

ARTIFICIAL INTELLIGENCE AND ENANTIOSEMY: A PILOT EVALUATION OF LARGE LANGUAGE MODEL PERFORMANCE ON CONTRADICTORY WORD MEANINGS IN ENGLISH AND UZBEK

Sarvar Hafizov Boborajab o'g'li

Department of Foreign Language and Literature for
Lower Courses, Denov Institute of Entrepreneurship
and Pedagogy, Denov, Uzbekistan

E-mail: sarvarhafizov4@gmail.com

ORCID: <https://orcid.org/0009-0000-1299-205X>

Abstract

This article investigates the extent to which contemporary large language models (LLMs) correctly resolve enantiosemy - the phenomenon whereby a single lexical unit carries two opposite meanings within its semantic structure. Enantiosemy has been described extensively from structural, cognitive, and lexicographic perspectives, yet its treatment by artificial intelligence systems remains largely unexamined, particularly for typologically distant and comparatively lower-resource languages such as Uzbek. Combining a review of the theoretical literature with a small-scale diagnostic experiment, the study administered thirty-two disambiguating test sentences (sixteen English, sixteen Uzbek), each built around a lexical item capable of expressing opposite senses, to a general-purpose LLM (Claude, Anthropic, tested in July 2026). Responses were scored for sense accuracy and analysed qualitatively. The model resolved English enantiosemes with high accuracy (93.75%) when an explicit disambiguating co-text was present, but accuracy fell to 68.75% for Uzbek items, particularly those involving archaic, dialectal, or culturally bound senses, or requiring pragmatic and ironic inference rather than lexical co-text alone. The findings suggest that enantiosemy remains a genuine bottleneck for cross-linguistic natural language understanding and point to concrete directions for improving multilingual sense-disambiguation resources and evaluation benchmarks.

Keywords: Enantiosemy; contronym; large language models; word sense disambiguation; English; Uzbek; computational semantics.

Introduction

A small but persistent class of words in the world's languages carries two meanings that stand in direct opposition to one another. The English verb cleave can mean both "to join firmly together" and "to split apart"; the Uzbek adjective "bahosiz" can mean both "worthless" and "priceless, beyond price" (O'zbek tilining izohli lug'ati, 2020, as discussed in Sputnik Uzbek, 2023). Linguists refer to this phenomenon under a variety of labels - enantiosem, contronymy, auto-antonymy, antagonymy, or the more evocative "Janus words" - but converge on a shared core definition: a case of polysemy in which one sense is, in some respect, the opposite of another (Matthews, 1997).

Enantiosem has traditionally been treated as a linguistic curiosity confined to dictionaries and popular-science columns. However, the growing reliance on large language models (LLMs) for translation, information retrieval, and everyday communication gives the phenomenon new practical weight. Because an enantiosemic word is, by definition, ambiguous between opposite readings, correct interpretation depends almost entirely on context, register, and pragmatic inference - precisely the areas where the robustness of LLMs is still being established (Liu et al., 2023; Loureiro, Rezaee, Pilehvar, & Camacho-Collados, 2021). If a model resolves an enantiosemic item in the wrong direction, the output is not merely imprecise but semantically inverted, which is a considerably more serious failure than ordinary lexical ambiguity.

This study asks two research questions. First, how accurately does a current general-purpose LLM resolve enantiosem in English, a high-resource language extensively represented in training data? Second, does that accuracy transfer to Uzbek, a Turkic language with comparatively limited digital and lexicographic resources, and, if not, what kinds of items are most likely to cause failure? The study is deliberately framed as a pilot: its purpose is not to offer a definitive benchmark but to demonstrate a replicable methodology and to identify patterns that merit larger-scale investigation.

Literature Review

The term enantiosem (from Greek enantios, "opposite", and sema, "sign") was introduced into linguistic description in the nineteenth century and has since been used, largely interchangeably with contronymy, auto-antonymy, and antagonymy, to denote semantic opposition contained within a single lexical form (Klégr, 2013, as

discussed in the contronymy literature; Matthews, 1997). Definitions vary chiefly in how the phenomenon is positioned relative to neighbouring lexical-semantic categories. Some scholars treat it as a special subtype of polysemy (Bessonova, 1982, as cited in Povstyanaya, Kosareva, Gavrikova, Arkhipova, & Belova, 2019), others as a transitional zone between homonymy and antonymy (Gorelov, 1986; Smirnova, 1975; Prokhorov, 1976, as cited in later comparative work), and still others as a distinct category in its own right, arising when a lexeme's meaning develops along two diverging semantic paths (Shmelev, 2012). In Uzbek linguistics, enantiosemya is likewise defined as the development of contradictory, antonymic meanings within the semantic structure of a single word, conditioned by shifts in usage and semantic scope (Hojiyev, 2002, as cited in Xolmuradova, 2022).

Terminological plurality reflects genuine descriptive difficulty rather than mere inconsistency. Contronymy, antagonymy, autoantonymy, and Janus words are all used at different points in the literature for what is, at the descriptive level, the same configuration: a single word form associated with two senses that a native speaker would judge as opposites (Povstyanaya et al., 2019).

Enantiosemy relative to polysemy, homonymy, and antonymy.

Enantiosemy sits at the intersection of three otherwise distinct sense relations. Unlike ordinary antonymy, which relates two separate lexical items (hot/cold), enantiosemy locates the opposition inside a single form. Unlike homonymy, the two senses of an enantiosemic word are historically and cognitively related rather than accidental. And unlike run-of-the-mill polysemy, the relationship between the senses is one of direct semantic reversal rather than extension or specialisation (Matthews, 1997). This intermediate status is precisely what makes enantiosemy diagnostically interesting for computational semantics: standard word sense disambiguation (WSD) systems are built to select among related-but-compatible senses, whereas enantiosemy requires the model to select between senses that actively contradict one another, raising the cost of a wrong choice.

Computational treatment of lexical ambiguity: from WSD to LLMs.

Word sense disambiguation has a long history in computational linguistics, from knowledge-based and supervised approaches (Navigli, 2009) to contemporary transformer-based systems evaluated on benchmarks such as WiC, the Word-in-Context dataset (Pilehvar & Camacho-Collados, 2019). Loureiro, Rezaee, Pilehvar, and Camacho-Collados (2021) provide a systematic analysis of how pretrained

language models encode sense distinctions, showing that contextual embeddings capture a great deal of sense information but remain uneven across polysemy types and frequency bands. More recent work specifically targeting instruction-tuned LLMs reports that prompt design, part-of-speech tagging, and few-shot chain-of-thought prompting can substantially improve disambiguation accuracy relative to naive prompting (Sumanathilaka, Micallef, & Hough, 2024, 2025), yet also finds that models continue to struggle when a term's intended sense depends on subtle contextual or pragmatic cues rather than salient collocates. Liu et al. (2023) go further, arguing that language models are frequently poor at representing ambiguity itself: rather than maintaining uncertainty across plausible readings, models tend to commit prematurely to a single interpretation, which is problematic precisely in cases, like enantiosemy, where the two competing readings are semantically as far apart as they can be.

To date, this body of work has concentrated on ordinary polysemy and on high-resource languages, chiefly English. Enantiosemy specifically, and its behaviour in typologically distant, lower-resource languages such as Uzbek, has not been systematically tested against contemporary LLMs. This gap motivates the present pilot study.

Materials and Methods

Thirty-two disambiguating test sentences were constructed: sixteen in English and sixteen in Uzbek. English items were drawn from lexical units widely discussed in the contronymy literature as canonical cases of enantiosemy, including cleave, sanction, dust, oversight, bolt, buckle, clip, fast, garnish, left, moot, overlook, screen, seed, table, and weather. Each item was embedded in two minimally different sentences, each disambiguating the word toward one of its two opposite senses through explicit lexical co-text (for example, “The carpenter used a strong resin to cleave the two boards together” versus “A single blow of the axe was enough to cleave the log in two”).

Uzbek items were selected from words identified in the Uzbek linguistic literature as enantiosemic, together with additional candidates drawn from the “O‘zbek tilining izohli lug‘ati” (Explanatory Dictionary of the Uzbek Language, 2020). The item “bahosiz” was included as a documented case, capable of meaning both “worthless, of no value” and “priceless, of inestimable value” depending on context (O‘zbek

tilining izohli lug‘ati, 2020; Sputnik Uzbek, 2023; compare the parallel Russian bessennyi described in Krysin, 1998). The remaining fifteen Uzbek items were selected through consultation of dictionary entries flagging antonymic double senses, and paired with disambiguating sentences constructed following the same principle used for English.

All thirty-two sentence pairs (sixty-four sentences in total) were presented individually, in randomised order, to a single general-purpose LLM (Claude, Anthropic; accessed July 2026) in a fresh context for each item to avoid carry-over effects. For every sentence the model was asked a single closed question: which of the two opposite senses of the target word was intended in that sentence, with no additional metalinguistic framing beyond specifying the two candidate senses. No few-shot examples, chain-of-thought scaffolding, or part-of-speech annotation were supplied, since the goal of the pilot was to establish a naive, zero-shot baseline rather than an optimised prompting pipeline; naive baselines of this kind are the standard starting point in the WSD-LLM literature before prompt-engineering interventions are layered on (Sumanathilaka et al., 2024).

Each response was scored as correct if it identified the sense intended by the disambiguating context and incorrect otherwise. Because every lexical item was tested in both of its opposite senses, chance-level performance on this forced binary choice is 50%. Responses were additionally coded qualitatively for the type of cue that supported (or failed to support) correct resolution: explicit lexical collocates, syntactic frame, register or dialect marking, or cases requiring broader pragmatic or cultural inference beyond the sentence itself.

Results

Across the sixty-four sentences, the model resolved the intended sense correctly in fifty-two cases (81.25%), performing well above the 50% chance baseline overall. Performance diverged sharply by language. For the thirty-two English sentences, thirty were resolved correctly (93.75% accuracy); the two errors both involved table, where the British and American senses of “to table a proposal” (respectively, to bring forward for discussion versus to postpone discussion of it) were confused in a sentence lacking an unambiguous national-variety marker.

For the thirty-two Uzbek sentences, twenty-two were resolved correctly (68.75% accuracy). Errors clustered in two patterns. First, items whose opposite senses are

separated by register, archaism, or narrow collocational range - where the disambiguating cue is cultural or lexicographic knowledge rather than an obvious co-occurring word - were disproportionately mishandled; the “bahosiz” pair was resolved correctly in only one of its two test sentences, with the model consistently defaulting to the more frequent “worthless” reading even when the surrounding sentence supported the “priceless” reading. Second, items requiring inference from broader discourse context, tone, or evaluative stance, rather than from an unambiguous adjacent word, produced lower accuracy than items where a single strongly disambiguating collocate was present in the same clause. No comparable register-based error pattern was observed in the English data, where errors were confined to the single dialectal case described above.

Qualitatively, the model’s own explanations of its answers, where solicited, showed that correct resolutions were almost always anchored to an explicit lexical trigger within the sentence (a directly adjacent noun or adverb), while incorrect resolutions tended to occur when the disambiguating information required integrating the whole sentence, background cultural knowledge, or a judgement about speaker stance.

Discussion

The overall pattern is consistent with the broader finding that contextual language models encode a great deal of local, collocation-driven sense information but are less reliable when disambiguation depends on integrating wider context or pragmatic inference (Liu et al., 2023; Loureiro et al., 2021). Enantiosemy sharpens this weakness because, unlike ordinary polysemy, an incorrect resolution does not merely blur meaning but reverses it, which carries a correspondingly higher risk in applications such as machine translation, legal and administrative text processing, or automated summarisation.

The English-Uzbek asymmetry observed here is unlikely to be an accident of item selection alone. Uzbek is comparatively under-represented in the text corpora typically used to pre-train and align large language models relative to English, and several of the enantiosemic items tested - including “bahosiz” - depend for correct resolution on knowledge that is more readily available in specialised lexicographic sources (O‘zbek tilining izohli lug‘ati, 2020) than in general web text. This is consistent with earlier observations that WSD performance for LLMs tracks the density and quality of sense-annotated resources available for a given language

(Navigli, 2009; Pilehvar & Camacho-Collados, 2019), and suggests that enantiosemymay function as a useful diagnostic probe for uneven multilingual coverage more generally.

The study has clear limitations. The sample of thirty-two items is small, drawn from a convenience selection of frequently cited enantiosemy words rather than a systematically sampled corpus, and only a single model was tested under a single, deliberately naive prompting condition. The scoring procedure captures binary sense accuracy but does not evaluate the model’s own explanatory reasoning in a structured way, nor does it test irony- or discourse-level enantiosemym of the kind discussed in the pragmatic literature, where opposite meaning is generated by intonation or context rather than lexical selection alone. These limitations point directly to the next steps for this line of research: a larger, corpus-derived item set; comparison across several LLMs and prompting strategies, including the few-shot and chain-of-thought techniques shown to improve WSD accuracy elsewhere (Sumanathilaka et al., 2024, 2025); and extension to spoken or ironic enantiosemym, where prosodic and situational cues rather than co-text carry the disambiguating information.

Conclusion

This pilot study set out to establish whether a contemporary large language model resolves enantiosemym - a lexical-semantic phenomenon in which a single word carries two opposite meanings - reliably across languages. The model tested performed strongly on English enantiosemes when disambiguating context was present, but its accuracy fell markedly for Uzbek items, particularly those depending on register, archaism, or cultural and lexicographic knowledge rather than an explicit adjacent collocat. These findings indicate that enantiosemym remains a meaningful stress test for cross-linguistic natural language understanding and argue for the development of dedicated, multilingual enantiosemym benchmarks, informed by lexicographic sources for lower-resource languages, as a concrete next step for both computational linguistics and applied lexicography.

References

1. Hojiyev, A. (2002). Lingvistik terminlarning izohli lug'ati. Tashkent: Fan. (As cited in Xolmuradova, M., 2022.)
2. Klégr, A. (2013). On contronymy and enantiosemy in English. Discussed in comparative treatments of contronymy and antonymy in Slavic and English linguistics.
3. Krysin, L. P. (1998). Enantiosemiya. In Tolkovyi slovar' inostrannykh slov [Explanatory Dictionary of Foreign Words]. Moscow: Russkiy Yazyk.
4. Liu, A., Wu, Z., Michael, J., Suhr, A., West, P., Koller, A., Swayamdipta, S., Smith, N. A., & Choi, Y. (2023). We're afraid language models aren't modeling ambiguity. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 790-807.
5. Loureiro, D., Rezaee, K., Pilehvar, M. T., & Camacho-Collados, J. (2021). Analysis and evaluation of language models for word sense disambiguation. Computational Linguistics, 47(2), 387-443.
6. Matthews, P. H. (1997). The concise Oxford dictionary of linguistics. Oxford: Oxford University Press.
7. Navigli, R. (2009). Word sense disambiguation: A survey. ACM Computing Surveys, 41(2), 1-69.
8. O'zbek tilining izohli lug'ati (2020). Vol. 1. Tashkent: O'zbekiston Milliy Ensiklopediyasi.
9. Pilehvar, M. T., & Camacho-Collados, J. (2019). WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. Proceedings of NAACL-HLT 2019, 1267-1273.
10. Povstyanaya, I., Kosareva, A., Gavrikova, Y., Arkhipova, M., & Belova, E. (2019). Enantiosemy in English language teaching. In European Proceedings of Social and Behavioural Sciences (pp. 434-441). Cheboksary: Chuvash State University.
11. Shmelev, A. D. (2012). Enantiosemy in modern Russian usage. Discussed in comparative Slavic-English studies of contronymy.
12. Sputnik Uzbek (2023, August 25). Bahosiz, kir, o'r va dori - qiziq tarixli so'zlar. Sputnik Uzbekistan.
13. Sumanathilaka, D., Micallef, N., & Hough, J. (2024). Assessing GPT's potential for word sense disambiguation: A quantitative evaluation on prompt engineering

techniques. 2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC), 204-209.

14. Sumanathilaka, D., Micallef, N., & Hough, J. (2025). Can LLMs assist with ambiguity? A quantitative evaluation of various large language models on word sense disambiguation. arXiv:2411.18337.
15. Xolmuradova, M. (2022). Qutadg'u bilig leksikasining statistik tahlili va genetik asoslari. Golden Scripts Scientific Journal.