

## (Don't) Take This Literally: Human and LLM Judgments in Atypical Idiom Contexts

Irene Pagliai (irene.pagliai@tu-braunschweig.de)<sup>1</sup>

<sup>1</sup>Department of English Linguistics, TU Braunschweig

Idioms invite us not to take language too literally—except when we must. Although idioms typically convey figurative meanings [4], their literal components can become relevant in two atypical contexts: *literal-compositional contexts*, where an idiomatic string is interpreted compositionally, and *figurative-creative contexts*, where idiomatic wordplay requires a literal continuation to be semantically and pragmatically reinterpreted [5, 10]. Building on hybrid models of idiom processing [6], a recent behavioral experiment on Italian idioms [8] showed that the integration of literal continuations is modulated by the Potential Idiomatic Ambiguity (PIA) index, a composite measure of an idiom's literal plausibility, decomposability, and transparency. Across both context types, high-PIA idioms facilitated the integration of literal continuations more than low-PIA idioms.

The present study asks whether Large Language Models (LLMs) reproduce humans' sensitivity to idiom-specific properties when evaluating literal continuations in these two contexts. We replicate the experiment in [8] using the same 16 Italian idioms, split into high- and low-PIA groups, and the same dialogue-based stimuli (Table 1). Speaker A biased the interpretation toward either a literal or figurative reading; speaker B produced a fixed response containing the idiom and a literal continuation, whose acceptability was rated on a 1–7 scale. We elicited the same ratings from GPT-5 mini and GPT-4.1 mini, with 42 repeated runs for each idiom in each context to match the number of human participants. To explore whether human–model alignment varies across inference settings, GPT-5 mini was tested with low vs. high reasoning effort and GPT-4.1 mini at temperature 0 vs. 0.7. This experiment places LLMs under a double stress test, idiomatic language and linguistic creativity [1, 3, 7], while contributing to debates on LLMs as virtual participants [2, 9].

Item-level Spearman correlations show that human–model alignment was stronger in literal-compositional contexts ( $\rho = .62-.73$ ) than in figurative-creative contexts ( $\rho = .44-.62$ ), where literal continuations required creative semantic-pragmatic reinterpretation. However, correlations captured only part of the pattern. As shown in Figure 1, mean ratings revealed systematic differences in both scale use and PIA sensitivity across sources. A cumulative link mixed model confirmed this pattern, revealing a significant  $\text{PIA} \times \text{context} \times \text{source}$  interaction. In literal-compositional contexts, all models reproduced the human direction of the PIA effect, assigning higher ratings to literal continuations after high- than low-PIA idioms. However, this qualitative alignment concealed model-specific differences: GPT-4.1 mini showed an exaggerated PIA effect, whereas GPT-5 mini showed an optimism bias, assigning overall higher ratings than humans. In figurative-creative contexts, PIA effects weakened and the cross-source pattern changed: humans showed the strongest sensitivity to idiom-specific properties, whereas models showed reduced or absent PIA sensitivity.

Taken together, these results suggest that the tested models partly approximate human idiom interpretation in context, but do not fully reproduce human idiomatic sensitivity. In literal-compositional contexts, models appear too ready to accept literal continuations, especially for high-PIA idioms; in figurative-creative contexts, they show reduced sensitivity to the idiom-specific properties that guide human judgments. More broadly, these findings caution against treating human–model correlations alone as evidence of shared interpretive mechanisms.

Table 1: Examples of dialogue-based stimuli. Speaker A provides either a literal-compositional (LIT) or figurative-creative (FIG) context; speaker B’s response is fixed within each idiom and contains the idiom followed by a literal continuation. Examples are English translations of the original Italian materials.

| Idiom                                                                | PIA  | Context | Speaker A                                                                                                                                             | Speaker B                                                                       |
|----------------------------------------------------------------------|------|---------|-------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| <i>prendere per mano</i><br>("take by hand")                         | High | LIT     | You know, next week we are going with the whole family to Tokyo. I’m terrified of those huge street crossings, especially for my 4-year-old daughter. | Well, if you have to take her by hand, make sure you have a firm grip.          |
|                                                                      |      | FIG     | You know, at school I have to follow a child with learning disabilities. I have to guide her in learning basic math skills.                           |                                                                                 |
| <i>fare acqua da tutte le parti</i><br>("make water from all sides") | Low  | LIT     | You know, yesterday a pipe burst in our bathroom and it’s continuing to leak. The floor is a mess, it’s completely wet!                               | Well, if it has to make water from all sides, make sure you have large buckets. |
|                                                                      |      | FIG     | You know, I thought I had a good idea for the doctoral project. But at the moment, I don’t think it’s working.                                        |                                                                                 |

Interaction of idiom PIA and condition across humans and language models

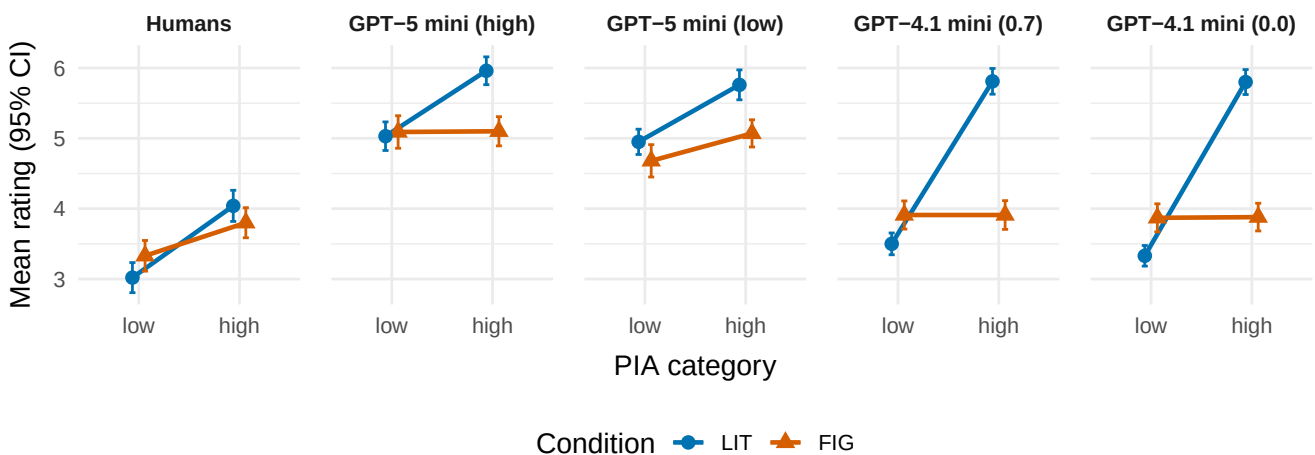


Figure 1: Mean ratings by source (humans and language models), contextual condition (literal vs. figurative), and PIA category (high vs. low). Error bars represent 95% confidence intervals around the mean.

## References

- [1] Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. A Hard Nut to Crack: Idiom Detection with Conversational Large Language Models. In Debanjan Ghosh, Smaranda Muresan, Anna Feldman, Tuhin Chakrabarty, and Emmy Liu, editors, *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid), June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.figlang-1.5.

- [2] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, July 2023. ISSN 1364-6613. doi: 10.1016/j.tics.2023.04.008.
- [3] Anca Dinu and Andra-Maria Florescu. Testing Language Creativity of Large Language Models and Humans. In Mika Härmäläinen, Emily Öhman, Yuri Bizzoni, So Miyagawa, and Khalid Alnajjar, editors, *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 426–436, Albuquerque, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-234-3. doi: 10.18653/v1/2025.nlp4dh-1.37.
- [4] Rachel Giora. *On Our Mind: Salience, Context, and Figurative Language*. Oxford University Press, May 2003. ISBN 9780195136166. doi: 10.1093/acprof:oso/9780195136166.001.0001.
- [5] Andreas Langlotz. *Idiomatic Creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. John Benjamins Publishing Company, April 2006. ISBN 9789027293763. doi: 10.1075/hcp.17.
- [6] Maya R. Libben and Debra A. Titone. The multidetermined nature of idiom processing. *Memory & Cognition*, 36(6):1103–1121, September 2008. ISSN 1532-5946. doi: 10.3758/mc.36.6.1103.
- [7] Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.362. URL <https://aclanthology.org/2025.acl-long.362/>.
- [8] Irene Pagliai. ‘well, if you want to play with fire, make sure the flames are tamed’: investigating the integration of idiom literal completions across literal and creative contexts. *Language and Cognition*, 17, 2025. ISSN 1866-9859. doi: 10.1017/langcog.2025.10013.
- [9] Sarah Schröder, Thekla Morgenroth, Ulrike Kuhl, Valerie Vaquet, and Benjamin Paaßen. Large Language Models Do Not Simulate Human Psychology, August 2025.
- [10] Wiltrud Wagner. *Idioms and Ambiguity in Context: Phrasal and Compositional Readings of Idiomatic Expressions*. De Gruyter, November 2020. ISBN 9783110685459. doi: 10.1515/9783110685459.