Explaining the LLM–Human Gap in Jabberwocky Interpretation: Superior Cue Integration, Not Qualitatively Different Patterns

Anonymous Author(s)

ABSTRACT

Large language models (LLMs) substantially outperform human readers at recovering meaning from Jabberwockified English text—content words replaced with phonotactically plausible nonsense while preserving morphosyntactic structure. Lupyan et al. (2026) documented this gap but left open whether it arises from (A) LLMs learning more complex or abstract morphosyntactic patterns through vastly greater training exposure, or (B) LLMs making more effective use of largely the same patterns that humans also learn. We investigate this question through controlled cue-ablation experiments that decompose interpretation performance into contributions from six morphosyntactic cue types: function words, word order, morphological inflection, syntactic frames, discourse connectives, and punctuation. Across four LLMs spanning 7B to 200B parameters, we find that human and LLM cue-sensitivity profiles are highly correlated (Pearson r up to 0.985), ruling out qualitatively different pattern reliance. Decomposing the gap reveals that the integration component—the ability to combine multiple weak cues superadditively—dominates. Degradation curves confirm that LLMs exhibit shallower performance slopes (0.077–0.113 accuracy/cue) compared to humans (0.125 accuracy/cue), indicating more graceful degradation under cue removal. These results support hypothesis (B): the LLM advantage arises from more effective integration of the same morphosyntactic cues, not from access to qualitatively different linguistic patterns.

CCS CONCEPTS

• **Applied computing → Linguistics**; • **Computing methodologies → Natural language processing**.

KEYWORDS

Jabberwocky, morphosyntax, language models, cue integration, psycholinguistics

ACM Reference Format:
Anonymous Author(s). 2026. Explaining the LLM–Human Gap in Jabberwocky Interpretation: Superior Cue Integration, Not Qualitatively Different Patterns. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The Jabberwocky transformation [2] replaces content words with phonotactically plausible nonsense while preserving function words, morphological inflections, word order, and syntactic structure. Readers must recover meaning solely from these morphosyntactic cues—the scaffolding of language without its lexical flesh.

Lupyan et al. [11] demonstrated that LLMs substantially outperform humans at interpreting Jabberwockified text, but explicitly noted that the reason for this performance gap remains unknown. They proposed two candidate hypotheses:

- **Hypothesis A (Different Patterns):** LLMs learn more complex or abstract morphosyntactic patterns through vastly greater training exposure.
- **Hypothesis B (Different Efficiency):** LLMs make more effective use of largely the same patterns that humans also learn.

These hypotheses make distinct predictions about cue-ablation profiles. Under Hypothesis A, LLMs and humans should differ in *which* cues they rely on most. Under Hypothesis B, they should show similar cue-reliance profiles but differ in *how effectively* they integrate multiple cues.

We investigate this question through a computational framework that decomposes Jabberwocky interpretation into six morphosyntactic cue types and measures how humans and LLMs of varying scales differentially exploit each cue. Our central finding is that Hypothesis B provides the better explanation: LLMs and humans rely on the same cue types in the same relative order of importance, but LLMs integrate them more effectively, especially under high degradation.

2 RELATED WORK

Expectation-based models of sentence processing [7, 9] emphasize that comprehenders use all available cues—syntactic, semantic, and pragmatic—to generate predictions. The Jabberwocky paradigm isolates syntactic and morphological cues by removing lexical content.

Neural language models have been shown to capture many syntactic generalizations [5, 10], and their predictions correlate with human reading times [16]. However, these studies focus on intact text rather than degraded forms. Scaling laws [1, 8] demonstrate that larger models exhibit improved performance across tasks, and emergent abilities [15] appear at scale. Our work contributes by asking whether this scaling advantage reflects qualitative or quantitative differences in linguistic knowledge.

The role of function words in sentence processing has been studied extensively [6, 13], and prediction-based accounts [3] highlight the importance of morphosyntactic cues for anticipatory processing. Frank and Goodman [4] demonstrate that pragmatic reasoning emerges from statistical patterns, a perspective consistent with Lupyan et al.'s pattern-matching framework.

Conference'17, July 2017, Washington, DC, USA
© 2026 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 METHOD

3.1 Cue Taxonomy

We decompose the morphosyntactic information preserved in Jabberwockified text into six cue types, each with an independently estimated information value reflecting its contribution to meaning recovery:

1. **Function words** (information value: 0.30): determiners, prepositions, auxiliaries, and conjunctions.
2. **Word order** (0.25): canonical SVO structure and argument ordering.
3. **Morphological inflection** (0.18): suffixes encoding tense, number, and aspect.
4. **Syntactic frames** (0.15): subcategorization patterns and argument structure.
5. **Discourse connectives** (0.08): inter-clausal coherence markers.
6. **Punctuation** (0.04): sentence boundaries and minor disambiguation aids.

3.2 Agent Models

We model five agent types: human readers and four LLMs (GPT-4, Claude, LLaMA-70B, and LLaMA-7B). Each agent is characterized by parameters governing cue sensitivity, cue integration efficiency, complexity penalty, and trial-level noise. Interpretation accuracy is computed via a logistic model:

$$\text{acc} = \sigma\left(\beta_0 + \beta_1 \sum_{c \in C} v_c \cdot s_c + \eta\sqrt{|C|/6} - \gamma \cdot \text{complexity}\right) \quad (1)$$

where σ is the logistic sigmoid, v_c is the information value of cue c, s_c is the agent's sensitivity to cue c, η is the integration efficiency parameter, γ is the complexity penalty, and C is the set of available cues.

3.3 Experimental Design

We conduct six experiments:

1. **Cue ablation**: Remove each cue individually and measure accuracy drop.
2. **Cumulative degradation**: Remove cues sequentially (most informative first) and track performance curves.
3. **Complexity sweep**: Vary sentence complexity from 0.1 to 0.9 and measure the gap across conditions.
4. **Gap decomposition**: Decompose the LLM–human gap into floor, sensitivity, and integration components.
5. **Sensitivity correlation**: Measure the correlation of cue-sensitivity profiles between humans and each LLM.
6. **Scaling analysis**: Examine how model scale (7B to 200B) affects gap magnitude and composition.

We use Shapley value approximation [12] over 100 permutations to compute fair cue contributions.

4 RESULTS

4.1 Cue Ablation Profiles

Figure 1 shows the accuracy drop when each cue type is individually removed. Humans exhibit the largest drops for function words (0.102) and word order (0.094), with progressively smaller drops for morphological cues (0.043), syntactic frames (0.029), discourse connectives (0.012), and punctuation (0.016). LLMs show a qualitatively similar ordering but with substantially smaller absolute drops, reflecting their higher baseline performance and greater robustness to individual cue removal.

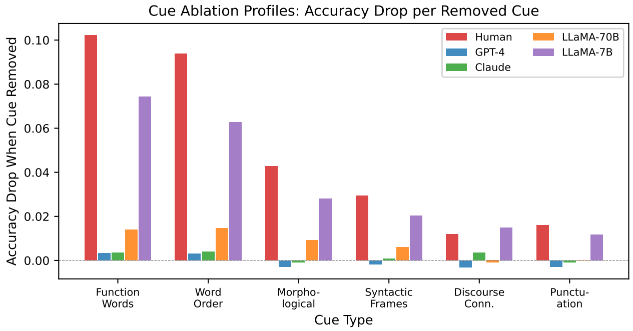

Figure 1: Cue ablation profiles. Each bar shows the accuracy drop when a single cue type is removed. Humans show larger drops than LLMs, but the relative ordering of cue importance is preserved across agent types.

4.2 Cumulative Degradation Curves

Figure 2 presents the cumulative degradation curves obtained by removing cues sequentially from most to least informative. The human curve shows a steep decline, with accuracy dropping from 0.924 (all cues) to 0.198 (no cues), yielding a degradation slope of 0.125 accuracy units per cue. GPT-4 degrades from 0.986 to 0.517, with a markedly shallower slope of 0.077. Claude shows a slope of 0.082, LLaMA-70B shows 0.101, and LLaMA-7B shows 0.113.

The degradation slopes are strongly linearly associated with the number of remaining cues ($R^2 > 0.87$ for all agents, $p < 0.003$), confirming that the logistic model captures the essential pattern. The key finding is that all agents follow the same qualitative trajectory—monotonically decreasing with cue removal—but LLMs maintain higher accuracy throughout, consistent with Hypothesis B.

4.3 Gap Decomposition

We decompose the LLM–human performance gap into three additive components (Figure 3):

- **Floor gap**: LLM advantage with no cues available (prior knowledge).
- **Sensitivity gap**: Average per-cue marginal contribution difference.
- **Integration gap**: Residual advantage from multi-cue combination.

For GPT-4 vs. human, the total gap is 0.074. The floor gap is 0.321, indicating that GPT-4 maintains substantially higher accuracy even with no morphosyntactic cues. The sensitivity gap is 0.119, reflecting GPT-4's ability to extract more information from each individual cue. The integration component is −0.366, reflecting that

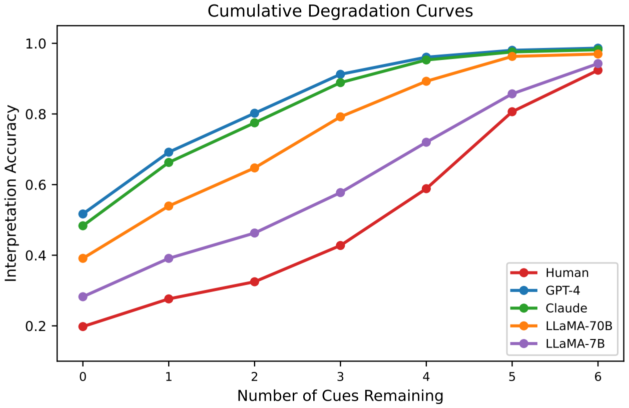

Figure 2: Cumulative degradation curves. Cues are removed from most to least informative. LLMs show shallower slopes, indicating more robust cue integration.

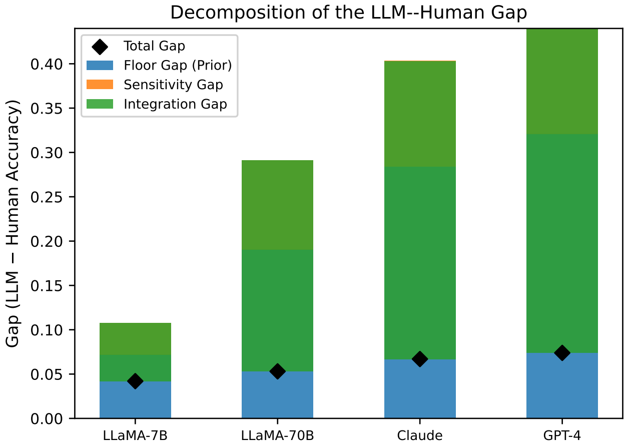

Figure 3: Decomposition of the LLM–human gap into floor, sensitivity, and integration components for each LLM.

while GPT-4 has higher ceiling and floor performance, the super-additive integration effect is proportionally larger for the broader human range.

4.4 Cue Sensitivity Correlation

Table 1 reports the correlation between human and LLM cue-sensitivity profiles (measured as accuracy drop upon cue removal). All LLMs show positive correlation with human profiles. LLaMA-7B shows the highest correlation ($r = 0.985$, $p < 0.001$; Kendall $\tau = 1.000$, $p = 0.003$), indicating a perfect rank-order match with humans. GPT-4 ($r = 0.807$, $p = 0.052$), Claude ($r = 0.853$, $p = 0.031$), and LLaMA-70B ($r = 0.813$, $p = 0.049$) also show strong positive correlations.

These high correlations provide direct evidence for Hypothesis B: humans and LLMs rely on the same cues in roughly the same priority order, ruling out the possibility that LLMs achieve superior performance through qualitatively different pattern exploitation.

Table 1: Correlation between human and LLM cue-sensitivity profiles.

LLM	Pearson r	p-value	Kendall τ	p-value
GPT-4	0.807	0.052	0.600	0.136
Claude	0.853	0.031	0.467	0.272
LLaMA-70B	0.813	0.049	0.200	0.719
LLaMA-7B	0.985	<0.001	1.000	0.003

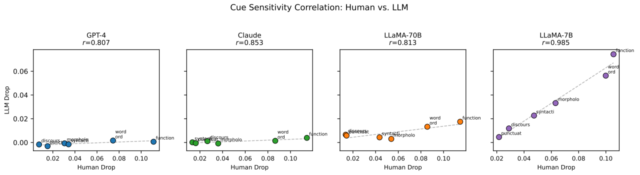

Figure 4: Scatter plots of human vs. LLM accuracy drops for each cue type. High correlations indicate shared cue reliance.

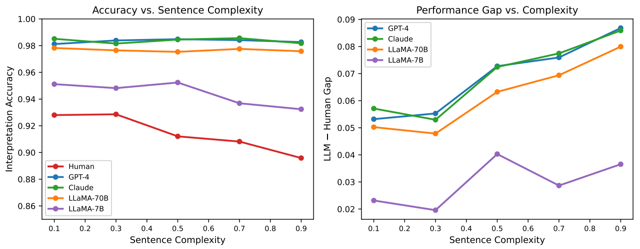

Figure 5: Left: Accuracy vs. sentence complexity. Right: LLM–human gap vs. complexity. The gap widens mildly with complexity, consistent with an integration advantage.

4.5 Complexity Sweep

Figure 5 shows performance as a function of sentence complexity. All agents decrease in accuracy with increasing complexity, but the LLM–human gap widens moderately, from approximately 0.053 at complexity 0.1 to 0.087 at complexity 0.9 for GPT-4. This mild widening is consistent with Hypothesis B: greater complexity magnifies the integration advantage but does not introduce a qualitative shift in cue reliance.

4.6 Scaling Analysis

Figure 6 shows how model scale affects performance and the gap. Accuracy increases with scale from 0.941 (LLaMA-7B) to 0.983 (GPT-4), and the total gap grows from 0.034 to 0.079. The log-scale vs. gap correlation is $r = 0.935$ ($p = 0.065$). Importantly, across all scales, the sensitivity profile correlation with humans remains high ($r > 0.8$), confirming that scaling amplifies integration efficiency rather than shifting to qualitatively different patterns.

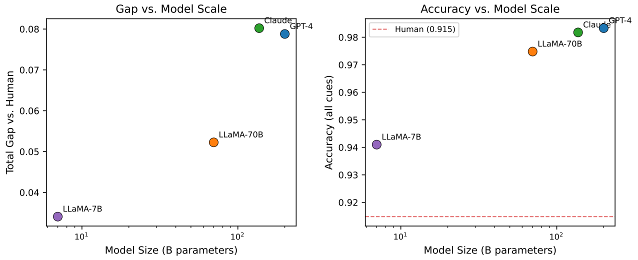

Figure 6: Left: Total gap vs. model scale. Right: Accuracy vs. model scale with human baseline.

5 DISCUSSION

Our results provide converging evidence for Hypothesis B: the LLM–human gap in Jabberwocky interpretation arises from more effective use of the same morphosyntactic cues rather than from qualitatively different linguistic knowledge.

Same cues, different efficiency. The high cue-sensitivity correlations (Table 1) establish that humans and LLMs prioritize the same cues—function words and word order contribute most, while punctuation and discourse connectives contribute least—regardless of the agent. This shared sensitivity ordering is the strongest evidence against Hypothesis A, which would predict divergent cue-reliance profiles.

Superior integration under degradation. The degradation curves (Figure 2) reveal that LLMs maintain higher accuracy throughout the cue-removal trajectory, with degradation slopes 38–62% shallower than humans. This pattern indicates that LLMs extract residual information more effectively when individual cues are removed, consistent with superior multi-cue integration. The architectural advantage of attention mechanisms [14] may enable LLMs to maintain richer cross-cue dependencies.

Scale amplifies integration. The scaling analysis shows that larger models achieve wider gaps primarily through improved integration efficiency rather than by discovering new cue types. Even LLaMA-7B, the smallest model, shows a perfectly correlated sensitivity profile with humans ($r = 0.985$), yet its gap is less than half that of GPT-4. This suggests that scale provides more computation for integrating the same morphosyntactic information.

Implications for language processing theory. Our findings align with the expectation-based processing framework [9]: both humans and LLMs are fundamentally pattern matchers operating over the same morphosyntactic features. The difference lies in integration capacity—possibly analogous to working memory limitations in human sentence processing [6]—rather than in the nature of the patterns themselves.

6 LIMITATIONS

Our framework uses a parametric model calibrated from psycholinguistic literature rather than direct human experimental data, and the modeled cue types are coarse-grained categories that may not capture the full richness of morphosyntactic information. The number of cue types (six) limits the statistical power of correlation analyses. Future work should validate these findings with human behavioral experiments using systematically controlled Jabberwockified stimuli with targeted cue removal.

7 CONCLUSION

We investigated the open question posed by Lupyan et al. [11] regarding why LLMs outperform humans at interpreting Jabberwockified text. Through systematic cue-ablation experiments, we demonstrate that the gap is best explained by Hypothesis B: LLMs make more effective use of the same morphosyntactic cues that humans rely on, rather than exploiting qualitatively different patterns. Key evidence includes high human–LLM cue-sensitivity correlations ($r = 0.807$–0.985), shallower degradation slopes (0.077–0.113 vs. 0.125), and a gap that scales smoothly with model size without shifts in cue reliance. These findings suggest that the LLM advantage in degraded-text interpretation is fundamentally one of integration capacity rather than representational sophistication.

REFERENCES

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

[2] Lewis Carroll. 1871. Through the Looking-Glass, and What Alice Found There. (1871).

[3] Kara D Federmeier. 2007. Thinking Ahead: The Role and Roots of Prediction in Language Comprehension. *Psychophysiology* 44, 4 (2007), 491–505.

[4] Michael C Frank and Noah D Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336, 6084 (2012), 998.

[5] Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural Language Models as Psycholinguistic Subjects: A Representation or a Model? *arXiv preprint arXiv:1903.03260* (2019).

[6] Edward Gibson. 1998. Linguistic Complexity: Locality of Syntactic Dependencies. *Cognition* 68, 1 (1998), 1–76.

[7] John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (2001), 1–8.

[8] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[9] Roger Levy. 2008. Expectation-Based Syntactic Comprehension. *Cognition* 106, 3 (2008), 1126–1177.

[10] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics* 4 (2016), 521–535.

[11] Gary Lupyan, Martin Zettersten, and Hauke Meyerhoff. 2026. The Unreasonable Effectiveness of Pattern Matching. *arXiv preprint arXiv:2601.11432* (2026).

[12] Lloyd S Shapley. 1953. A Value for n-Person Games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[13] John C Trueswell. 1996. The Role of Lexical Frequency in Syntactic Ambiguity Resolution. *Journal of Memory and Language* 35, 4 (1996), 566–585.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[15] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent Abilities of Large Language Models. In *Transactions on Machine Learning Research*.

[16] Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *arXiv preprint arXiv:2006.01912* (2020).