed of 95% monolingual English speakers, with the majority of participants identifying as White (53%) and Black (39%) (Mirman et al., 2010).

Data were filtered to contain only productions that were words in the Merriam-Webster Online English Dictionary, which yielded 1095 unique target (object in image) production (transcribed patient response) pairs. An example target-production pair is (*apple, pear*), as seen in Figure 1. The seven semantic similarity classes, which include the class of *unrelated*, are provided in Table 1. A ground truth semantic similarity classification was obtained for each target-production pair. The ground truth label was derived from the majority vote of three independent human raters with relevant clinical expertise (human rater prompt provided in [Supplementary Material](#)). For the 27 pairs that did not have a majority consensus, an expert speech-language pathologist was consulted to select a single semantic similarity ground truth.

3.2 Experiments

The following experiments were repeated on GPT-4o-mini (GPT), as well as on locally-hosted 27B parameter LLaMA-3.3 Instruct (Llama) (Touvron et al., 2023) and 27B parameter MedGemma (Sellergren et al., 2025). The default temperatures, 1 for GPT and 0.8 for Llama and MedGemma, were selected to facilitate characterization of model stability across dimensions of the data. A 1B parameter Llama was also experimented with, but could not reliably produce structured outputs and was therefore removed from analysis. These models were selected due to infrastructure constraints, as well as to investigate the behavior of smaller models that can potentially be used on-device in a clinical context.

All models were accessed through an Open-AI compatible API interface, and the Instructor Python library (version 1.7.2) was used to enforce structured JSON output (Liu, 2024). All Llama and MedGemma models were hosted as HTTP servers and accessed via a locally-hosted API using the Llama.cpp backend. The HPC environment supported 4 CPU cores and two A100 NVIDIA GPUs. One run of all samples for one model took approximately 1 GPU hour with the specified setup. All code was written in Python 3.13.1 and R 4.2.

Typically, the binary classification of re-

lated/unrelated is sufficient for most diagnoses (Roach et al., 1996), though there exist clinical scenarios where the fine-grained differences in semantic similarity are clinically relevant (Reilly et al., 2011). As such, we still analyzed these more granular cases (Table 1), as these distinctions are useful for future clinical decision support interpretability and explainability, as well as elucidating potential differences in the "understanding" of semantic similarity between models and humans.

All models were queried with the same prompt, which contained the definitions of the seven categories of semantic similarity alongside examples, the target-production pair, and the context that the model should emulate a trained speech-language pathologist. The model was then asked to return the correct category of semantic similarity, or if the target-production pair is unrelated. The full prompt is provided in the [Supplementary Material](#). Please see Figure 1 for a schematic of the data collection and the experimental design.

Contrast	Class	Estimate	SE	<i>p</i>
gemma-gpt	A	1.423	0.137	<0.001
gemma-llama	A	-0.139	0.129	0.53
gpt-llama	A	-1.562	0.138	<0.001
gemma-gpt	C	3.022	0.268	<0.001
gemma-llama	C	-1.212	0.188	<0.001
gpt-llama	C	-4.234	0.288	<0.001
gemma-gpt	B	-0.434	0.665	0.791
gemma-llama	B	0.194	0.617	0.947
gpt-llama	B	0.628	0.655	0.603
gemma-gpt	P	0.761	0.516	0.303
gemma-llama	P	-3.366	0.422	<0.001
gpt-llama	P	-4.127	0.509	<0.001
gemma-gpt	S	5.864	0.473	<0.001
gemma-llama	S	2.739	0.285	<0.001
gpt-llama	S	-3.125	0.417	<0.001
gemma-gpt	U	-3.200	0.168	<0.001
gemma-llama	U	-0.595	0.123	<0.001
gpt-llama	U	2.605	0.162	<0.001

Table 3: Pairwise model comparisons with Tukey’s Adjustment by Semantic Category. A=associated, C=category coordinate, B=subordinate, P=superordinate, S=synonym, U=unrelated

3.2.1 Analysis 1: Inter-Run Variability

Each model was queried five independent times for each target-production pair. This experiment was done to estimate the inter-run agreement per model

that can be attributed to the stochastic nature of each LLM. Krippendorff’s α (Krippendorff, 2018), a well-established metric for inter-rater agreement, was calculated per semantic category. In this analysis, each of the separate 5 queries was treated as a separate "rater", as defined by Krippendorff’s α . A higher α score for a given category would indicate more stable behavior by the model for that category, while lower levels of agreement would indicate the opposite. A threshold of 0.8 is typically accepted as indicating "good agreement" between raters (Krippendorff, 2018). Krippendorff’s α was calculated using the R `irr` package (Gamer et al., 2019).

3.2.2 Analysis 2: Generalized Linear Mixed Effects Models

While descriptive statistics can give a general idea of model performance, they are usually insufficient and indeed sometimes misleading, especially for nested, clustered or hierarchical experimental designs. Hence, we employed Generalized Linear Mixed Effects Models (GLMMs) as a flexible method of isolating fixed effects in clustered experimental designs (Schad et al., 2020), where the goal is to isolate sources of variation. GLMMs are an underutilized but powerful method for evaluating NLP experiments, where frequently there is a need to account for model, item or patient-level effects (van Miltenburg et al., 2019; Gantt et al., 2020). In particular, we are interested in the **fixed effects** of model selection (do certain models outperform others?) and semantic category (are some categories more difficult to classify than others?), while accounting for **random effects** stemming from each individual target-production pair and model run.

The model and semantic category were treated as crossed fixed effects and were sum contrast coded, and accuracy was treated as the outcome. We also included random intercepts for target-production pairs and model runs. After descriptive analysis, all models performed with 100% accuracy on the diminutive category, so diminutive samples were dropped due to the 100% accuracy leading to unconstrained coefficient growth. The `lme4` R package was used to fit the GLMM (Bates et al., 2015). Tukey-adjusted pairwise testing, marginal means, and conditional means were calculated using the R `emmeans` package (Russell V. Lenth and Piaskowski, 2026). More detailed model specifications are available in the [Supplementary Material](#).

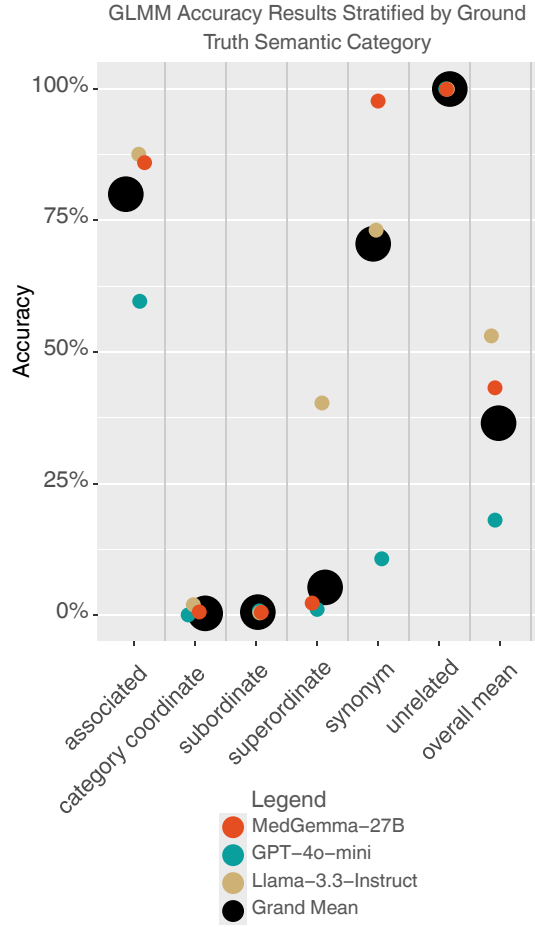


Figure 2: Accuracy of models in relation to grand mean derived from GLMM Analysis. Black circles represent the grand mean, or average accuracy across models.

3.2.3 Analysis 3: UQ with Conformal Prediction

Conformal prediction (CP) is a distribution-free UQ method that outputs confidence bounds with strong statistical guarantees (Vovk et al., 2005). For categorical outputs, these confidence bounds are represented by prediction sets that are marginally guaranteed to contain the correct class with a user-specified confidence. Inputs that are harder to classify, and therefore more uncertain, will result in larger prediction sets and vice versa. Comparing the size of prediction sets between classes is a method of comparing if certain classes are more uncertain than others.

CP starts with the selection of a scoring function that encodes the model’s built-in notion of uncertainty (Angelopoulos and Bates, 2023). In this case, the scoring function selected was $s(x, y) =$

Term	Estimate / Var	SE / SD	Lower 95% CI	Upper 95% CI
Random Effects				
Target-Production:Semantic Category	32.073	5.663		
Semantic Category	0	0		
Model Run	0	0		
Fixed Effects				
(Intercept)	-0.555	0.54	-1.613	0.502
Model				
MedGemma	0.281	0.088	0.108	0.454
GPT	-0.958	0.102	-1.158	-0.758
Llama	0.677	0.085	0.510	0.843
Class				
associated	1.94	0.659	0.648	3.233
category coordinate	-5.175	0.817	-6.775	-3.574
subordinate	-4.625	1.884	-8.318	-0.931
superordinate	-2.335	1.37	-5.02	0.349
synonym	1.429	1.342	-1.2	4.058
unrelated	8.766	0.636	7.52	10.012

Table 4: GLMM Summary. Variation (Var) and standard deviation (SD) refer to Random Effects. The Fixed Effects Intercept represents the grand mean. All estimates on the log-odds scale. A negative log odds indicates a worse-performing model on average or a more difficult to classify category on average, respectively.

$1 - \text{softmax}_y(x)$, as it represents an LLM’s inherent notion of uncertainty for a single token. A calibration dataset is used to build a new distribution of these scores.

In adaptive prediction set CP, for each calibration point, the scores of each class are ordered smallest to largest, and added together until the true class is reached (Romano et al., 2020). This creates a distribution of scores, and the quantile, which is equal to $1 - \alpha$, where α is the specified confidence, is calculated. For new test data points, the same procedure of calculating scores and ordering in increasing order is repeated. The final prediction set contains all classes such that their cumulative scores crosses the quantile threshold. In this way, prediction sets for each input are guaranteed to contain the correct class at a specified confidence. Please see Figure 5 in the Supplementary Material for a visual representation.

The CP procedure was run 100 times with an 80/20 randomized calibration/test split, stratified by patient. The CP procedure itself produces prediction sets at a 95% confidence level, while the re-sampling produces 95% confidence bounds on the size of said sets.

4 Results

4.1 Analysis 1

All tested models had high inter-run agreement across runs regardless of the ground truth semantic class. All models exceeded the 0.8 Krippendorff’s α threshold of "good agreement". Please see Supplementary Material Figure 6 for a histogram of the specific values across semantic categories.

4.2 Analysis 2

Table 2 provides a table highlighting the accuracy and 95% CI across models and semantic categories over 5 runs. The accuracy for binary related vs. unrelated classification was 78.6%, 76.5%, and 83.2%, for MedGemma, GPT, and Llama respectively. When stratifying by specific semantic categories, accuracy and confidence intervals had high variability among the models. Of the seven semantic categories, the diminutive category had the highest average accuracy among all three models at 100%. All models also had high accuracy in identifying *unrelated* target-production pairs, with an average accuracy of 87.36%. However, for the remaining categories, model accuracy dropped and varied from model to model.

The GLMM model converged with no singular-

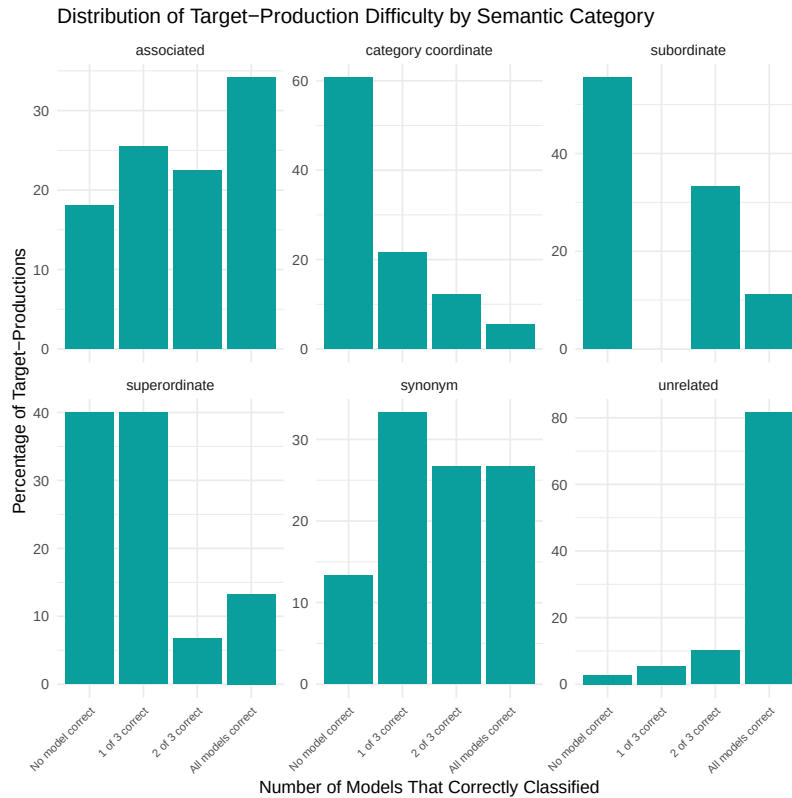


Figure 3: Distribution of Target-Pair Difficulty Stratified by Semantic Category.

ity warnings and with reasonable standard errors. Table 3 shows the results of pairwise testing between models stratified by semantic class. Pairwise analysis reveals statistically significant differences between models for all classes except for the *subordinate* class. Table 4 provides a summary of the GLMM, including variances, intercept estimates, and 95% confidence intervals.

4.3 Analysis 3

The class with the smallest set size was the *diminutive* category with a set size of 1, indicating high certainty. The *unrelated* category also shows small average set sizes across models, with the average size across models being 1.18. In general, the remaining categories showed variation among average set sizes across models. Some categories in particular showed high variation in set sizes, especially concerning the 95% confidence bound obtained by varying the calibration set. For example, the *subordinate* category has very large confidence intervals.

5 Discussion

The inter-run variability analysis showed that all the LLMs tested were highly likely to select the

same class each time, despite the relatively high temperature settings. While the inter-run variability between models was comparable, the performance, whether from an accuracy or uncertainty perspective, was not uniform across models. Despite all models being of similar size (the size of GPT was estimated to be roughly similar to the local models), the performance varied widely not only across model but also across semantic category. While Analysis 1 confirmed that models are internally consistent, this does not address whether they are consistently correct.

All models performed relatively well on the main clinical task of binary classification of related vs. unrelated. However, it is clear even from the descriptive analysis that the LLMs tested struggle with certain types of semantic relationships (ex. *subordinate*), suggesting that LLMs may struggle with distinguishing human-defined categories. Given Table 2 and the p-values from the pairwise test, it would seem some models do indeed outperform others in certain categories at a statistically significant level. For example, for the *superordinate* category, it would seem that the LLama model clearly outperforms the other models in this category (Table 2).

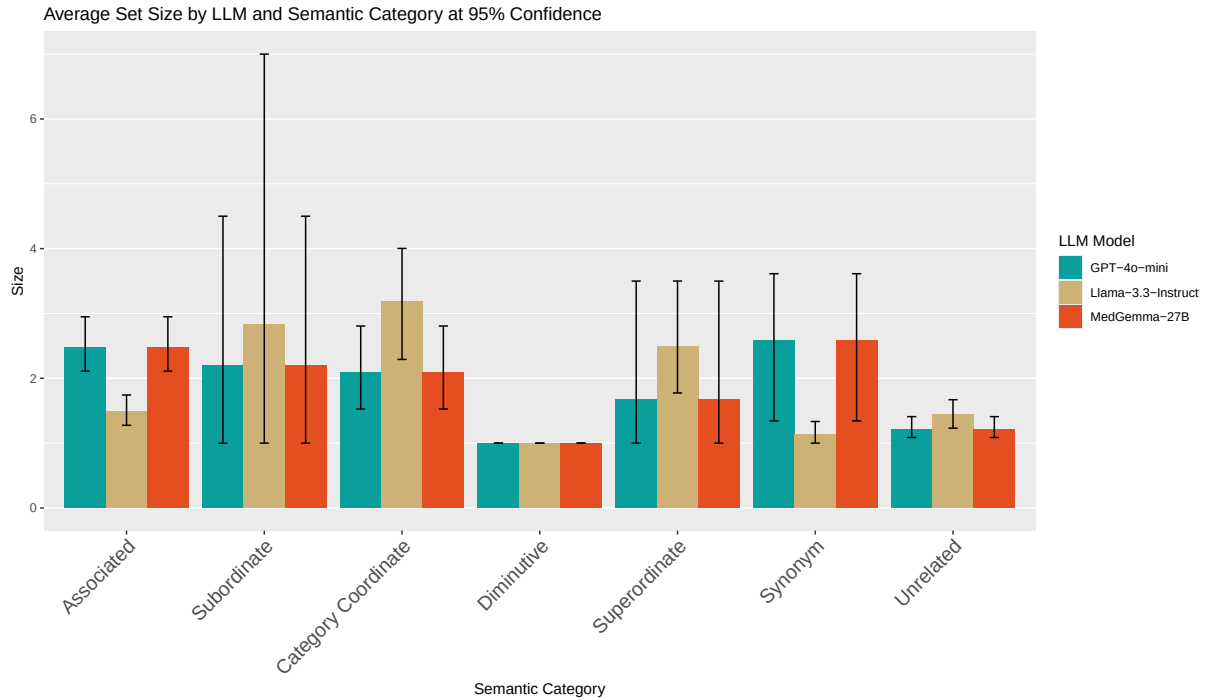


Figure 4: Adaptive Conformal Prediction results stratified by semantic category with 95% confidence intervals and 95% confidence level prediction sets. A smaller set size indicates an easier to classify category, where a set size of 1 is considered the ideal and smallest confidence bound.

However, the GLMM analysis reveals a more nuanced picture. Firstly, there is a high amount of variance stemming from the target-production pair level and the semantic category level, showing that there is both high within-semantic category and between-semantic category variance (Table 4). The discrepancy between the descriptive accuracy and the GLMM accuracy per semantic category highlights an important point: for some semantic categories, the descriptive accuracy can be misleading.

From Figure 3, we see that for classes that had a grand mean accuracy of near zero (Figure 2), the target-production pairs in these classes are far more likely to be difficult for all models, and only a subset of pairs are easily classifiable. This skewed distribution is likely what creates the discrepancy: the variation in accuracy that might seem significant with descriptive analysis and with traditional statistical hypothesis testing, is largely driven by a few easy questions that might flip between models. The reality is all models struggled with the majority of the distribution, or what would be considered the "average" question in that class. When accounting for the fact that the average difficulty target-production pairs are misclassified, the accuracy drops and we argue that despite significant

p-values, the meaningful difference between models disappears. The unbalanced nature of the data also increases the coarse accuracy, as there is simply more data for easier classes such as *unrelated*.

The variation in model performance is also revealed via UQ. The CP analysis suggests that for this particular task, some semantic relations are inherently more difficult to classify. For example, the *unrelated* category has both high accuracy and confidence across all models. This suggests that regardless of model type, simply identifying if there is a semantic relation can be done with high confidence, whereas identifying the specific type of similarity is more challenging. We also observe differences across models in the variability of prediction set sizes. The *synonym* category, for instance, has differences in the width of confidence intervals around set sizes (Figure 4). This indicates that some models are more sensitive to random change in the calibration sets and therefore exhibit higher uncertainty. We note that small sample sizes (see Table 2) can produce unstable results, and therefore CP results for categories with small sample size also contribute to higher variance between calibration sets. However, in general, the CP results align with the GLMM results, as calibration sets composed of more difficult queries lead to

larger prediction sets.

6 Conclusions

This work demonstrates that there is high variability in LLM performance when it comes to classifying semantic similarity for PNT tasks, both across model and across semantic category. This variation across semantic category seems to be driven by varying distributions of inherent difficulty between categories, making differences in accuracy artificially inflated. Additionally, different semantic classes have a higher amount of uncertainty surrounding predictions. The LLMs evaluated in this analysis show difficulty in classifying semantic relations that would be trivial for humans, such as *subordinate* and *superordinate* relations, suggesting an inherent difference in classification of semantic relationships compared to that of humans.

This work does not aim to firmly suggest one model over another for this task, especially as all of these models will not be state-of-the-art by the time of publication. Rather, we highlight the importance of rigorous quantification of LLM outputs beyond traditional metrics such as descriptive accuracy for complex clinical NLP tasks. We find that for this particular PNT task, looking only at descriptive accuracy metrics leads to incomplete or even incorrect conclusions about the relative performance of the different models. This work is an example of the value of the use of GLMMs or other mixed effects models when evaluating NLP experiments, which frequently incorporate a hierarchical or nested structure that, if left unaccounted for, can bias results.

Additionally, our results suggest that while LLMs are able to perform at a clinically satisfactory level at the core binary classification problem inherent to scoring the PNT, they would be unreliable at producing accurate granular categorizations beyond a binary decision. Depending on the scenario of use, this issue could represent a significant limitation to their clinical applicability for this task.

7 Limitations

There are several limitations to this project. Firstly, we were unable to test local models larger than 27B parameters in size due to infrastructure constraints. Secondly, these models do not represent the current state of the art in LLMs; it is possible that larger or otherwise more capable models

would exhibit more consistent behaviors.¹ Thirdly, only $\sim 1/6$ of the dataset was labeled by human raters to produce ground truth labels, limiting the statistical power of downstream analyses. Finally, the classes were highly unbalanced, though this is representative of clinical practice due to the frequency at which patients with paraphasia produce certain types of semantic errors, meaning that the behaviors observed during our evaluation are likely representative of those that would be seen in an applied setting.

8 Acknowledgments

The research reported in this publication used computational infrastructure supported by the Office of Research Infrastructure Programs, Office of the Director, of the National Institutes of Health under Award Number S10OD034224, as well as the NLM T15 grant under award number LM007088. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We also wish to thank the anonymous reviewers for their constructive and helpful comments.

References

- Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.
- Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. [Position: Medical large language model benchmarks should prioritize construct validity](#). In *Forty-Second International Conference on Machine Learning Position Paper Track*.
- Anastasios N. Angelopoulos and Stephen Bates. 2023. [Conformal Prediction: A Gentle Introduction](#). *Foundations and Trends® in Machine Learning*, 16(4):494–591.
- Marianna Apidianaki. 2023. [From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation](#). *Computational Linguistics*, 49(2):465–523.

¹That said, the evaluation methodology presented in this work would be of value in exploring this question.

- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of Statistical Software*, 67(1):1–48.
- Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. [Semantics derived automatically from language corpora contain human-like biases](#). *Science*, 356(6334):183–186.
- Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. [From Word To Sense Embeddings: A Survey on Vector Representations of Meaning](#). *Journal of Artificial Intelligence Research*, 63:743–788.
- Marianne Casilio, Gerasimos Fergadiotis, Alexandra C. Salem, Robert C. Gale, Katy McKinney-Bock, and Steven Bedrick. 2023. [ParAlg: A Paraphasia Algorithm for Multinomial Classification of Picture Naming Errors](#). *Journal of speech, language, and hearing research: JSLHR*, 66(3):966–986.
- Andor Diera and Ansgar Scherp. 2026. [Do Language Models Encode Semantic Relations? Probing and Sparse Feature Analysis](#). *arXiv preprint*. ArXiv:2603.17624 [cs.CL].
- Nicola Ferro and Gianmaria Silvello. 2016. [A General Linear Mixed Models Approach to Study System Component Effects](#). In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16*, pages 25–34, New York, NY, USA. Association for Computing Machinery.
- Matthias Gamer, Jim Lemon, and Ian Fellows Puspendra Singh. 2019. [irr: Various Coefficients of Interrater Reliability and Agreement](#).
- William Gantt, Benjamin Kane, and Aaron Steven White. 2020. [Natural Language Inference with Mixed Effects](#). In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 81–87, Barcelona, Spain (Online). Association for Computational Linguistics.
- Harold Goodglass and Arthur Wingfield. 1997. *Anomia: neuroanatomical and cognitive correlates*. Academic Press.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. [SimLex-999: Evaluating semantic models with \(genuine\) similarity estimation](#). *Computational Linguistics*, 41(4):665–695.
- Edith Kaplan, Harold Goodglass, and Sandra Weintraub. 2001. *Boston naming test*, 2 edition. Lea & Febiger.
- Diane L. Kendall, John C. Rosenbek, Kenneth M. Heilman, Tim Conway, Karen Klenberg, Leslie J. Gonzalez Rothi, and Stephen E. Nadeau. 2008. [Phoneme-based rehabilitation of anomia in aphasia](#). *Brain and Language*, 105(1):1–17.
- Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. [Conformal prediction with large language models for multi-choice question answering](#). In *ICML 2023 (neural conversational AI TEACH) workshop*.
- Jason Liu. 2024. [Instructor: A library for structured outputs from large language models](#).
- Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. [Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation](#). *Journal of Memory and Language*, 92:57–78.
- Daniel Mirman, Ted J. Strauss, Adelyn Brecher, Grant M. Walker, Paula Sobel, Gary S. Dell, and Myrna F. Schwartz. 2010. [A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function](#). *Cognitive Neuropsychology*, 27(6):495–504.
- Viktor Moskvoretskii, Alexander Panchenko, and Irina Nikishina. 2024. [Are Large Language Models Good at Lexical Semantics? A Case of Taxonomy Learning](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1498–1510, Torino, Italia. ELRA and ICCL.
- OpenAI. 2024. [ChatGPT \(GPT-4o mini version\) \[Large language model\]](#).
- Jamie Reilly, Jonathan E. Peelle, Sharon M. Antonucci, and Murray Grossman. 2011. [Anomia as a Marker of Distinct Semantic Memory Impairments in Alzheimer’s Disease and Semantic Dementia](#). *Neuropsychology*, 25(4):413–426.
- April Roach, Myrna F. Schwartz, Nadine Martin, Rita S. Grewal, and Adelyn Brecher. 1996. The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, 24:121–133.
- Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. 2020. Classification with valid and adaptive coverage. In *Proceedings of the 34th international conference on neural information processing systems, Nips '20*, Vancouver, BC, Canada. Curran Associates Inc.
- Herbert Rubenstein and John B. Goodenough. 1965. [Contextual correlates of synonymy](#). *Commun. ACM*, 8(10):627–633.
- Russell V. Lenth and Julia Piaskowski. 2026. [emmeans](#).
- Alexandra C. Salem, Robert Gale, Marianne Casilio, Mikala Fleegle, Gerasimos Fergadiotis, and Steven Bedrick. 2023. [Refining semantic similarity of paraphasias using a contextual language model](#). *Journal of Speech, Language, and Hearing Research*, 66(1):206–220.

Daniel J. Schad, Shravan Vasishth, Sven Hohenstein, and Reinhold Kliegl. 2020. [How to capitalize on a priori contrasts in linear \(mixed\) models: A tutorial](#). *Journal of Memory and Language*, 110:104038.

Myrna F. Schwartz and Adelyn Brecher. 2000. [A model-driven analysis of severity, response characteristics, and partial recovery in aphasics' picture naming](#). *Brain and Language*, 73(1):62–91.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, C{\i}an Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Sussanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. [MedGemma technical report](#).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and efficient foundation language models](#). ArXiv: 2302.13971 [cs.CL].

Emiel van Miltenburg, Merel van de Kerkhof, Ruud Koolen, Martijn Goudbeek, and Emiel Krahmer. 2019. [On task effects in NLG corpus elicitation: a replication study using mixed effects modeling](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 403–408, Tokyo, Japan. Association for Computational Linguistics.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. [Algorithmic learning in a random world](#). In *Algorithmic Learning in a Random World*. Springer New York, NY.

A Supplementary Material

A.1 Prompts

LLM propmt:

You are a speech-language pathologist scoring responses on a confrontation naming test. You have to say whether a word production is semantically related to a target word. Two words are related if and only if they share one of the following relationships:

- synonym e.g., toilet -> commode
- category coordinate e.g., banana -> apple
- superordinate e.g., apple -> fruit
- subordinate e.g., flower -> rose
- associated e.g., bench -> park
- diminutives e.g., dog -> doggie

I will give you a target and a production pair, and you will apply the above criteria to determine whether they are related. If the target and production words do not share one of those relationships,

they are unrelated. First, what is the relationship type? If not one of the above relations say unrelated. Second, is the pair related? Here is the production pair: *prod, target*

Human raters were given the following prompt: "Review the target-response pairs in the sheet and determine, based on your judgment and knowledge of the PNT guidelines, whether the two are semantically related. If you decide the target-response pair are semantically related, please (1) categorize the type of the semantic relationship using one or more of the number codes listed below, and (2) write one sentence describing why the target and response are similar".

A.2 GLMM Design

The final GLMM model is given by:

$$a \sim model * class + (1|class/tp) + (1|run)$$

where:

1. *a* is a binary indicator of whether the LLM classified the correct class
2. *model* = {*GPT*, *Llama*, *MedGemma*}
3. * indicates a crossed effect
4. *class* is the ground truth semantic category
5. (1|*class/tp*) indicates a nested random intercept for the *tp* target-production pair
6. (1|*run*) is random intercepts for the 5 runs of each model.

A.3 Code

The code repository for this project is available upon request.

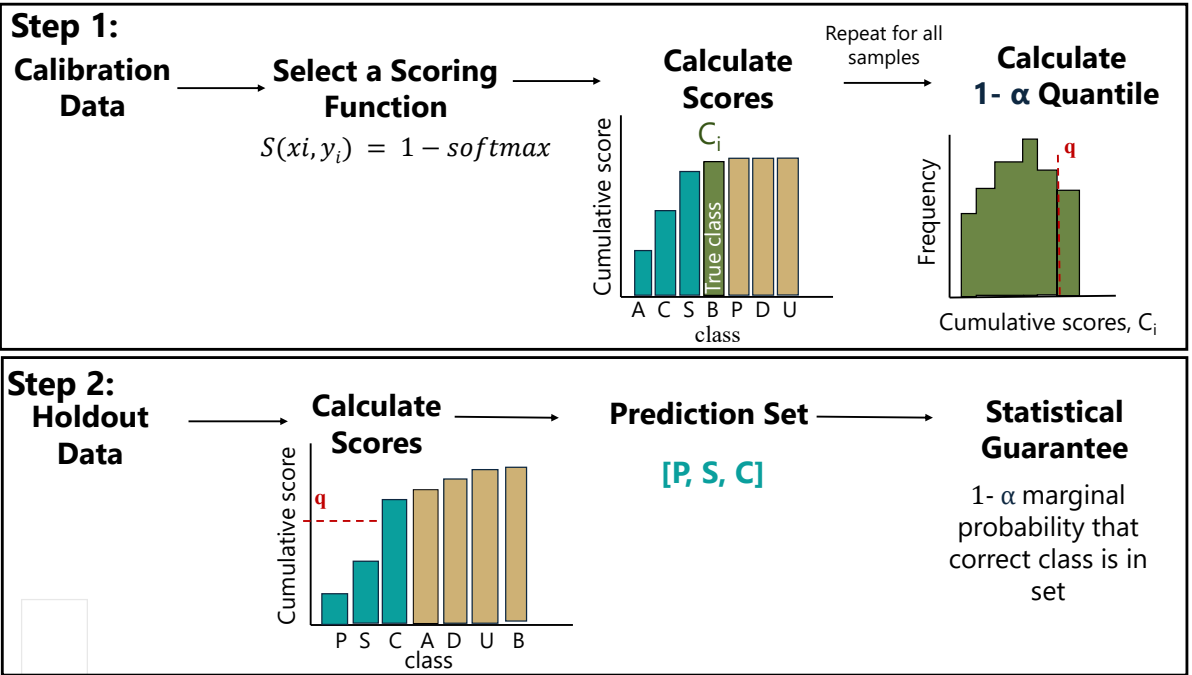


Figure 5: A visual representation of adaptive prediction sets conformal prediction.

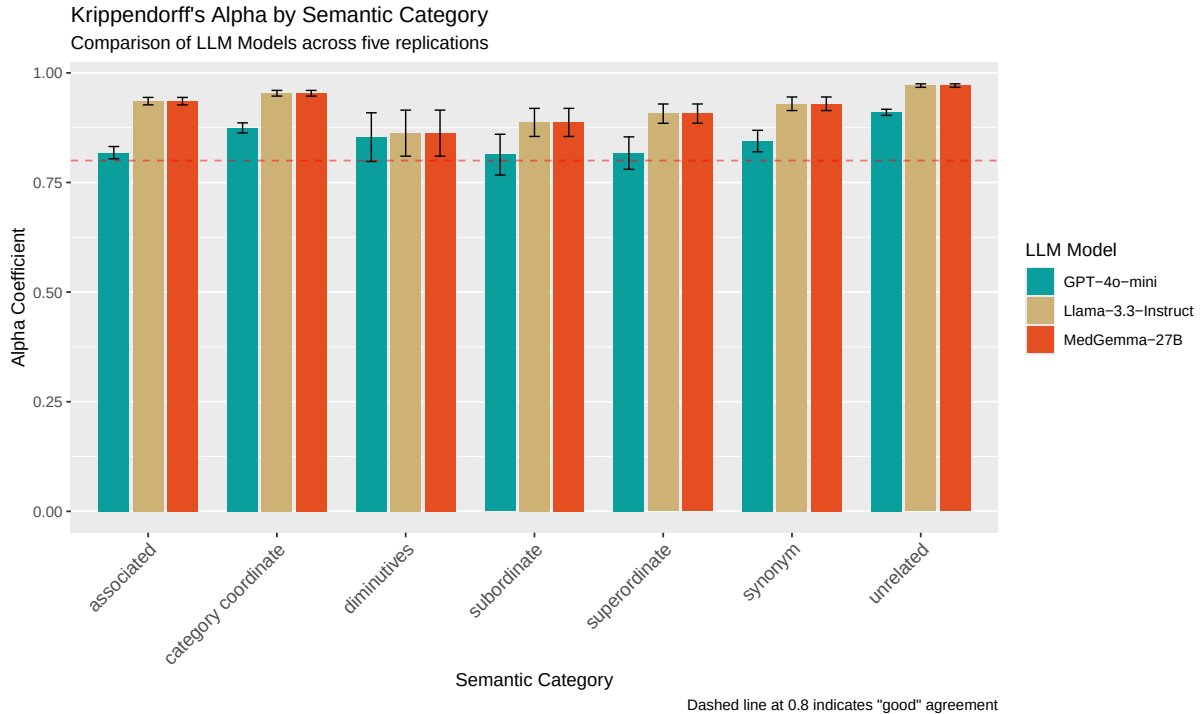


Figure 6: Krippendorff's alpha across semantic categories. Red dashed line at 0.8 indicates "good" agreement.