Can LLMs Interpret Figurative Language as Humans Do?: Surface-level vs. Representational Similarity

Anonymous ACL submission

Abstract

Large language models generate judgments that resemble those of humans. Yet the extent to which these models align with human judgments in interpreting figurative and socially grounded language remains uncertain. To investigate this, human participants and four instruction-tuned LLMs of different sizes (GPT-4, Gemma-2-9B, Llama-3.2, and Mistral-7B) rated 240 dialogue-based sentences representing six linguistic traits: conventionality, sarcasm, funny, emotional, idiomacy, and slang. Each of the 240 sentences was paired with 40 interpretive questions, and both humans and LLMs rated these sentences on a 10-point Likert scale. Results indicated that humans and LLMs aligned at the surface level with humans, but diverged significantly at the representational level, especially in interpreting figurative sentences involving idioms and Gen Z slang. GPT-4 most closely approximates human representational patterns, while all models struggle with context-dependent and socio-pragmatic expressions like sarcasm, slang, and idiomacy.

1 Introduction

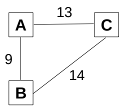
Human language is the primary medium through which individuals convey thoughts, emotions, and subjective interpretations of the world. In figurative and pragmatic language, meaning is shaped by contextual cues such as social environment, shared background knowledge, prior experiences, and tone. These contextual factors are important for interpreting pragmatically enriched forms of language, including humor, sarcasm, emotions, and slang. Large language models (LLMs) are known to generate fluent, human-like text despite lacking direct access to such contextual grounding or subjective experience.

Studies report that LLMs exhibit partial human-like behavior. Cai et al. (2024) demonstrate that ChatGPT and Vicuna reproduce a range of human linguistic effects, from phonology to pragmatics, suggesting that LLMs can approximate some aspects of human judgment. Comparing LLMs with humans across broader tasks also shows mixed strengths. Karanikolas et al. (2023) describes how LLMs balance natural language understanding and generating (NLU and NLG) like abilities in ways that echo aspects of human language use, whereas studies by Alsajri et al. (2024), Akter et al. (2023), and Atox and Clark (2024) show that ChatGPT excels in grammatical precision while Gemini often performs better in contextual comprehension or reasoning. Together, these findings suggest that LLMs may match humans on surface-level judgments while differing in deeper interpretive processes, and this varies with each model.

Although LLMs produce coherent and contextually appropriate responses, it remains unclear how well they capture human-like linguistic structures, particularly in domains that require pragmatic inference. Several authors argue that LLMs do not possess the cognitive grounding needed for genuine comprehension. Cuskley et al. (2024) state that although LLMs produce human-like text, they should not be treated as models of human linguistic or cognitive functioning. Durt et al. (2023) similarly note that human language is rooted in subjective experience and preconscious processes, which current LLMs lack. This perspective is reinforced by Dentella et al. (2024), who show that instability in responses by LLMs in basic comprehension tasks indicates reliance on surface-level pattern matching rather than structured semantic processing.

To address ongoing uncertainty regarding the depth of LLMs' language understanding, we assess whether their sentence-level interpretations resemble those of humans, using Representational Similarity Analysis (RSA)(Kriegeskorte et al. (2008); (Kriegeskorte and Kievit, 2013); Yamauchi and Wang (2025)). Consider sentences A, B, and C in Figure 1. A common approach to assessing whether humans and LLMs interpret these sentences similarly is to compare their ratings along specific dimensions (e.g., positivity, provocativeness, and dramaticity). However, ratings alone are insufficient, as semantic similarity is dimension-dependent: B and C may appear similar on one dimension, while A and C or A and B align on others. As a result, raw ratings do not reveal the underlying interpretive structure.

	Positive	Provocative	Dramatic
A: I have measured out my life with coffee spoons.	3	8	1
B: I celebrate myself, and sing myself.	7	3	1
C: I am the master of my fate, I am the captain of my soul.	8	8	9



|A – B| = 9 **|B - C|** = 14
|A – C| = 13

***B** and **C** are similar in one dimension (Positive) but they are most dissimilar overall.*

Figure 1: Hypothetical semantic similarities of sentences A, B, and C.

However, Representational Similarity Analysis (RSA) addresses this limitation by examining pairwise distance relationships among sentences. For example, computing distances (|A - B| = 9, |A - C| = 13, |B - C| = 14) reveals a stable relational structure in which A is closer to both B and C, while B and C are most distinct. By comparing such representational structures derived from human and LLM judgments, RSA allows us to test whether both systems organize meaning within a shared semantic space, rather than merely assigning similar judgment ratings.

1.1 Related Work

A central question in current research is the extent to which large language models (LLMs) process language in ways comparable to humans. One line of work examines whether LLMs align with human judgments in pragmatic interpretation tasks. For example, Bojić et al. (2025) reports that GPT-4 can outperform humans on structured dialogue-based pragmatic benchmarks, but cautions that these gains rely on highly controlled datasets that may not reflect naturalistic language use. Similarly, Hu et al. (2023) show that LLMs can make human-like choices in some zero-shot pragmatic tasks, yet systematically fail in contexts requiring theory of mind or sensitivity to speaker intent. These findings suggest that apparent alignment with human judgments may depend strongly on task structure.

At the same time, several studies highlight persistent limitations in LLMs' semantic and figurative reasoning, raising questions about whether surface-level agreement reflects deeper interpretive similarity. Shani et al. (2025) find that although LLMs can form broad conceptual categories resembling those of humans, they fail to capture fine-grained distinctions such as typicality. Likewise, Ye et al. (2025) report inconsistent metaphor comprehension, particularly under syntactic manipulations, suggesting fragility in how figurative meaning is internally represented. These results point to a potential dissociation between surface-level performance and underlying representational structure.

Taken together, these mixed findings leave unresolved whether LLMs' apparent pragmatic competence reflects genuine alignment with human interpretive representations or merely surface-level similarity (SLS), as aggregate agreement may mask failures to capture the fine-grained distinctions required for nuanced phenomena such as sarcasm, humor, emotional tone, and idiomaticity.

Recent work in model interpretability further motivates examining internal representations rather than behavioral accuracy. Fedzechkina et al. (2025) show that deeper layers in Pythia models encode hierarchical semantic information that aligns with human judgments than embedding-level similarity alone, suggesting that representational analyses can reveal alignment not visible at the surface level. This line of work supports the view that meaningful human–LLM comparison requires examining how language is internally organized, not just how outputs are rated.

Theoretical perspectives similarly emphasize why representational alignment matters for understanding human–LLM comparability. Poliak et al. (2025) argue that human comprehension relies on structural priors that guide meaning construction independently of surface word order, implying processing mechanisms that may be absent from current LLM architectures. Broader accounts by Bayzayeva (2025) and Preda (2012) conceptualize language as both communication and cognition, suggesting that LLMs optimized for prediction rather than understanding and should not be expected to fully replicate human interpretive processes.

These perspectives directly motivate the use of Representational Similarity Analysis (RSA) to assess whether alignment between humans and LLMs emerges at the level of behaviorally derived representations for pragmatic and figurative language like idiomacy, sarcasm, and Gen Z slang, rather than comparing only surface-level responses such as whether humans and models assign similar ratings to a sentence with sarcasm (e.g., "you look great!, today"). RSA evaluates whether humans and models organize sentences in structurally similar ways within their respective semantic spaces. Using this approach, Ogg et al. (2024) found that GPT-4 exhibits the closest representational alignment with humans among the models tested. However, Giallanza and Campbell (2024) demonstrates that behavioral similarity can appear human-like even when underlying representational geometry diverges, cautioning against equating surface-level alignment with shared internal structure.

Together, these studies suggest that although LLMs may align with human judgments at a surface level, it remains unclear to what extent their representations mirror human interpretation for figurative language, and how this alignment varies across models. The present work addresses this gap by jointly examining surface-level judgments and internal representational structure across multiple LLMs, enabling a systematic comparison of how closely different models process figurative and socially grounded language in ways comparable to humans.

1.2 Overview

This study examines how closely large language models (LLMs) align with human interpretations of figurative language by comparing both surface-level judgments and representational structures. Human participants and LLMs rated sentences spanning multiple semantic categories: conventional, idiomatic, emotional, funny, sarcastic, and slang on a 10-point Likert scale across 40 expression-based questions (e.g., *is this relevant to you?, Is this sarcastic?, Does this concern you?, etc*; Figure 2). The models evaluated were GPT-4, Gemma-2-9B-IT, Mistral-7B-Instruct-v0.3, and Llama-3.2-3B-Instruct.

Surface-level similarity (SLS) was assessed by computing Pearson correlations between human and model ratings for each sentence category (Figure 3a). To examine representational similarity, sentence-level representational spaces were constructed separately for humans and models by computing pairwise Euclidean distances between sentence rating vectors, yielding representational dissimilarity matrices that capture relative patterns of interpretation (Figure 3b). Representational Similarity Analysis (RSA) quantified alignment by correlating these matrices across humans and models, allowing comparison of interpretive structure independent of absolute rating values. Higher correlations indicate closer alignment in how interpretations are organized, whereas lower correlations reflect systematic differences in linguistic processing. Two studies employed different zero-shot prompting strategies to assess model sensitivity to prompt formulation.

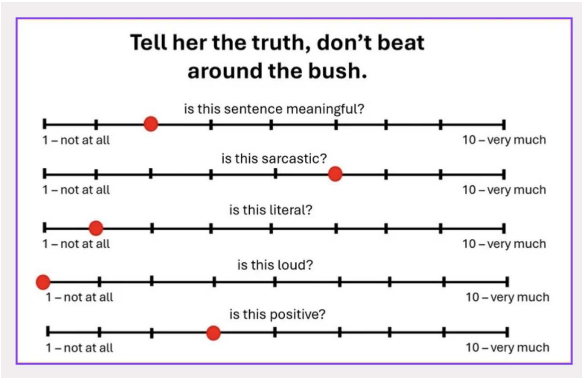


Figure 2: The figure shows an example of how a sentence and questions were presented to humans to rate on a 10-point Likert scale.

2 Method

2.1 Materials

A total of 240 sentences were compiled, comprising 120 sensible dialogues commonly used in everyday communication and 120 nonsensical counterparts (see Table 4). Sentences were classified into six categories based on their defining characteristics: conventional, idiomatic, emotional, funny, sarcastic, and Gen Z slang, with each category containing 20 sensible and 20 nonsensical sentences. To check if the models would be able to identify the linguistic trait in the sentence, Sensible sentences contained one of these defining characteristics; in contrast, non-sensible sentences maintained similar structures but intentionally omitted the key characteristics that contribute to the linguistic trait (Table 1). For example, *"bite the bullet"* was classified as a sensible idiomatic sentence, whereas *"eat the bullet,"* despite its structural similarity, was categorized as a non-sensible variant. Similarly, *"the dog is outside"* was treated as a sensible conventional sentence, while *"the outside is the dog"* was classified as its non-sensible counterpart.

Each of the 240 sentences was paired with 40 interpretive questions designed to assess meaning, tone, and expressive quality (*e.g., loud, sentimental,* *concerning, positive, etc*) (Table 5). This design yielded 9,600 sentence–question pairs, enabling a comprehensive evaluation of how linguistic features influence interpretation.

The majority of sentences and questions were created manually by the authors, with a subset extracted from Rabagliati et al. (2018); For a complete list, look at Tables 4, 5

2.2 Participants (Humans and Language Models)

Humans. A total of 235 undergraduate participants were recruited for a course credit. Data from 14 participants were excluded due to incomplete responses, resulting in a final sample of 211 participants (137 female, 73 male, 1 non-binary; M_{age} = 18.82), all of whom reported English as their primary spoken language.

Language Models. Four decoder-only, instruction-tuned large language models (LLMs) were evaluated to examine language understanding and interpretation: GPT-4 (OpenAI), Llama-3.2-3B-Instruct (Meta), Gemma-2-9B-IT (Google), and Mistral-7B-Instruct-v0.3 (Mistral). The models developed by Meta, Google, and Mistral were accessed through the Hugging Face platform, while GPT-4 was accessed via an OpenAI API key. All models were pretrained on various publicly available datasets with different parameter sizes and then instruction-tuned for better alignment with user prompts (Table 2.

Language models were provided with the same dataset of 9,600 sentence–question pairs, identical to the human task, and evaluated under zero-shot prompting conditions. The prompts were formatted in a role–content structure, with input given in the user role.

2.3 Procedure

Humans. Participants rated a subset of sentences drawn from the full set of 240 stimuli. Each sentence was paired with all 40 expressive questions and was rated on a 10-point Likert scale ranging from 1 (not likely) to 10 (very much). Participants were instructed to read each sentence carefully and provide ratings based on the displayed questions (Figure 2). Each participant evaluated approximately 32 sentences, randomly sampled from the full set and balanced across all sentence categories.

Language Models. We employed two zero-shot prompting conditions. This design allowed us to assess models' sensitivity to prompt wording and to examine whether explicit instructions led to systematic changes or improvements in rating patterns and overall performance. All models required approximately four hours to generate outputs for the full dataset and were executed on high-performance computing resources.

In **Study 1**, each model received a zero-shot prompt instructing it to read a sentence and rate it on a 10-point Likert scale in response to the accompanying question, the same as given to the participants. Each of the 9,600 sentence–question pairs was given as input individually, requiring the model to process the same prompt format repeatedly across all trials. The model outputs included both a numerical rating and a justification for each response, returned under the assistant role. This prompt was identical to the instructions given to human participants.

Prompt template 1= "Read the following sentence , and rate it on a Likert scale of 1-10 based on the given question ."

Study 2 was designed to assess whether model–human alignment improved under refined prompt conditions. It followed the same procedure, with an additional constraint added to the zero-shot prompt. Models were instructed to (a) provide ratings strictly between 1 and 10, and (b) justify their choice in a single sentence. As in Study 1, all 9,600 inputs were processed; however, the outputs followed a more structured format, consisting of a numerical rating and a concise justification.

Prompt template 2= "Read the following sentence and rate it on a Likert scale of 1-10 based on the given question and justify the rating in a sentence. Do not respond with a number other than [1, 10]."

2.4 Analysis

Surface-level Similarity(SLS): Similar outputs do not necessarily imply similar underlying representations; therefore, we distinguish between surface-level and representational similarity in our analyses. This distinction allows us to isolate cases where models match human judgments at the level of observable responses without assuming shared internal structures or interpretive mechanisms.

Surface-level similarity captures the extent to which humans and models produce comparable rating patterns, independent of how those judgments are internally organized. Pearson's correlation coefficients were computed between human ratings and each LLM's ratings from both studies, across all

Category	Type	Sentences
Conventional	Sensible Non-sensible	I adopted a dog today The dog adopted me today
Idiomatic	Sensible Non-sensible	I'm feeling under the weather The weather is over me
Emotional	Sensible Non-sensible	Max eagerly unwrapped a mysterious gift Max quietly wrapped a mysterious gift
Funny	Sensible Non-sensible	I used to be a baker because I kneaded dough I used to be a baker, and had to knead dough
Sarcastic	Sensible Non-sensible	I love your shirt, for now I love your shirt
Gen Z Slang	Sensible Non-sensible	This food is gas The gas is food

Table 1: Sample sentences from each sentence category presented to humans and LLMs for judgment.

questions within each category. This approach provides a direct measure of output-level agreement between human and model language judgments (Figure 3a).

Representational Similarity(RSA): Humans and each model were examined in terms of how they organized sentences relative to one another. For each group, pairwise Euclidean distances were calculated between all sentence pairs using their ratings across all questions, resulting in a distance matrix that captured how similar or different the sentences were in the judgment space. Because the distance matrix was symmetrical, only the upper triangle containing the unique sentence pairs was retained. Correlation coefficients were then computed between the humans' and each model's distance matrix, separately for each category of sentences (Figure 3b). This analysis quantified how similarly humans and models structured and represented relationships among sentences based on their ratings. Intuitively, sentences that were closer together in this space were interpreted more similarly, whereas sentences that were farther apart were interpreted more differently Kriegeskorte et al. (2008); Yamauchi and Wang (2025).

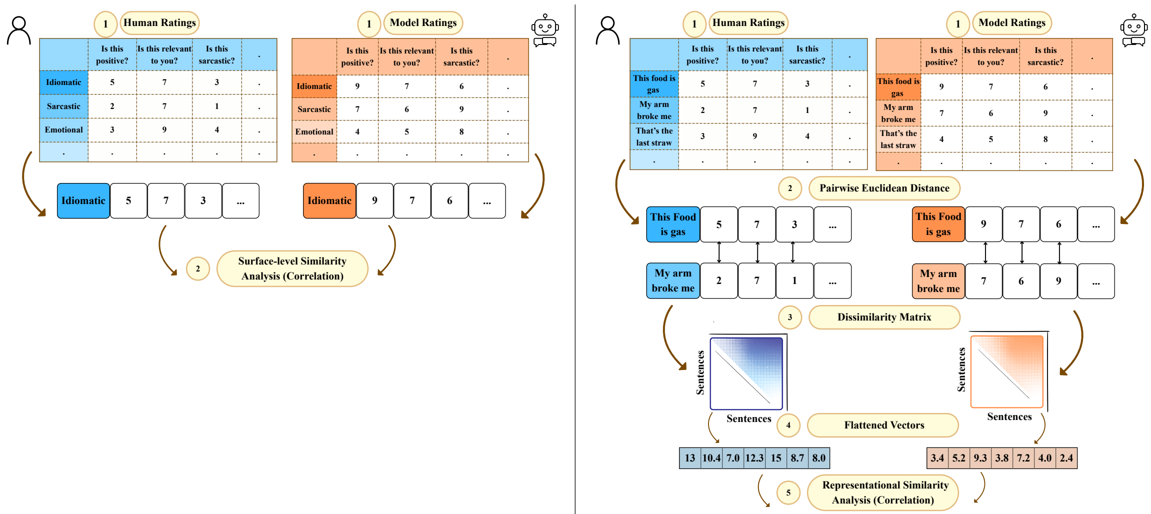
Humans vs. Models To evaluate similarities in language interpretation between humans and models, all sentences were clustered according to their defining linguistic characteristics (conventional, idiomatic, emotional, funny, sarcastic, and Gen Z slang), and both similarity analyses were conducted.

Humans vs. Humans To establish a **baseline** for similarity analysis, human participants were randomly divided into two groups, and both SLS and RSA were computed between these groups. This comparison quantified the internal consistency of human judgments and provided an upper bound for evaluating model–human alignment. Computing these measures separately for each sentence category further allowed us to examine whether human agreement varied across different types of linguistic traits.

3 Results

Figures 4 and 5 illustrate surface-level similarity and Representational similarity results, comparing human judgments with LLM's judgments and within human judgments across sentence categories for sensible and non-sensible sentences in Studies 1 and 2.

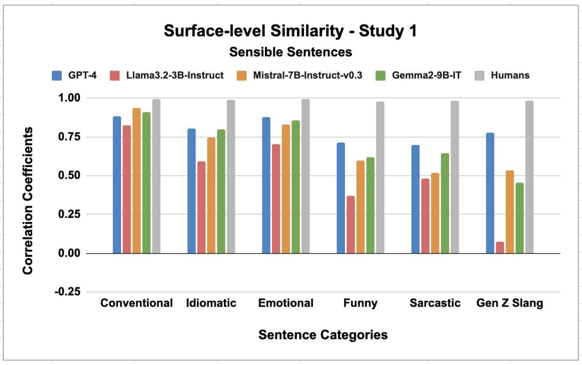
Surface Similarity Analysis. Across both studies, GPT shows the strongest and most consistent alignment with human judgments across sentence categories. For sensible sentences, GPT–human correlations typically range from r = .60–.90 (Figures 4a, 5a), while for non-sensible sentences correlations remain lower but substantial (r = .55–.75) (Figures 4c, 5c). Gemma and Mistral exhibit moderate alignment with humans, with correlations generally ranging from r = .50–.80 depending on sentence type. In contrast, Llama consistently shows the weakest alignment, particularly for figurative categories such as humor, sarcasm, and Gen-Z slang, where correlations often fall below r = .40. Across all models, conventional and emotional sentences yield higher surface-level correlations than humor, sarcasm, and slang. Correlations are generally higher in Study 2 than in Study 1, indicating improved agreement between models and humans under revised prompting. However, non-sensible sentences remain challenging for all models. Corre-



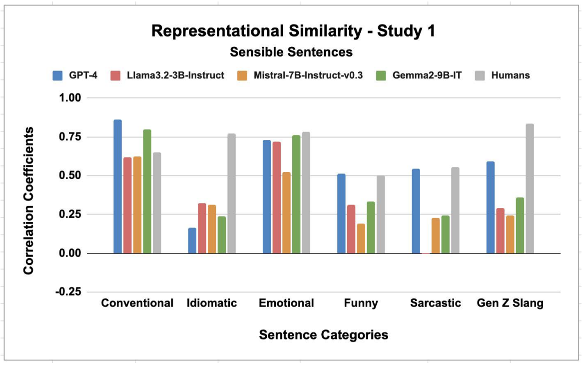
(a) Surface-level similarity analysis

(b) Representational similarity analysis

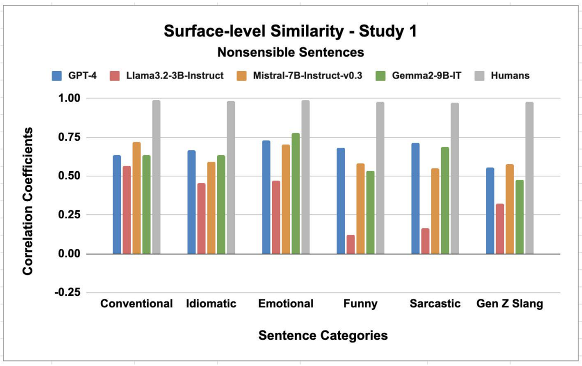
Figure 3: Human participants and LLMs rated all sentences on a 10-point Likert scale. In (a), ratings were aggregated by sentence category (e.g., idiomatic sentences) for humans and models, and Pearson correlations were computed between category-level mean ratings across the 40 dimensions (questions)(1×40) to assess surface-level alignment. In (b), sentence-level ratings across all categories (e.g., Gen Z slang *"The food is gas"* and idiomatic *"My arm broke me"*) were used to compute pairwise Euclidean distances across 40-dimensional rating vectors, yielding a 240×240 representational dissimilarity matrix. Representational alignment between humans and models was assessed by correlating the upper-triangular entries of their matrices, with correlations reported separately for each sentence category.



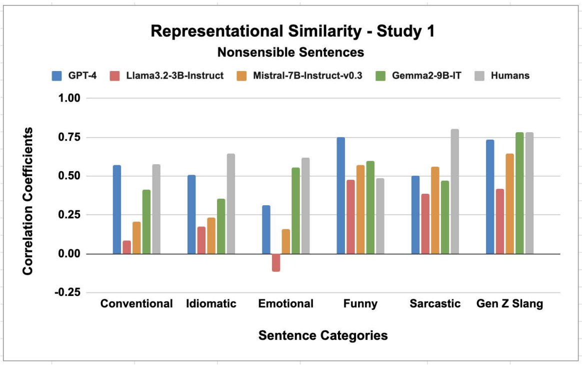
(a) SLS — Sensible Sentences



(b) RSA — Sensible Sentences



(c) SLS — Non-sensible Sentences



(d) RSA — Non-sensible Sentences

Figure 4: The Surface-level similarity (a, c) and Representational similarity (b, d) analyses among humans and models for sensible and non-sensible sentences across Study 1.

lations between human groups remain high across all categories (r = .95–1.00), indicating an upper bound for model performance.

Representational Similarity Analysis. For sensible sentences, RSA reveals moderate to high alignment between human and model representations across both studies. In Study 1 (Figure 4b), GPT exhibits the strongest representational similarity with humans, with correlations ranging approximately from r = .50–.85 across categories, and the highest values observed for conventional and emotional sentences. Gemma and Mistral show moderate representational alignment (r = .55–.80), again performing better on conventional and emotional sentences than on humor, sarcasm, or Gen-Z slang. Llama demonstrates consistently low representational alignment with humans (r = .10–.35), particularly for figurative categories like idiomacy, sarcasm, and Gen Z slang.

Across both studies, representational similarity analyses reveal clear prompt sensitivity and model-wise differences. In Study 2, representational alignment for sensible sentences increases overall, with GPT maintaining high alignment with humans (r = .60–.85) and Gemma and Mistral showing notable improvements (r = .60–.80), though figurative categories consistently lag behind conventional and emotional sentences. Human–human RSA remains high and stable across studies (r = .75–.85), indicating shared representational structure. For non-sensible sentences, model–human similarity is lower and more variable: GPT again shows the strongest alignment (r = .45–.75), followed by Gemma and Mistral (r = .40–.60), while Llama exhibits weak or near-zero correlations. Although Study 2 shows improvements for sarcasm and Gen-Z slang (r = .55–.75), model–human alignment remains consistently below human–human similarity. Overall, a stable hierarchy emerges, with GPT most closely aligned to human representations for all categories of sentences, followed by Gemma and Mistral, and Llama showing limited alignment.

4 Discussion

At the surface level, LLMs, particularly GPT-4 approximated human judgments for all categories of sentences, indicating that current models can capture coarse regularities in how language is evaluated. However, the persistent gap between model–human and human–human agreement for idiomatic and Gen Z slang suggests that surface-level alignment is inherently fragile. This implies that output-level agreement alone is insufficient as evidence of human-like processing. These results support prior critiques that behavioral similarity can emerge from distributional pattern matching rather than stable semantic grounding (Cuskley et al., 2024; Durt et al., 2023).

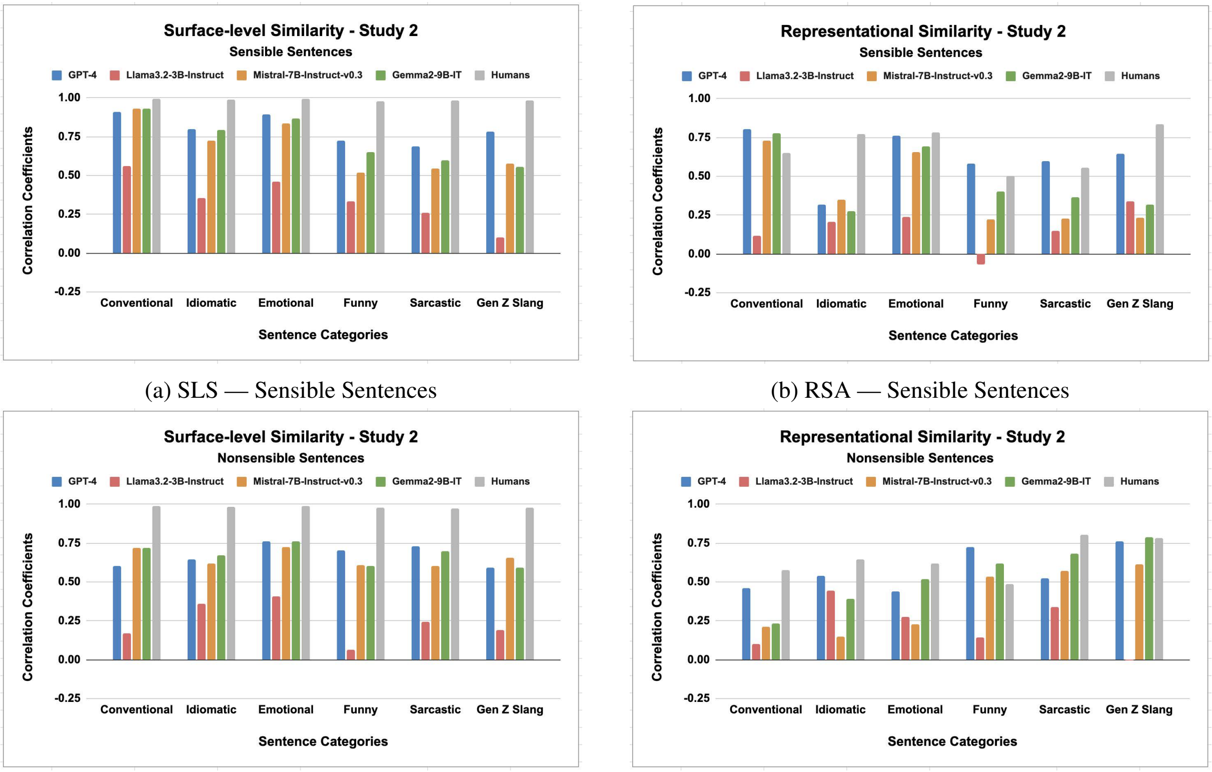
Despite moderate to high correlations in Surface-level similarity, Representational Similarity Analysis revealed systematic divergence between model and human representational spaces. This indicates that current training methods, next-token predictions, do not sufficiently constrain the internal organization of meaning, even when output behavior appears human-like.

The contrast between emotional and idiomatic language further sharpens this result. Emotional sentences show relatively stronger representational alignment, likely because emotional meaning is often conveyed through explicit lexical and affective cues. In contrast, idiomatic expressions require non-compositional reinterpretation, where meaning cannot be derived directly from surface form. The weaker RSA alignment observed for idioms suggests that model representations privilege literal form meaning over pragmatically inferred meaning, consistent with prior observations of limited pragmatic abstraction in LLMs (Karanikolas et al., 2023; Dentella et al., 2024).

Comparisons across models reveal a consistent hierarchy, with GPT-4 exhibiting stronger alignment with humans than smaller models at both surface and representational levels. GPT-4 fails to reach human–human representational consistency, indicating that improved alignment does not imply convergence toward human-like semantic organization. This suggests that LLM's internal representation reflects partial approximation with humans rather than shared structure.

The weaker representational alignment observed for smaller models, such as Llama and Mistral, is especially pronounced for pragmatically rich categories like idiomatic, sarcastic, and Gen Z slang. This pattern suggests limitations in the models' ability to encode context-sensitive cues within a single representational space. Rather than attributing these differences solely to scale, current architectures do not enforce stable clustering of pragmatic meaning, even when surface-level behavior improves.

Although revised prompting yielded modest gains in human–model alignment, representa-



(a) SLS — Sensible Sentences

(b) RSA — Sensible Sentences

(c) SLS — Non-sensible Sentences

(d) RSA — Non-sensible Sentences

Figure 5: The Surface-level similarity (a, c) and Representational Similarity (b, d) analyses among humans and models for sensible and non-sensible sentences across Study 2.

tional similarity remained consistently below human–human baselines, indicating that prompt engineering can influence observable behavior without substantially restructuring internal semantic organization, an effect that was especially pronounced for the Llama model. Given that human interpretation of sarcasm and humor depends on discourse context, speaker intent, and situational knowledge (Hu et al., 2023; Shani et al., 2025), these results suggest that missing contextual input alone does not fully explain representational divergence. Instead, these results suggest that context-sensitive supervision is required to induce stable pragmatic representations aligned with human interpretation, as models otherwise rely on surface heuristics that generalize poorly across pragmatic contexts.

Consistent with findings that LLMs align with human judgments on core grammatical constructions (Hu et al., 2024), our results suggest that such alignment extends to surface-level interpretive structure, but becomes less stable when evaluation depends on deeper representational and pragmatic similarity rather than explicit grammatical form. By explicitly contrasting behavioral alignment with internal semantic structure, this work underscores the need for evaluating frameworks that probe how meaning is organized, not just what output is generated (Cuskley et al., 2024; Poliak et al., 2025).

4.1 Conclusion

Using Representational Similarity Analysis, this study shows that current large language models encode figurative language like idiomacy and Gen Z slang in representational spaces that are only partially aligned with human semantic organization. The models fell short of human–human reliability, with alignment highest for lexically driven categories like emotional and conventional, and markedly weaker for pragmatically rich expressions like sarcasm and funny that require contextual and social inference. The results suggest that limitations in LLM performance arise not merely from output variability or prompting sensitivity, but from differences in how linguistic meaning is internally structured, underscoring the need for training regimes that support context-sensitive and socially grounded semantic representations.

4.2 Limitations

Notable limitations are: First, the dataset size of 240 sentences paired with 40 questions was relatively small, which may limit the generalizability of conclusions about the language behavior of both humans and models. In addition, the sentences were presented without contextual cues, which are especially important for pragmatic phenomena such as humor and sarcasm, and this absence of context may have contributed to the observed differences. Second, the study relied on Likert-scale ratings to capture language interpretation, which necessarily reduced complex and nuanced judgments to single numerical values and may have obscured differences in the underlying reasoning processes of humans and models. Finally, the analysis implicitly assumed that humans and LLMs share comparable internal representational frameworks for interpreting language; however, this assumption may not hold in practice. Importantly, these limitations do not undermine the core comparison between surface-level and representational similarity, as both humans and models were evaluated under identical context-free conditions; instead, they delimit the scope of the claims to context-independent pragmatic interpretation.

4.3 Ethical Considerations

This study was approved by the Institutional Review Board (IRB), Protocol Number IRB2022-1126. The ethical standards in the Declaration of Helsinki for research with human subjects were followed. Informed consent was obtained from all participants before the study. Data were collected from undergraduate participants recruited for course credit. Demographic information was collected in anonymized form and used only for aggregate analysis.

References

S. N. Akter, Z. Yu, A. Muhamed, T. Ou, A. Bäuerle, Á. A. Cabrera, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *Preprint*, arXiv:2312.11444.

Ali Alsajri, Haider A. Salman, and Ahmed Steiti. 2024. Generative models in natural language processing: A comparative study of chatgpt and gemini. *Babylonian Journal of Artificial Intelligence*, pages 134–145.

N. Atox and M. Clark. 2024. Evaluating large language models through the lens of linguistic proficiency and world knowledge: A comparative study. Authorea Preprints.

Saltanat Bayzayeva. 2025. The science of language: Exploring the core of linguistics. *Problems and Solutions of Scientific and Innovative Research*, 2(6):32–35.

Luka Bojić, Petar Kovačević, and Miljan Čabarkapa. 2025. Does GPT-4 surpass human performance in linguistic pragmatics? *Humanities and Social Sciences Communications*, 12(1):1–10.

Zhaoyang Cai, Xinyu Duan, Daniel Haslett, Shuai Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56.

Christine Cuskley, Robert Woods, and Megan Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Valentina Dentella, Fritz Günther, Elliot Murphy, and 1 others. 2024. Testing ai on language comprehension tasks reveals insensitivity to underlying meaning. *Scientific Reports*, 14:28083.

Christoph Durt, Tom Froese, and Thomas Fuchs. 2023. Large language models and the patterns of human language use: An alternative view of the relation of ai to understanding and sentience.

Maria Fedzechkina, Elena Gualdoni, Scott Williamson, Kimberly Metcalf, Serena Seto, and Barry J. Theobald. 2025. Analyze the neurons, not the embeddings: Understanding when and where llm representations align with humans. *Preprint*, arXiv:2502.15090.

Thomas Giallanza and D. Ian Campbell. 2024. Context-sensitive semantic reasoning in large language models. In *ICLR 2024 Workshop on Representational Alignment*.

Jennifer Hu, Samuel 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.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Nikolaos Karanikolas, Effrosyni Manga, Nefeli Samaridi, Efthymia Tousidou, and Michael Vassilakopoulos. 2023. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290.

Nikolaus Kriegeskorte and Rogier A. Kievit. 2013. Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17:401–412.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

Michael Ogg, Rohan Bose, Jordan Scharf, Christopher Ratto, and Michael Wolmetz. 2024. Turing representational similarity analysis (rsa): A flexible method for measuring alignment between human and artificial intelligence. *Preprint*, arXiv:2412.00577.

Adam Poliak, Shiri Malik-Moraleda, and Edward Gibson. 2025. Rational language comprehension depends on priors about both meaning and structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Augustin Preda. 2012. Psycholinguistics and language processing. *Studia Universitatis Babes-Bolyai Psychologia-Paedagogia*, 57(2):125–132.

Hugh Rabagliati, Amy Robertson, and David Carmel. 2018. The importance of awareness for understanding language. *Journal of Experimental Psychology: General*, 147(2):190.

Chen Shani, Lior Soffer, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. 2025. From tokens to thoughts: How llms and humans trade compression for meaning. *Preprint*, arXiv:2505.17117.

Takashi Yamauchi and Kunxia Wang. 2025. Reading moods by mouse-cursor tracking: Representational similarity analysis. *IEEE Transactions on Affective Computing*, (01):1–8.

Fengying Ye, Shanshan Wang, Lidia S Chao, and Derek F Wong. 2025. Unveiling llms' metaphorical understanding: Exploring conceptual irrelevance, context leveraging and syntactic influence. *arXiv preprint arXiv:2510.04120*.

Supplementary Materials

Model	Parameters	Primary Training Objective
GPT-4	Not publicly disclosed	General-purpose language understanding and generation, including reasoning, dialogue, and instruction following
Llama-3.2-3B-Instruct	3 billion	Instruction-tuned language generation with an emphasis on efficiency and conversational alignment
Gemma-2-9B-IT	9 billion	Instruction-tuned model optimized for safe, high-quality text generation and reasoning
Mistral-7B-Instruct-v0.3	7 billion	Instruction-following language generation with strong performance on reasoning and dialogue tasks

Table 2: Overview of Large Language Models Used in the Study

Models	Study 1	Study 2
GPT-4	0.74	0.76
Llama-3.2-3B-Instruct	0.31	0.30
Mistral-7B-Instruct-v0.3	0.68	0.69
Gemma-2-9B-IT	0.68	0.71

(a) Overall Surface-level similarity (SLS)

Models	Study 1	Study 2
GPT-4	0.63	0.64
Llama-3.2-3B-Instruct	0.49	0.19
Mistral-7B-Instruct-v0.3	0.49	0.49
Gemma-2-9B-IT	0.61	0.65

(b) Overall Representational Similarity (RSA)

Table 3: Overall surface-level similarity (left) and representational similarity (right) between humans and models across Study 1 and Study 2.

A Complete Sentence Set

Category	Type	Sentence
Conventional	Sensible	The dog is outside.
Conventional	Sensible	I opened the window for some fresh air.
Conventional	Sensible	My sister is playing the piano.
Conventional	Sensible	The fork is in the drawer.
Conventional	Sensible	I made coffee this morning.
Conventional	Sensible	She had a sandwich for lunch.
Conventional	Sensible	The cat drank the water.
Conventional	Sensible	He ironed his shirt.
Conventional	Sensible	Ally went to the mall.
Conventional	Sensible	I broke my arm.
Conventional	Sensible	He bought a computer at the store.
Conventional	Sensible	I microwaved a bowl of soup.
Conventional	Sensible	He folded the laundry this morning.
Conventional	Sensible	I adopted a dog today.
Conventional	Sensible	I washed the dishes after dinner.
Conventional	Sensible	The kids built a sandcastle on the beach.
Conventional	Sensible	There are 32 kids in the class.
Conventional	Sensible	The chef prepared a delicious meal.
Conventional	Sensible	The sun rises in the east and sets in the west.
Conventional	Sensible	Plants need sunlight for photosynthesis.
Idiomatic	Sensible	Do not cry over spilled milk.
Idiomatic	Sensible	Break a leg today!
Idiomatic	Sensible	I'll do that when pigs fly!
Idiomatic	Sensible	Just bite the bullet.
Idiomatic	Sensible	Tell her the truth, don't beat around the bush.
Idiomatic	Sensible	I was going to go to class, but it's raining cats and dogs.
Idiomatic	Sensible	She's going through a lot, cut her some slack.
Idiomatic	Sensible	That's the last straw.
Idiomatic	Sensible	I'm feeling under the weather.
Idiomatic	Sensible	Now we're back to square one.
Idiomatic	Sensible	The project is not rocket science.
Idiomatic	Sensible	It's time to hit the sack.
Idiomatic	Sensible	I think he's pulling my leg.
Idiomatic	Sensible	You're barking up the wrong tree.
Idiomatic	Sensible	This is a wild goose chase.
Idiomatic	Sensible	The homework is a piece of cake.
Idiomatic	Sensible	You can kill two birds with one stone.
Idiomatic	Sensible	Take it with a grain of salt.
Idiomatic	Sensible	The devil is in the details.
Idiomatic	Sensible	She's burning bridges.
Emotional	Sensible	Sam is so enthusiastic that he jumps out of the bed and begins to cheer.
Emotional	Sensible	Filled with pride, Emily beamed as she watched her child's first steps.
Emotional	Sensible	Bursting with excitement, Alex's heart raced when he opened his first college acceptance letter.

Continued on next page

Category	Type	Sentence
Emotional	Sensible	Susan cried for Lisa when she heard that her dog passed away.
Emotional	Sensible	She sobbed when she heard the news about her aunt.
Emotional	Sensible	Her jaw dropped when she saw who was at the door.
Emotional	Sensible	Max's face turned red and he clenched his fist.
Emotional	Sensible	Emily took a deep breath to calm her nerves before going on stage
Emotional	Sensible	Ava paced back and forth waiting for her exam results.
Emotional	Sensible	Max eagerly unwrapped a mysterious gift.
Emotional	Sensible	Sarah smiled warmly as she revisited cherished childhood photos
Emotional	Sensible	David sighed in frustration during a traffic jam.
Emotional	Sensible	Alex watched as the long-awaited event was canceled.
Emotional	Sensible	Alex's heart fluttered, anticipating his partner's arrival.
Emotional	Sensible	Mark teared up, moved by a heartfelt movie.
Emotional	Sensible	Jake twirled with joy at news of his dream job.
Emotional	Sensible	Mike angrily crumpled up his homework after getting a bad grade.
Emotional	Sensible	Sally smiled and shed a tear at her brother's graduation.
Emotional	Sensible	Sofia blushed and giggled when he told her a joke.
Emotional	Sensible	Her face turned white with shame when he read her diary.
Funny	Sensible	If money does not grow on trees, why do banks have branches?
Funny	Sensible	I used to be a baker because I kneaded dough.
Funny	Sensible	Parallel lines have so much in common, it's a shame they'll never meet.
Funny	Sensible	An apple a day keeps the doctor away, if you throw it hard enough.
Funny	Sensible	I sold my vacuum, all it was doing was gathering dust.
Funny	Sensible	Life is a bowl of soup, and I'm a fork.
Funny	Sensible	You're not adopted, but we've placed an ad.
Funny	Sensible	My jeans says eat a salad, but my heart says eat a pizza.
Funny	Sensible	My brain has too many tabs open.
Funny	Sensible	I have a food baby.
Funny	Sensible	Everyone has a right to be stupid, but some abuse that privilege.
Funny	Sensible	I'm on a seafood diet; I see food and I eat it.
Funny	Sensible	My doctor diagnosed me with too smart syndrome.
Funny	Sensible	I'm jealous of my parents because I'll never have a kid as cool as theirs.
Funny	Sensible	I get enough exercise pushing my luck.
Funny	Sensible	The first five days after the weekend are the hardest.
Funny	Sensible	The road to success is always under construction.
Funny	Sensible	Finally, spring is here! I am so thrilled I wet my plants.
Funny	Sensible	I'm never late, others are just too early.
Funny	Sensible	Hey dude, pick your brain, sometimes.
Sarcastic	Sensible	I love your shirt, for now.
Sarcastic	Sensible	Nice perfume. How long did you marinate in it?
Sarcastic	Sensible	You find your patience before I lose mine.

Continued on next page

Category	Type	Sentence
Sarcastic	Sensible	Light travels faster than sound. This is why some people appear bright until they speak.
Sarcastic	Sensible	If you find me offensive. Then I suggest you quit finding me.
Sarcastic	Sensible	Unless your name is Google, stop acting like you know everything.
Sarcastic	Sensible	I do not have the energy to pretend to like you today.
Sarcastic	Sensible	I love your shoes. Great shoes. What a surprise!
Sarcastic	Sensible	I envy your commitment to being fashionably late.
Sarcastic	Sensible	I'm so glad you posted another photo of your lunch. I was on the edge of my seat wondering what you ate today.
Sarcastic	Sensible	It's truly impressive how you make the easiest tasks look incredibly complicated.
Sarcastic	Sensible	Thank you for the unsolicited advice; it's exactly what I wanted to hear.
Sarcastic	Sensible	I love your new hair, I never thought I'd see chunky highlights in this decade.
Sarcastic	Sensible	I failed my exam, that's exactly what I needed this week!
Sarcastic	Sensible	I love working 40 hours a week for minimum wage!
Sarcastic	Sensible	I would agree with you, but then we would both be wrong.
Sarcastic	Sensible	I don't know why I have a headache; all I've done is forget to drink water and stare at a screen all day.
Sarcastic	Sensible	I am busy right now, can I ignore you some other time?
Sarcastic	Sensible	Life is good, you should get one.
Sarcastic	Sensible	If had a dollar for every smart thing you say, I would be poor.
Gen Z slang	Sensible	I got an A in beer pong.
Gen Z slang	Sensible	Have a seat. You can just move that laundry pile.
Gen Z slang	Sensible	She is so boujee with that Louis Vuitton bag.
Gen Z slang	Sensible	Those potato chips are bussin.
Gen Z slang	Sensible	He got so salty after I did not text back right away.
Gen Z slang	Sensible	Her speech was woke.
Gen Z slang	Sensible	I am so amped for tonight's basketball game!
Gen Z slang	Sensible	She will be OK after she blows off some steam.
Gen Z slang	Sensible	I bombed that exam.
Gen Z slang	Sensible	I heard the Greek life on campus is pretty fun.
Gen Z slang	Sensible	Wake up!. It's time to hit the books.
Gen Z slang	Sensible	Where is your girlfriend? I dunno.
Gen Z slang	Sensible	The movie was mid.
Gen Z slang	Sensible	That 50% off sale at the campus can't be legit.
Gen Z slang	Sensible	I'm so bored, I need to hear some good tea.
Gen Z slang	Sensible	I'm beefing with my roommate.
Gen Z slang	Sensible	Why is your friend on this date with us? He is kind of a third wheel.
Gen Z slang	Sensible	This food is gas.
Gen Z slang	Sensible	I was gonna stay in tonight, but I don't wanna get FOMO.
Gen Z slang	Sensible	Hey Bro. You talk too much. Keep your shirt on.
Conventional	Non-sensible	The outside is the dog.
Conventional	Non-sensible	The fresh air opened the window.

Continued on next page

Category	Type	Sentence
Conventional	Non-sensible	The piano is playing my sister.
Conventional	Non-sensible	The drawer is in the fork.
Conventional	Non-sensible	I made morning this coffee.
Conventional	Non-sensible	He had a cat for lunch.
Conventional	Non-sensible	The cat drank the shirt.
Conventional	Non-sensible	His shirt ironed him.
Conventional	Non-sensible	The mall went to Ally.
Conventional	Non-sensible	My arm broke me.
Conventional	Non-sensible	He bought a store at the computer.
Conventional	Non-sensible	I microwaved a soup of bowl.
Conventional	Non-sensible	The laundry folded him this morning.
Conventional	Non-sensible	The dog adopted me today.
Conventional	Non-sensible	After dinner, the dishes washed me.
Conventional	Non-sensible	On the beach, the sandcastle built the kids.
Conventional	Non-sensible	There are 32 classes in the kids.
Conventional	Non-sensible	A delicious meal prepared the chef.
Conventional	Non-sensible	The east rises in the sun and the west sets in the moon.
Conventional	Non-sensible	Sunlight needs plants for photosynthesis.
Idiomatic	Non-sensible	Don't cry over an orange juice spill.
Idiomatic	Non-sensible	Break your arm today.
Idiomatic	Non-sensible	I'll do that when the pigs dance.
Idiomatic	Non-sensible	Eat the bullet.
Idiomatic	Non-sensible	Tell the truth and don't beat up the tree.
Idiomatic	Non-sensible	I was going to go to class, but it's raining fish and monkeys.
Idiomatic	Non-sensible	She's going through the slack, cut her a lot.
Idiomatic	Non-sensible	The straw is last.
Idiomatic	Non-sensible	The weather is over me.
Idiomatic	Non-sensible	Square one is back to me.
Idiomatic	Non-sensible	The rocket science is not the project.
Idiomatic	Non-sensible	It's time for the sack to hit me.
Idiomatic	Non-sensible	I think my leg is pulling him.
Idiomatic	Non-sensible	You're barking up the right flower.
Idiomatic	Non-sensible	This run for ducks is wild.
Idiomatic	Non-sensible	A piece of cake is the homework.
Idiomatic	Non-sensible	You can murder two chickens with one rock.
Idiomatic	Non-sensible	Take a grain of rice with it.
Idiomatic	Non-sensible	The details are in the devil.
Idiomatic	Non-sensible	The bridges burned her.
Emotional	Non-sensible	Sam is so sad that he jumps and cheers.
Emotional	Non-sensible	Filled with anger, Emily beamed as her child took her first steps.
Emotional	Non-sensible	Sadness burst Alex when he opened his college letters.
Emotional	Non-sensible	Susan weeped gleefully for Lisa when she heard that her dog passed away.
Emotional	Non-sensible	She heard the news about her aunt when she sobbed.
Emotional	Non-sensible	Her jaw clenched when she saw who was at the door.
Emotional	Non-sensible	Max's face turned white and he tighened his fist.
Emotional	Non-sensible	Emily took a deep breath to stir her nerves before going on stage

Continued on next page

Category	Type	Sentence
Emotional	Non-sensible	Ava ran around the exam results.
Emotional	Non-sensible	Max quietly wrapped a mysterious gift.
Emotional	Non-sensible	Sarah cried as she revisited cherished childhood photos
Emotional	Non-sensible	David smiled in frustration during a traffic jam.
Emotional	Non-sensible	Alex canceled the long-awaited event he watched.
Emotional	Non-sensible	Alex's heart sank, anticipating his partner's arrival.
Emotional	Non-sensible	The movie moved Mark and teared up.
Emotional	Non-sensible	Jake scratched his head at news of his dream job.
Emotional	Non-sensible	Mike crumpled up his homework after getting a nasty grade.
Emotional	Non-sensible	Sally scowled and shed a tea at her brother's graduation.
Emotional	Non-sensible	Sofia turned red when he told her a joke.
Emotional	Non-sensible	Her face turned down when he read her diary.
Funny	Non-sensible	If branches don't grow on banks, why do trees have money?
Funny	Non-sensible	I used to be a baker, and had to knead dough.
Funny	Non-sensible	Parallel lines will never meet.
Funny	Non-sensible	If you throw it at the doctor, the watermelon keeps the doctor away.
Funny	Non-sensible	I sold my vacuum that used to gather dust.
Funny	Non-sensible	I'm a bowl of soup with a fork.
Funny	Non-sensible	You're not adopted, but we've seen an ad.
Funny	Non-sensible	My jeans says buy a salad, my heart says sell a pizza
Funny	Non-sensible	My computer has too many tabs open.
Funny	Non-sensible	The baby is food.
Funny	Non-sensible	Everyone has a right to be stupid, but some are really stupid.
Funny	Non-sensible	I'm on a seafood diet; I eat fish every day
Funny	Non-sensible	My doctor diagnosed me with too nasty syndrome.
Funny	Non-sensible	I envy my parents because they have a cool car.
Funny	Non-sensible	I get enough exercise pushing my lucky chair.
Funny	Non-sensible	The first five days before the weekend are the hardest.
Funny	Non-sensible	The road to success is always under consideration
Funny	Non-sensible	Finally, spring is here! I am so thrilled I watered my plants.
Funny	Non-sensible	I'm never late, others are always late.
Funny	Non-sensible	Dude, sometimes your brain picks you.
Sarcastic	Non-sensible	I love your shirt.
Sarcastic	Non-sensible	Nice perfume. How long have you had that for?
Sarcastic	Non-sensible	The patience got lost and couldn't be found.
Sarcastic	Non-sensible	Light travels faster than sound. Some people appear bright until they smile.
Sarcastic	Non-sensible	If you find me offensive. Then I suggest you quit smoking.
Sarcastic	Non-sensible	Unless your name is Alphabet, stop acting like you know everything.
Sarcastic	Non-sensible	I do not like you today.
Sarcastic	Non-sensible	I love your shoes. Great shoes.
Sarcastic	Non-sensible	I envy your commitment to being fashionable.
Sarcastic	Non-sensible	I'm so glad you posted a photo of your lunch. I was wondering what you ate today.
Sarcastic	Non-sensible	It's truly impressive how you make the incredibly complicated tasks look easy.
Sarcastic	Non-sensible	Thank you for the advice; it's exactly what I needed.

Continued on next page

Category	Type	Sentence
Sarcastic	Non-sensible	I love your new hair, I never thought I'd see beautiful hair like that.
Sarcastic	Non-sensible	I needed to fail my exam this week.
Sarcastic	Non-sensible	I hate working 40 hours a week for minimum wage.
Sarcastic	Non-sensible	I would agree with you, we would both be right.
Sarcastic	Non-sensible	I don't know why I have a headache; all I've done is to drink a lot of water.
Sarcastic	Non-sensible	I'm ignoring right now, can I be busy some other time?
Sarcastic	Non-sensible	Life's one, you should get good.
Sarcastic	Non-sensible	If had a dollar for every smart thing you say, I would be rich.
Gen Z slang	Non-sensible	I got an F in Economics 101.
Gen Z slang	Non-sensible	Have a laundry pile, you can just move that seat.
Gen Z slang	Non-sensible	She is so boujee with that Wal Mart bag.
Gen Z slang	Non-sensible	Those bussin' are potato chips.
Gen Z slang	Non-sensible	He bought salt after I didn't text back.
Gen Z slang	Non-sensible	She woke her speech.
Gen Z slang	Non-sensible	Tonight's basketball game was exciting.
Gen Z slang	Non-sensible	She will be OK after she eats steamed rice.
Gen Z slang	Non-sensible	That exam bombed me.
Gen Z slang	Non-sensible	I heard the campus on greek life is pretty fun.
Gen Z slang	Non-sensible	Wake up! It's time to smack the books.
Gen Z slang	Non-sensible	Where is her boyfriend? No clue
Gen Z slang	Non-sensible	The mid was movie.
Gen Z slang	Non-sensible	That 50% off sale at the campus is great.
Gen Z slang	Non-sensible	I need tea to be bored.
Gen Z slang	Non-sensible	My roommate is with beef.
Gen Z slang	Non-sensible	Why is your friend on this date with us? He's kind of a fourth tire.
Gen Z slang	Non-sensible	The gas is food.
Gen Z slang	Non-sensible	I don't wanna eat FOMO.
Gen Z slang	Non-sensible	Hey Bro. You smile too much. Keep your pants on.

Table 4: Complete list of all 240 sentences used in the study for the judgments from human participants and LLMs, organized by category and sentence type.

B Interpretive Questions

Questions
Is this meaningful?
Do you find this funny?
Is this surprising?
Is this relevant (to you)?
Is this grammatically correct?
Is this insulting?
Do you find this exciting?
Is this sentence dull?
Is this sarcastic?
Is this honest?
Does this sentence show attitude?
Is this ironic?
Is this objective?
Is this mocking you?
Is this frustrating?
Does this sound serious?
Is this literal?
Do you find this stupid?
Is this convincing?
Is this active?
Do you think this is dramatic?
Is this arrogant?
Is this sentimental?
Is this polite?
Is this demeaning to you?
Is this positive?
Is this informative?
Is this logical?
Is this intriguing?
Does this sound confident?
Is this ambiguous?
Is this concerning?
Is this provocative?
Does this sound like praise?
Does this sound loud?
Does this sound negative?
Is this sentence weird?
Do you find this suspicious?
Do you think this is melodramatic?
Does this sound sympathetic?

Table 5: List of 40 interpretive questions used to collect judgments from human participants and LLMs.