HUMAN LANGUAGE AND MACHINE INTELLIGENCE: THE CHALLENGES OF INTERPRETING METAPHOR, IRONY, AND CONTEXT IN AI-GENERATED DISCOURSE

Fariza Ergasheva O‘lmas qizi
EFL Teacher, Sharq University, Navoi, Uzbekistan
fara0040@gmail.com

Annotatsiya. *Ushbu maqolada metafora, kinoya yoki bevosita aytilmagan kontekst ma’noni belgilaganida sun’iy intellekt yaratgan ravon nutq nima uchun pragmatik jihatdan nomuvofiq bo‘lishi mumkinligi o‘rganiladi. Tadqiqot o‘n ikkita ilmiy nashrning tanlanma sharhi va e’lon qilingan benchmark statistikasining tavsifiy tahliliga asoslanadi. Ko‘chma ma’noni aniqlash, izohlash sifati va yaratilgan javobning vaziyatga mosligi alohida ko‘rib chiqiladi. Kinoyani aniqlashga doir tarixiy natijalar hamda keyingi pragmatik sinovlar baholash vazifalarining xilma-xilligini ko‘rsatadi. Mualliflik misollari bir xil jumlaning kontekst o‘zgarganda turlicha talqin qilinishini ochib beradi. Kontekstlarni qarama-qarshi qo‘yishga asoslangan baholash yondashuvi taklif etiladi. Maqolaning hissasi benchmark tuzilishini diskurs tahlili va ingliz tilini xorijiy til sifatida o‘qitish bilan bog‘lashdan iborat. Xulosalarda kontekstual baholash va inson tekshiruvi tavsiya qilinadi, biroq taklif etilgan yondashuv tajribada tasdiqlangan usul sifatida taqdim etilmaydi.*

Kalit so‘zlar: *sun’iy intellekt; yaratilgan diskurs; metafora; kinoya; pragmatika; kontekstual xulosa; katta til modellari; ingliz tilini o‘qitish.*

Аннотация. *В статье рассматривается, почему беглая речь, генерируемая искусственным интеллектом, может оставаться прагматически неуместной, когда смысл определяется метафорой, иронией или невыраженным контекстом. Исследование сочетает выборочный обзор двенадцати публикаций с описательным анализом опубликованной статистики тестовых наборов. Разграничиваются распознавание переносного значения, качество объяснения и уместность сгенерированного ответа. Исторические результаты распознавания иронии и более поздние прагматические тесты показывают разнообразие оценочных задач, а не единую траекторию машинного интеллекта. Авторские примеры демонстрируют, как изменение контекста меняет интерпретацию одной и той же реплики. Предлагается подход к оцениванию на основе сопоставления*

контекстов, учитывающий использование доказательств, пересмотр интерпретации и неопределённость. Вклад статьи состоит в объединении анализа тестовых заданий, дискурсивной оценки и преподавания английского языка как иностранного. Выводы обосновывают необходимость контекстуальной оценки и проверки человеком, не представляя предложенный подход как экспериментально подтверждённую методику.

***Ключевые слова**: искусственный интеллект; генерируемый дискурс; метафора; ирония; прагматика; контекстуальный вывод; большие языковые модели; преподавание английского языка.*

***Abstract.** This article examines why fluent AI-generated discourse can remain pragmatically inappropriate when metaphor, irony, or unstated context determines meaning. It combines a selective review of twelve publications with descriptive analysis of published benchmark statistics. The discussion distinguishes figurative recognition, explanation quality, and the suitability of a generated response. Historical irony-detection results and later pragmatic benchmarks illustrate the diversity of evaluation tasks rather than a single trajectory of model intelligence. Constructed examples show how an unchanged utterance can require different interpretations after its context changes. The article proposes a context-contrastive evaluation framework that checks evidence use, interpretive revision, and uncertainty. Its contribution is an integrated analytical account linking benchmark design to discourse-level evaluation and EFL teaching. The conclusions support contextual assessment and human review without treating the proposed framework as an experimentally validated intervention.*

***Keywords:** artificial intelligence; generated discourse; metaphor; irony; pragmatics; contextual inference; large language models; EFL education.*

Introduction

A sentence may be grammatically impeccable and still fail as a response. When a student describes an examination as “a mountain I cannot climb,” an assistant that explains mountain safety has recognized familiar words but missed the communicative problem. A more useful response identifies academic difficulty, asks what feels unmanageable, and offers support. The difference concerns the relation between an utterance and its situation. This relation becomes especially visible when AI systems produce explanations, feedback, and dialogue that readers are inclined to treat as socially informed.

The object of this study is AI-mediated discourse involving nonliteral and context-dependent meaning. Its subject is the relationship between interpretive

evidence and generated responses. The aim is to explain what existing evaluation practices establish and what they leave unresolved. Three questions guide the analysis: which abilities are tested by figurative-language benchmarks; how should their statistics be interpreted; and what additional checks are needed when an AI system must respond appropriately rather than select a label?

Bender and Koller distinguish linguistic form from communicative meaning and question whether training on form alone can establish the latter [1]. Their argument provides a conceptual caution, not an experimental verdict on every contemporary model. This article adopts a narrower operational position: a system demonstrates useful interpretive competence when its response follows the available evidence, changes appropriately after a relevant contextual change, and acknowledges unresolved ambiguity. Such behaviour can be evaluated without making claims about machine consciousness.

Literature review

Metaphor processing has long been a computational research problem. Shutova surveys approaches to modelling metaphor in natural language processing [2], while the VUA and TOEFL shared task establishes metaphor identification in both a general corpus and learner writing [3]. These contributions make identification measurable, but a token-level judgement does not itself establish whether a dialogue system can respond helpfully to the figurative statement. In EFL settings, an unusual expression may also reflect developing proficiency. Interpreters should therefore test several readings before deciding that a departure from conventional wording is either an error or a deliberate metaphor.

Irony research introduces a related problem of intention. The SemEval-2018 task evaluates binary and multiclass irony classification [4]. SARC provides self-annotated Reddit material at a much larger scale [5]. Oprea and Magdy make a further distinction between sarcasm intended by the author and sarcasm perceived by an audience [6]. This distinction matters because reader agreement cannot automatically be treated as direct access to a speaker’s intention. The appropriate target depends on the application: recovering an intended criticism and predicting whether readers will find a generated sentence insulting are different tasks.

FLUTE links figurative natural language inference to textual explanations and contains 9,000 instances covering sarcasm, simile, metaphor, and idioms [7]. The associated FigLang shared task asks systems to supply both an inference label and a convincing explanation [8]. Together, they bring the reasoning offered in support of an answer into evaluation. Nevertheless, an explanation should be examined for its dependence on the supplied situation. A plausible account can introduce

assumptions that were never present, so explanatory fluency should not be accepted as its own evidence.

Pragmatic evaluation extends beyond figurative language. PUB contains fourteen tasks across implicature, presupposition, reference, and deixis [9]. MultiPragEval evaluates pragmatic inference in English, German, Korean, and Chinese [10]. These studies broaden the range of tested inferences and languages; neither establishes performance for Uzbek-speaking EFL classrooms. MUStARD additionally draws attention to audiovisual and conversational cues in sarcasm [11]. The literature therefore supports an evaluation agenda sensitive to annotation targets, language, and available context. The present article connects these dimensions to generated responses through a proposed context-contrastive framework rather than introducing another detection model.

Research methodology

The study is a selective analytical review with descriptive secondary-data analysis. Twelve publications were selected purposively to cover conceptual foundations, metaphor identification, irony and sarcasm annotation, explanation-based inference, multilingual pragmatics, multimodality, and context use. Publication records and relevant benchmark documentation were checked through ACL Anthology. The corpus is bounded to the cited 2010–2024 publications and is not an exhaustive or systematic review of research available in 2026. No participants were recruited, no language model was experimentally tested, and no new accuracy measurements were collected.

For each benchmark, the analysis records its unit of observation, annotation target, and relation to discourse interpretation. Numerical values are retained with their original measurement meaning. Counts of tweets, inference pairs, and question units are not pooled. Published irony results are compared only within the same official subtask. Table 1 reports dataset characteristics; Table 2 and Figure 1 reproduce selected historical system results. Constructed utterances in Table 3 and Figure 3 are analytical examples created for this article, not quotations from participants or outputs sampled from a model.

The synthesis follows three operations. First, it separates recognition from explanation and response generation. Second, it asks which missing information could alter each judgement. Third, it translates the resulting vulnerabilities into observable evaluation criteria. The resulting framework and rubric are proposals. They require independent annotation, inter-rater assessment, and empirical validation before they can support claims about teaching outcomes or model improvement.

Analysis and results

Benchmark scale and the meaning of an evaluation unit

Table 1. Selected benchmark resources and their interpretive scope

Resource	Published scale	Unit and target	Boundary
SemEval-2018 [4]	3,834 training; 784 test	Tweets; irony classification	Short social-media discourse
SARC [5]	1.3 million sarcastic statements	Self-annotated Reddit statements	Count excludes the many nonsarcastic statements
FLUTE [7]	9,000 instances	Figurative inference pairs with explanations	Inference is not a full dialogue response
PUB [9]	About 28,000 data points	14 pragmatic tasks; multiple-choice answers	Includes 6,100 newly annotated points
MultiPragEval [10]	1,200 question units	Pragmatic inference in four languages	No direct evidence for Uzbek

Table 1 shows why dataset size cannot serve as a general measure of pragmatic coverage. A large collection may repeatedly sample one communicative setting, whereas a smaller resource may distinguish several inference types. The relevant question is whether the evidence needed for an interpretation is represented and whether the annotation matches the intended use. SARC's reported figure counts sarcastic statements rather than the entire corpus [5]; interpreting it as a balanced dataset total would distort both scale and class prevalence.

A similar caution applies to transfer. Social-media irony, examination feedback, and professional correspondence differ in audience expectations. The review does not estimate how often a model will fail in any of these settings. Instead, it identifies a design gap: corpus-level recognition scores leave the appropriateness of subsequent generated discourse only partly specified. A useful benchmark for an educational assistant would retain the learner's question, the relevant prior exchange, and the response that must be assessed.

Historical scores and what they do not measure

Table 2. Selected official SemEval-2018 Task A results

System	Precision	Recall	F1
THU_NGN	0.630	0.801	0.705
NTUA-SLP	0.654	0.691	0.672
WLV	0.532	0.836	0.650

Source: Van Hee et al. [4], Table 3; official binary irony results. Values concern the ironic class. These are historical submitted systems, not contemporary LLMs.

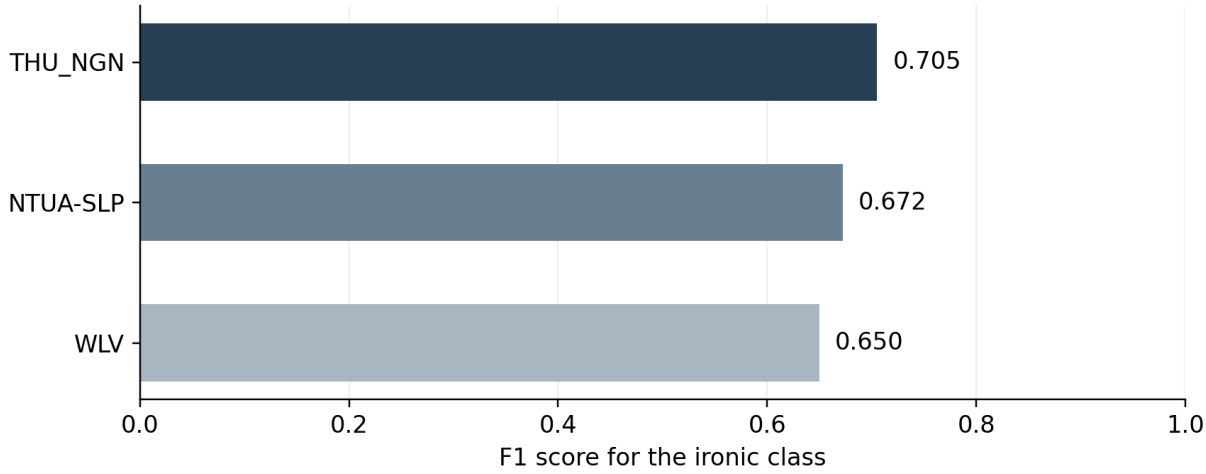


Figure 1. F1 scores for three selected official systems in the same binary irony task. Data: [4], Table 3. The axis spans 0–1; the figure is not a current model ranking.

The selected systems have different precision–recall balances. WLV retrieves a larger share of ironic cases but has lower precision; THU_NGN obtains the highest F1 among these entries [4]. An instructional tool may need to avoid falsely accusing a learner of sarcasm, while an exploratory corpus search may prioritize recall. The preferred balance therefore depends on the downstream cost of each error. F1 is the harmonic mean of precision and recall, $F1 = 2PR/(P + R)$; it is not the percentage of all statements correctly interpreted.

Rounded aggregate scores do not permit a significance test, a confidence interval, or reconstruction of paired prediction errors. Accordingly, no statistical significance is claimed here. Nor can these results quantify the ability to generate an appropriate apology, clarification, or explanation. Those behaviours require separate response-level assessment. This distinction prevents an apparently precise chart from supporting a broader claim than the underlying task allows.

Metaphor as a constraint on response content

Consider “The feedback opened a door.” A reasonable interpretation concerns a newly available opportunity or understanding, but the exact referent depends on what preceded the statement. The interpreter must preserve the relevant relationship while avoiding unnecessary physical details. An AI-generated answer that expands the sentence into an elaborate building narrative may be creative yet unhelpful. For classroom feedback, the useful continuation asks what became clearer or which next step is now possible.

Metaphor evaluation should therefore include paraphrase adequacy and response relevance. The expression should first be related to its local context, after which a literal paraphrase can test the proposed mapping. A second context can then

check whether the system is merely repeating a memorized explanation. If "opened a door" describes a literal action in a school corridor, an unchanged figurative interpretation would reveal overgeneralization. This contrast is an analytical test design, not an observed failure rate.

Table 3. Constructed diagnostic examples for discourse evaluation

Expression	Relevant situation	Supported reading	Response risk
"The exam is a mountain."	A learner describes difficulty preparing.	A demanding challenge	Advice about literal climbing
"Wonderful timing."	A reminder arrives after the deadline.	Possible ironic criticism	Unqualified agreement with praise
"It is cold in here."	An open window is within reach.	Possible indirect request	Missing a request; or assuming one without evidence
"She has done it again."	No prior referent or action is available.	Underdetermined reference	Inventing who "she" is or what happened

Note: all examples are constructed for this article. The suggested readings are contextual possibilities, not universally fixed interpretations.

Irony and uncertainty about social meaning

Irony often makes the literal evaluation unreliable as a guide to the intended stance. Yet opposite sentiment is not a sufficient decision rule. Praise after a mishap may be playful solidarity, irritation, self-directed humour, or a literal judgement about one successful aspect of an otherwise unsuccessful event. A system should not convert every positive word in a negative setting into hostility. The distinction between intended and perceived sarcasm documented by iSarcasm is especially relevant here [6].

The downstream response makes these distinctions consequential. If the user says "Wonderful, another rejected application," a cheerful congratulation would normally misread the event. A response that acknowledges disappointment is safer interpretively, while a categorical accusation about the user's emotional state would exceed the evidence. The system can recognize a likely ironic stance and still leave space for correction. Such calibrated language is a testable response property, even where annotators disagree on a single irony label.

Multimodal evidence can help resolve cases that text alone leaves open. MUStARD includes audiovisual utterances and preceding dialogue [11]. This does not imply that every text-based ambiguity can be repaired by adding more words. Some cues are unavailable because the interaction never recorded them. In such

cases, requesting clarification or offering a qualified reading is more defensible than generating a detailed account of the speaker's private intention.

Context selection and interpretive revision

Context has to be relevant, accessible, and used. Liu et al. show that the position of relevant information can affect performance on long-context question answering and retrieval [12]. These are not irony experiments. The appropriate inference for the present framework is methodological: a pragmatic evaluation should test where decisive cues appear, rather than assuming that inclusion in the prompt guarantees their use. No numerical long-context result is transferred to metaphor or irony in this article.

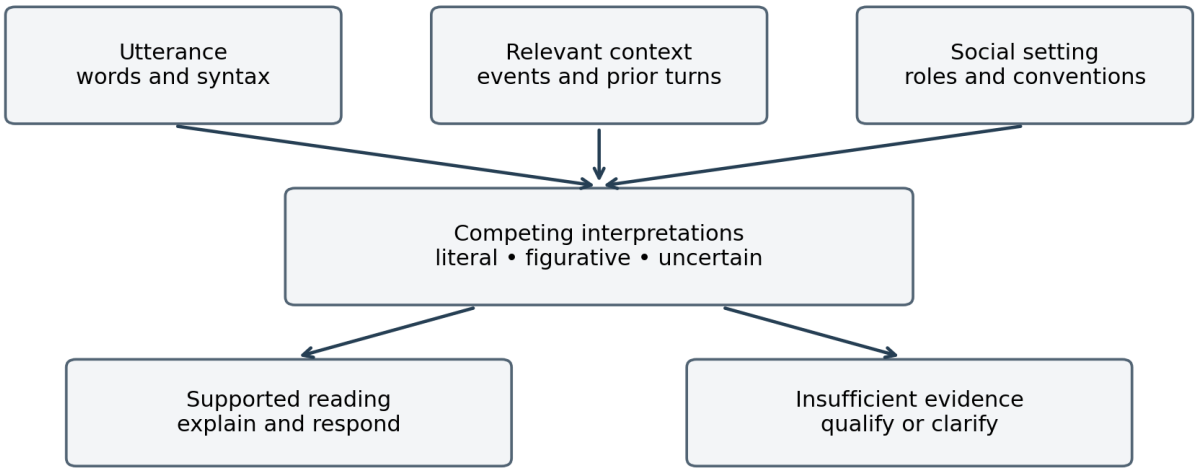


Figure 2. Proposed framework for context-sensitive discourse evaluation. Author's conceptual synthesis; arrows indicate interpretive dependencies, not a measured model architecture.

Figure 2 separates evidence gathering from commitment to a reading. The first stage considers the utterance, prior events, and social setting. The second compares interpretations rather than selecting the first plausible story. The final branch concerns response behaviour: where evidence is adequate, the system should explain and respond; where it is not, it should qualify its answer or ask a focused question. This structure also helps evaluators locate an error. A poor response may result from missing evidence, an unsupported inference, or an unsuitable communicative action.

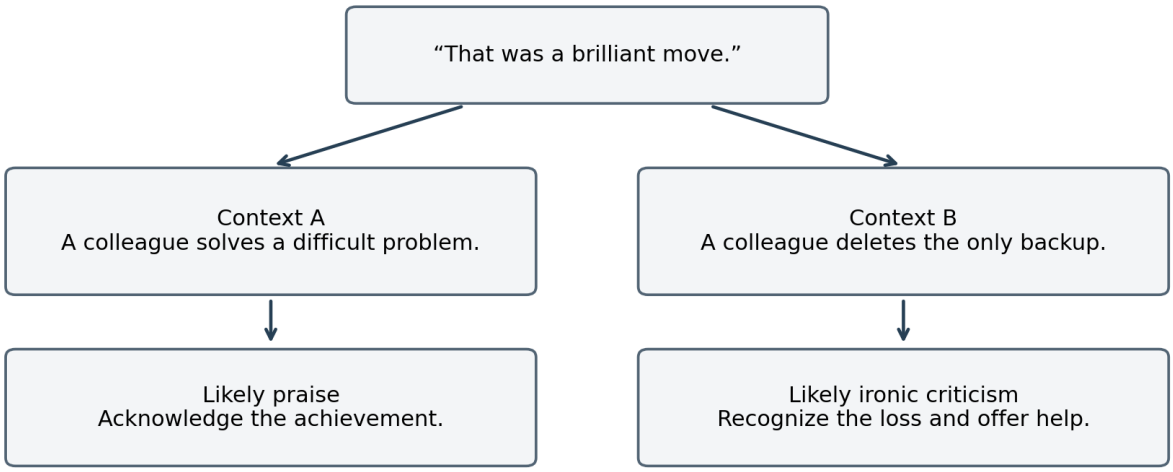


Figure 3. A constructed context contrast using identical wording. The two readings are likely under the stated situations, not guaranteed; tone or relationship could modify them.

The contrast in Figure 3 is deliberately minimal. The utterance remains unchanged while the preceding event changes. A successful system should alter both its interpretation and its response. Conversely, an irrelevant contextual change, such as replacing a room number, should not normally reverse its reading. Testing meaningful changes alongside irrelevant changes would distinguish contextual responsiveness from instability. Both requirements matter: a system that changes every answer is no more dependable than one that changes none.

A proposed rubric for generated discourse

Table 4. Proposed response-level rubric for future validation

Criterion	0 — inadequate	1 — partial	2 — adequate
Meaning preservation	Misreads the core meaning	Captures part of the meaning	Preserves the contextually supported meaning
Evidence use	Invents decisive context	Uses evidence incompletely	Links interpretation to available cues
Contextual revision	Ignores a decisive change	Revises inconsistently	Revises for relevant changes; stays stable otherwise
Response suitability	Socially or task-inappropriate	Relevant but poorly adapted	Fits the user's apparent communicative need
Uncertainty	Unjustified certainty	Vague hedging	Identifies specific ambiguity or asks a useful question

The rubric is intended to be reported as a profile across five dimensions. A combined 0–10 score may be calculated for exploratory use, but it would assume equal weighting and must not conceal a serious failure in one dimension. No scores are assigned to existing models here. A validation study should publish item-level prompts, model versions, generation settings, response samples, and annotation instructions. At least two independent evaluators should assess responses before adjudication, with agreement reported for each criterion. Ambiguous items should permit more than one justified interpretation.

A suitable experiment would pair each utterance with a context that supports its literal reading and another that supports a figurative reading. It should also include cases where the available evidence remains insufficient. Evaluation can then measure supported interpretation, appropriate revision, and unsupported certainty separately. Repeated generations are necessary when sampling introduces variation. Human reference responses should be collected under the same information conditions; otherwise, the comparison could reward human access to context that the model never received.

Implications for EFL teaching and limitations

In an EFL lesson, the proposed framework can organize a discussion around evidence rather than a single authoritative AI answer. Learners can identify which words permit a literal reading, which contextual facts favour another reading, and what response would be appropriate. They can then compare two generated explanations and revise the weaker one. The teacher’s role is to make the reasoning discussable, especially where a culturally familiar expression does not transfer directly into another language. This is a proposed instructional activity; no learning gain is claimed.

For an Uzbek-speaking classroom, translating an English irony item is only a starting point. A locally meaningful situation and plausible relationship between speakers should be checked by competent users of both languages. MultiPragEval demonstrates the relevance of multilingual assessment but does not cover Uzbek [10]. The present review offers no locally validated performance estimates or classroom survey data. Collecting such evidence would be a useful next step, with particular attention to how learners distinguish figurative intent from unfamiliar vocabulary.

The analysis has further limits. It is selective, predominantly English-oriented, and based on heterogeneous publications. Historical classification scores cannot describe all systems available at the time of writing. The constructed examples isolate particular interpretive features and do not reproduce the full complexity of

everyday conversation. Finally, the proposed framework has face validity as an analytical tool but lacks demonstrated reliability or predictive validity. These limits bound the contribution to synthesis, transparent statistical description, and an explicit evaluation proposal.

Conclusion and recommendations

The central challenge of metaphor, irony, and context in AI-generated discourse is the passage from a plausible reading to an appropriate communicative response. The reviewed benchmarks make different parts of this process measurable, but their counts and scores cannot be merged into a universal index of understanding. Recognition, explanatory justification, and response suitability should remain distinguishable in both research reports and educational use.

Future evaluations should preserve decisive context, test matched changes in that context, and examine generated responses alongside classification labels. Researchers should report metrics with their units, task definitions, and annotation targets; educators should require explanations grounded in the supplied situation and allow justified ambiguity. For Uzbek EFL applications, locally reviewed examples and independently rated classroom studies are needed before making claims of effectiveness. Progress is best demonstrated when a system changes its interpretation for a reason that evaluators can identify and produces a response that fits the resulting communicative need.

References

1. Bender E. M., Koller A. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – 2020. – P. 5185–5198. – DOI: 10.18653/v1/2020.acl-main.463.

2. Shutova E. Models of Metaphor in NLP // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. – 2010. – P. 688–697. – URL: https://aclanthology.org/P10-1071/ (accessed: 12.10.2026).

3. Leong C. W. (B.), Beigman Klebanov B., Hamill C., Stemle E., Ubale R., Chen X. A Report on the 2020 VUA and TOEFL Metaphor Detection Shared Task // Proceedings of the Second Workshop on Figurative Language Processing. – 2020. – P. 18–29. – DOI: 10.18653/v1/2020.figlang-1.3.

4. Van Hee C., Lefever E., Hoste V. SemEval-2018 Task 3: Irony Detection in English Tweets // Proceedings of the 12th International Workshop on Semantic Evaluation. – 2018. – P. 39–50. – DOI: 10.18653/v1/S18-1005.

5. Khodak M., Saunshi N., Vodrahalli K. A Large Self-Annotated Corpus for Sarcasm // Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). – 2018. – URL: https://aclanthology.org/L18-1102/ (accessed: 12.10.2026).

6. Oprea S., Magdy W. iSarcasm: A Dataset of Intended Sarcasm // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. – 2020. – P. 1279–1289. – DOI: 10.18653/v1/2020.acl-main.118.

7. Chakrabarty T., Saakyan A., Ghosh D., Muresan S. FLUTE: Figurative Language Understanding through Textual Explanations // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. – 2022. – P. 7139–7159. – DOI: 10.18653/v1/2022.emnlp-main.481.

8. Saakyan A., Chakrabarty T., Ghosh D., Muresan S. A Report on the FigLang 2022 Shared Task on Understanding Figurative Language // Proceedings of the 3rd Workshop on Figurative Language Processing. – 2022. – P. 178–183. – DOI: 10.18653/v1/2022.flp-1.26.

9. Sravanthi S., Doshi M., Tankala P., Murthy R., Dabre R., Bhattacharyya P. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities // Findings of the Association for Computational Linguistics: ACL 2024. – 2024. – P. 12075–12097. – DOI: 10.18653/v1/2024.findings-acl.719.

10. Park D., Lee J., Park S., Jeong H., Koo Y., Hwang S., Park S., Lee S. MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models // Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP. – 2024. – P. 96–119. – DOI: 10.18653/v1/2024.genbench-1.7.

11. Castro S., Hazarika D., Pérez-Rosas V., Zimmermann R., Mihalcea R., Poria S. Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper) // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – 2019. – P. 4619–4629. – DOI: 10.18653/v1/P19-1455.

12. Liu N. F., Lin K., Hewitt J., Paranjape A., Bevilacqua M., Petroni F., Liang P. Lost in the Middle: How Language Models Use Long Contexts // Transactions of the Association for Computational Linguistics. – 2024. – Vol. 12. – P. 157–173. – DOI: 10.1162/tacl_a_00638.