Overthinking doesn't help: Large language models on interpreting pseudowords

Jing Chen[1], *Marco Marelli*[1]

[1] *Department of Psychology, University of Milano-Bicocca, Italy*

Background: Humans assign meanings to pseudowords by exploiting sublexical form-meaning mappings stored in semantic memory [1, 2]. However, how humans segment a novel string of characters into (familiar) sublexical units to infer potential meanings remains an open question. Recent large language models (LLMs) provide a computational testbed for this question: engineered as probability distributions over token strings [3], LLMs segment input text into tokens determined by frequency-driven algorithms such as byte-pair encoding [4]. This offers a window to investigate whether LLMs, operating on such statistically derived sublexical units, can produce pseudoword interpretations that align with human intuitions.

Method: Here, we probed LLMs' ability to interpret Italian pseudowords using two tasks previously administered to human participants (60 per task) in [1]. Both tasks employed a two-alternative forced-choice (2AFC) design: in EXP1, participants selected the more semantically similar pseudoword for a given real word; in EXP2, they selected the more semantically similar real word for a given pseudoword (see Fig. 1a). Each task comprised 50 triplets, with a target and two alternatives paired via cosine similarity in fastText [5], which can derive vectors for pseudowords by summing character n-gram embeddings: one alternative was randomly selected among those with high cosine similarity to the target and the other among those with low similarity (see Fig. 1b). We instructed the same tasks to four state-of-the-art LLMs: two standard LLMs (GPT-4o, DeepSeek-v3) and two with reasoning capabilities (o4-mini, DeepSeek-R1; see Fig. 1c), performing four independent runs (i.e., reduce output randomness) at temperature 0.0 per task per model (i.e., give more deterministic output). We then assessed whether LLMs matched choices of human participants above chance (50%) using generalized linear mixed models (GLMMs) with random intercepts for items and participants (see Fig. 1d). We also compared against the alignment between fastText and human reported in [1] as baselines, examined whether reasoning LLMs outperformed standard ones, and whether reasoning costs (i.e., number of reasoning tokens) predicted alignment with human choices.

Results: LLMs achieved above-chance alignment with human participants in both experiments, with the best performance from a reasoning LLM, o4-mini, reaching 59.1% in EXP1 (z = 2.91, p < .01) and 62.5% in EXP2 (z = 4.35, p < .001). However, all LLMs fell substantially short of the fastText baseline in EXP1 reported in [1] (66.1%, z = 6.86, p < .001). In EXP2, while o4-mini and GPT-4o matched or exceeded fastText performance (58.0%, z = 2.49, p = .013), DeepSeek-R1 and DeepSeek-V3 did not consistently outperform the baseline. Furthermore, reasoning capabilities did not confer a consistent advantage, and their higher reasoning costs did not correlate with better human alignment in either experiment (all p > .05).

Discussion: Above-chance alignment from LLMs adds new empirical evidence that overall distributional information of (sublexical) forms predicts human pseudoword interpretation. However, lower performance relative to fastText, combined with no consistent benefit from reasoning capabilities, needs further investigation on LLMs tokenization strategies to identify the driving force behind the alignment.

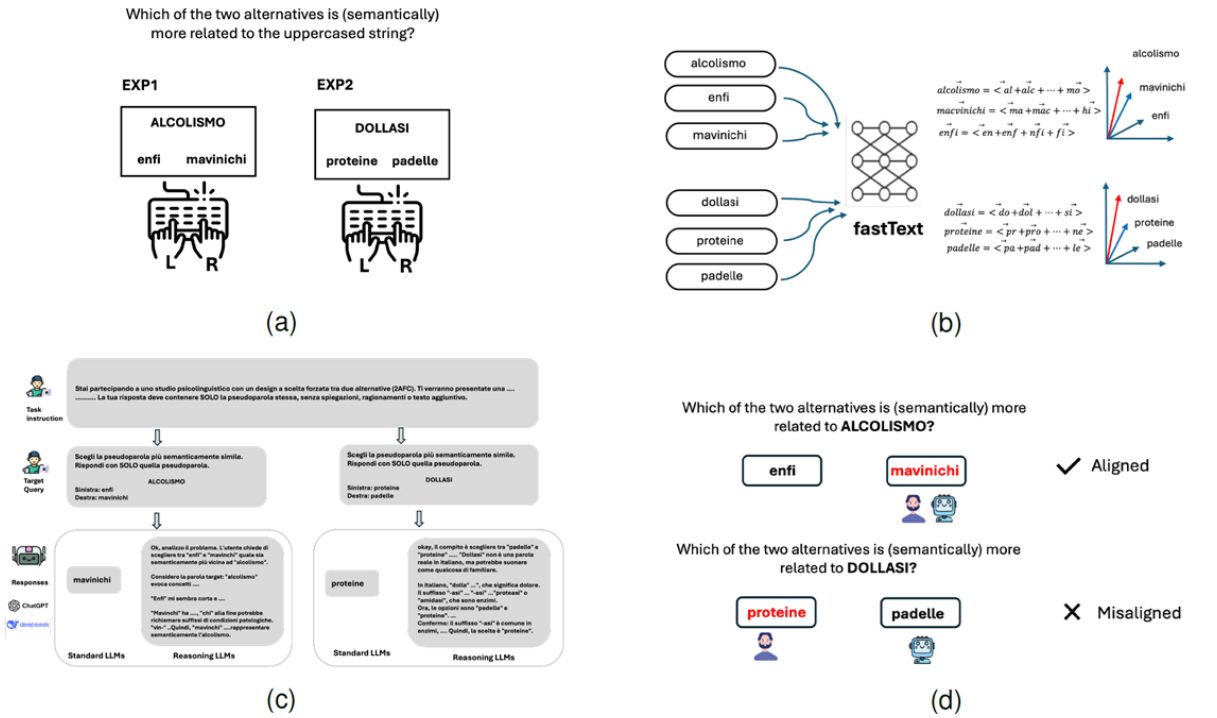

Figure 1: Overview. **a.** Example of the two-alternative forced-choice (2AFC) design in EXP1 and EXP2. Data sourced from [1]. **b.** Character n-gram based fastText provides cosine similarities between the target and each alternative, with higher scores indicating greater semantic similarity. Data sourced from [1]. **c.** Example of LLMs prompted to respond to each task in the 2AFC design. Standard LLMs only return their choice between the two options, while reasoning LLMs were instructed to produce a reasoning trace before the final decision. **d.** Examples where human and LLM select the same option (top) and different options (bottom), respectively.

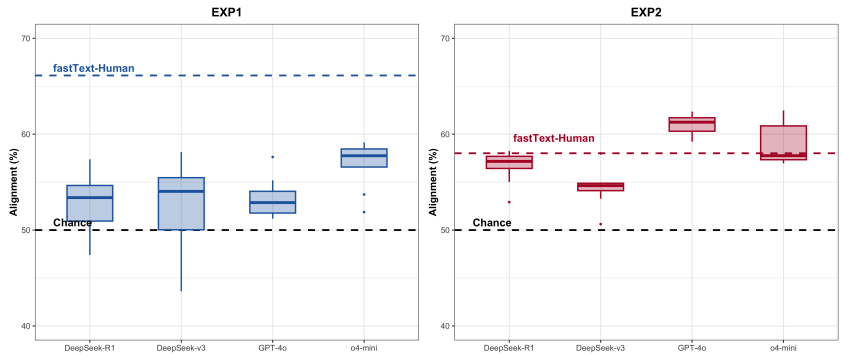

Figure 2: LLM-Human alignment on the two 2AFC tasks. The boxplots were plotted for each model across multiple runs. The dashed **black** line indicates the chance level (50%), while the dashed lines in **blue** and **red** refer to the baselines for EXP1 and EXP2, fastText predicted accuracy reported in [1], respectively.

References

[1] Gatti, D., Rodio, F., Rinaldi, L., & Marelli, M. (2024). On humans' (explicit) intuitions about the meaning of novel words. *Cognition*, *251*, 105882. https://doi.org/10.1016/j.cognition.2024.105882

[2] Bonandrini, R., Amenta, S., Sulpizio, S., Tettamanti, M., Mazzucchelli, A., & Marelli, M. (2023). Form to meaning mapping and the impact of explicit morpheme combination in novel word processing. *Cognitive Psychology*, *145*, 101594. https://doi.org/10.1016/j.cogpsych.2023.101594

[3] Vieira, T., Lebrun, B., Giulianelli, M., Gastaldi, J., Dusell, B., Terilla, J., O'donnell, T., & Cotterell, R. (2024). From language models over tokens to language models over characters.

[4] Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units* (Vol. 1). Association for Computational Linguistics.

[5] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching word vectors with subword information* (Vol. 5). Association for Computational Linguistics.