Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2025"

April 23–25, 2025

Evaluating the Pragmatic Competence of Large Language Models in Detecting Mitigated and Unmitigated Types of Disagreement

Valery Shulginov
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices,
HSE University / Moscow, Russia
MIPT / Moscow, Russia
shulginov.va@mipt.ru

Hasan Berkcan Şimşek
Centre for Sociocultural and
Ethnolinguistic Studies,
HSE University / Moscow, Russia
hasanberkcansimsek@gmail.com

Sergei Kudriashov
HSE University / Moscow, Russia
xenomirant@gmail.com

Renata Randautsova
Laboratory of Linguistic Conflict
Resolution Studies and Contemporary
Communicative Practices,
HSE University / Moscow, Russia
randovtsovarn@gmail.com

Sofya A. Shevela
HSE University / Moscow, Russia
MIPT / Moscow, Russia
sofyashevela@gmail.com

Abstract

This study presents a framework for evaluating the effectiveness of language models (LLMs) in detecting disagreement across a wide range of pragmatic strategies, from mitigated forms to overt verbal aggression. Special attention is given to complex cases of implicit manifestations of irony and sarcasm, which pose significant challenges for both automated analysis and interpersonal communication. Experimental testing of LLMs was conducted in two types of tasks: binary classification for identifying disagreement and classification of specific strategies for its expression. The results showed that large multilingual models outperformed other models, especially in binary classification. However, models that focus primarily on the Russian language, such as GigaChat and YaGPT, tend to interpret irony and sarcasm more accurately and have a higher result density. Comparative analysis with human judgments revealed that, despite progress, the accuracy of sarcasm detection by LLMs still lags significantly behind human judgments. The results suggest a need for further optimization of LLMs to improve their pragmatic competence in real communicative situations.

Keywords: LLMs evaluation; benchmark; pragmatics; disagreement; irony; sarcasm; QUD
DOI: 10.28995/2075-7182-2025-23-345-360

Оценка прагматической компетенции больших языковых моделей в выявлении смягченных и несмягченных типов несогласия

Валерий Шульгинов
Лаборатория лингвистической
конфликтологии и современных
коммуникативных практик,
НИУ ВШЭ / Москва, Россия
МФТИ / Москва, Россия
shulginov.va@mipt.ru

Хасан Беркджан Шимшек
Центр социокультурных и
этноязыковых исследований
факультета гуманитарных наук,
НИУ ВШЭ / Москва, Россия
hasanberkcansimsek@gmail.com

Сергей Кудряшов
НИУ ВШЭ/ Москва, Россия
xenomirant@gmail.com

Рената Рандовцова
Лаборатория лингвистической
конфликтологии и современных
коммуникативных практик,
НИУ ВШЭ / Москва, Россия
randovtsovarn@gmail.com

Софья Шевела
НИУ ВШЭ / Москва, Россия
МФТИ / Москва, Россия
sofyashevela@gmail.com

Аннотация

В настоящем исследовании представлена методика оценки эффективности больших языковых моделей (БЯМ) в области выявления несогласия, включая широкий диапазон стратегий его выражения — от смягченных форм до явной вербальной агрессии. Особое внимание уделяется сложным случаям имплицитных проявлений иронии и сарказма, представляющим значительные трудности как для автоматического анализа, так и для межличностного общения. Экспериментальное тестирование БЯМ проводилось в двух типах задач: бинарная классификация для идентификации несогласия и классификация конкретных стратегий его выражения. Результаты показали, что большие мультиязычные модели демонстрируют преимущество над другими моделями, особенно в рамках бинарной классификации. Тем не менее БЯМ, ориентированные преимущественно на русский язык, например, GigaChat и YaGPT, склонны более точно интерпретировать иронию и сарказм и характеризуются более высокой плотностью результатов. Сравнительный анализ с оценками ассесоров показал, что несмотря на достигнутый прогресс, точность определения сарказма у БЯМ по-прежнему существенно уступает человеческим оценкам. Результаты исследования указывают на необходимость дальнейшей оптимизации БЯМ для повышения их прагматической компетенции в реальных коммуникативных ситуациях.

Ключевые слова: оценка БЯМ; бенчмарк; прагматика; несогласие; ирония; сарказм; вопрос над дискуссией

1 Introduction

In recent years, there has been an accelerated advancement in artificial intelligence systems and large language models (LLMs). They are increasingly employed in tasks involving direct human interaction through chatbots, search engines, and web browsers (Min et al., 2023; Sravanthi et al., 2024). The rapid integration of LLMs into everyday life raises critical questions regarding the assessment of these models' communicative skills, particularly their ability to interpret not only the literal and direct meaning of an utterance, but also those expressed more implicitly through various pragmatic strategies.

In this context, a pivotal skill warranting thorough examination is pragmatic competence, which refers to the ability to understand and use language effectively in specific social situations (see Taguchi, 2009). Previous studies have assessed individual pragmatic skills such as the understanding of implicatures and presuppositions (Qi et al., 2023; Hu et al., 2023; Ruis et al., 2023) and the interpretation of figurative language (Tong et al., 2021; Liu et al., 2022; Gu et al., 2022). In a large-scale evaluation, the Pragmatic Understanding Benchmark (PUB) (Sravanthi et al., 2024) is introduced, which is composed of a total of fourteen tasks focusing on four major pragmatic phenomena: implicature, presupposition, reference, and deixis. As a result, aggregators of pragmatic tasks are formed, allowing for a comprehensive study of LLM competencies.

In this vein, we ground our approach in politeness theory, which posits that adult members of society navigate social interactions to maintain 'face'—their desired public self-image. As Brown and Levinson (1978/1987, p. 61) note, "people cooperate (and assume each other's cooperation) in maintaining face in interaction, such cooperation being based on the mutual vulnerability of face". Face consists of a positive aspect (the need to be accepted, respected, or valued by others) and a negative aspect (the desire for autonomy and freedom from imposition). In communication, various politeness strategies are employed to mitigate threats to these face needs, such as avoiding direct demands (protecting negative face) or offering compliments (reinforcing positive face).

Within this framework, we center our analysis on disagreement, as it exemplifies a speech act that can often trigger confrontation. In some contexts, disagreement is "consensual" (Bolander & Locher, 2017, p. 608). It is a natural and expected response that can enhance social interaction (Schiffrin, 1984), especially when it is preferred or permitted (Kakava, 2002). However, disagreement often "does not leave us cold," as it evokes negative emotions such as annoyance, anger, or contempt, and can lead to conflict (Langlotz & Locher, 2012, p. 1591). In its most basic form, disagreement involves the expression of an opposing opinion without intending negativity, using straightforward and unmitigated phrases that convey a contrary stance, such as 'I don't agree', 'I don't think so', or simply 'no', which can be termed direct disagreement. However, disagreement can also be expressed in many other ways. Understanding and recognizing different strategies for expressing disagreement can enable models to interact more effectively with users and to anticipate the direction of communication.

Assessing the degree of consensus and conflict in disagreement requires considering both the speaker's verbal behavior and an understanding of normative elements, relationships between parties, and community-level expectations (Angouri & Locher, 2012; Bolander & Locher, 2017). In this paper, we limit our analysis to verbal strategies used in disagreement, without addressing extra-linguistic cues related to consensus and conflict. Our goal is to evaluate how effectively different LLMs detect disagreement at the linguistic level, particularly when encountering its various forms of expression. To achieve this, we develop a nuanced benchmark that captures the distinct linguistic strategies used in disagreement.

2 Various Types of Disagreement

As noted by Benz and Jasinskaja (2017), each sentence in discourse typically responds to a (often implicit) Question Under Discussion (QUD), either by providing an answer or by introducing another question that helps in addressing the original one. The way a sentence is structured and interpreted is often influenced by the QUD it engages with. Similarly, each turn in a conflictual sequence addresses a specific QUD – whether directly answering the existing one or shifting it. For example, when Speaker A makes an arguable statement in Turn 1, they implicitly raise a QUD that Speaker B may either agree or disagree with in Turn 2. The structure of disagreement is shaped by the evolving QUDs, with each turn potentially redefining the scope of the issue at hand. While Turn 3 and subsequent turns are also important for the development of potential conflict episodes, Turns 1 and 2 constitute the minimum necessary sequence for the realization of a disagreement. Therefore, the two-turn structure forms the foundational pattern essential for LLMs to identify disagreement and its potential progression into conflict. For this reason, this study limits its analysis to the two-turn structure of disagreement, focusing on how the QUD initiated in Turn 1 is responded to in Turn 2.

Linguistic strategies in disagreement can range from unmitigated to mitigated (Kakava, 1993a, as cited in Kakava, 2002). Unmitigated disagreement is a direct and explicit expression of opposing viewpoints, with little or no attempt to soften the statement, making it potentially face-threatening. Choosing unmitigated disagreement can reflect a desire to be rude, disruptive, or hurtful (Locher, 2004), but it can also occur in contexts of consensual disagreement, where the primary concern is to engage in debate and defend one's opinions (e.g., Kotthoff, 1993). On the other hand, mitigated types of disagreements involve using politeness strategies to soften the potential face-threatening effect of the disagreement.

To develop a typology of disagreement strategies suitable for benchmarking, we first conducted a systematic literature review focusing on studies that propose taxonomies of disagreement and prioritize comprehensive classification within their respective contexts (e.g., Culpeper, 1996; Muntigl & Turnbull, 1998; Locher, 2004; Shum & Lee, 2013). The aim was to synthesize different patterns of disagreement and establish a consolidated framework reflecting well-documented disagreement types. This approach prioritizes cross-contextual robustness to support benchmarking applications, though we recognize that domain-specific adaptations and culturally situated disagreement practices may extend beyond the scope of this typology. Disagreement types used in the benchmark are outlined below (see Appendix 1 for two-turn examples for each type):

Polite disagreement. A person mitigates their speech to avoid or reduce the face-threatening effect. Mitigation strategies include using hedges (e.g., 'well', 'I think'), modal auxiliaries (e.g., 'may', 'could'), emphasizing subjectivity through personal reasoning (e.g., 'it's just hard for me') (Locher, 2004, pp. 114-133), or concessive constructions such as 'yes, but…' (Uzelgun et al., 2015). Additionally, one can offer an alternative claim without directly opposing (cf. *counterclaim*, Muntigl & Turnbull, 1998), express regret, be less direct (Leech, 1983), or frame disagreement as questions (Locher, 2004).

Interrogative Disagreement. A person uses interrogatives to challenge a claim, usually by demanding evidence and implying that the other party is unable to do so (cf. *challenge*, Muntigl & Turnbull, 1998; *objections in the form of a question*, Locher, 2004; *raising rhetorical questions*, Shum & Lee, 2013) through statements like "What do you know about it?" (Langlotz & Locher, 2012, p. 1594).

Repetitive Disagreement. A person shows disagreement by repeating or reconstructing previous comments. Repetition can emphasize negative attitudes and increase imposition on the listener, boosting impoliteness (Bousfield, 2008, see also Locher, 2004, p. 123). It helps sustain and "anchor" disagreement (Kakava, 2002, p. 1560).

Referential disagreement. A person bolsters their disagreement by drawing on personal experiences and/or presenting external evidence, such as expert opinions or statistics (cf. *giving personal experience* and *giving facts*, Shum & Lee, 2013).

Irrelevance-Based Disagreement. A person communicates to another through a meta-dispute-act that their previous statement is not relevant to the current discussion, using expressions such as "It doesn't matter" or "You're straying off topic" (Muntigl & Turnbull, 1998, p. 229). This type of disagreement challenges the social skill of making relevant claims (Langlotz & Locher, 2012).

Reprimand and Profanity. Strong disapproval is expressed through reprimands or profanity, fostering conflict (cf. *using short vulgar phrases* and *reprimanding, giving negative comments*, Shum & Lee, 2013). This involves asserting one's stance as the correct one, while accusing the other party in a personalized manner, typically using pronouns like 'I' and 'you/your' to cast them as wrongdoers.

Ironic Disagreement. An ironic remark passes a negative judgment, but it directs the disagreement not at the interlocutor but at the situation as a whole, does not involve a face-threatening act, is not aggressive, and can be open to interpretation (Witek 2022). For instance, the statement "What nice weather!" can be made when it is windy and pouring rain.

Sarcastic Disagreement. A person makes a sarcastic remark in response to another person's statement to express disagreement (cf. *sarcasm or mock politeness*, Culpeper, 1996). It is a face-threatening act "with the use of politeness strategies that are obviously insincere" (Culpeper, 1996, p. 356). Sarcasm is perceived as more deliberate than irony, as it typically has a target, is more aggressive and offensive, and is conveyed in a cutting tone that is rarely equivocal (Lee, 1998; Reyes et al., 2013; Aniruddha et al., 2015; Sykora et al., 2020).

3 Methodology

Our dataset involved eight categories of tasks that encompass various strategies for expressing disagreement, with each category represented by a set of 50 assignments. We evaluated the capabilities of several LLMs using our benchmark in a zero-shot setting. Prompts consisted of two parts: a) the task description with output options provided as numbers or distinct single words, b) a minimal communication context for model analysis and evaluation. The configuration of the test dataset was as follows:

Polite Disagreement (PolD) involved responses where disagreement is mitigated through politeness strategies, with the objective for LLMs to identify phrases containing disagreement.

Interrogative Disagreement (InterD) featured responses in the form of questions, where only some express disagreement, requiring identification of such intent.

Repetitive Disagreement (RepD) included consistently rephrased responses, with only a portion indicating disagreement, aiming to identify these instances.

Referential Disagreement (RefD) encompassed responses containing references to documents, past experiences, or general knowledge, with some expressing disagreement, necessitating LLMs to identify disagreement.

Irrelevance-Based Disagreement (IrrBD) involved responses expressing disagreement with the QUD or indicating its irrelevance, requiring classification by LLMs.

Mitigated, Direct, Reprimand, and Profanity (MDRP) always contained disagreement, which was either mitigated, direct, or reprimanding, with the objective to categorize disagreement.

Ironic Disagreement (IronyD) always contained disagreement, which can be ironic or direct, necessitating identification of irony.

Sarcastic Disagreement (SarcD) always contained disagreement, which was either sarcastic or direct, requiring LLMs to identify sarcasm.

The dataset was divided into two parts: (1) tasks that required a binary choice between the presence and absence of disagreement (Pold, InterD, RepD, RefD), and (2) tasks where disagreement was present but required correct classification (IrrBD, MDRP, IronyD, SarcD).

The models under investigation included *Claude 3.5 Sonnet*, *Gemini Pro*, *GPT-4o*, *Llama-3 405B* (Grattafiori et al., 2024), *Mistral Nemo*, *GigaChat-Pro*, *YandexGPT-Pro*, and *Saiga Llama-3 70B*. Models were prompted via the respective APIs with default generation parameters and *temperature set to zero* in order to ensure greedy generation of specified answers. We are aware of the potential compromises of this approach, as models' performance may be subject to change dependent on various context and prompt parameters. However, as we are more interested in evaluating discriminative abilities of the model at identifying certain pragmatic patterns, and various sampling strategies, despite their ability to provide more thorough investigation of real-world performance, tend to get lower scores on most discriminative tasks (Song et al., 2024), we have chosen to lean towards deterministic generation with a prompt fixed for all the models. As some APIs don't provide access to output logits, the answers were evaluated in a generative approach by parsing them for the correct output structure and target answer structure, specified in the prompt. Also, as the tasks in the benchmark required either binary classification or ternary classification (for one task) with balanced class distributions, we measured accuracy as the generalization score and normalized Levenshtein similarity as the measure of ability to follow the provided instructions and produce structured output. Testing was conducted in September 2024 with the latest model versions provided via API.

To evaluate the consistency between the LLMs and human judgments, sarcasm identification, a particularly interesting and debatable task, was selected. A survey was conducted using tasks from the benchmark, where 11 evaluators assessed whether the provided context contained a sarcastic disagreement or not across 50 samples. The evaluators, mostly university students fluent in Russian, were presented with the same task description used for zero-shot LLM prompting.

4 Results and Analysis

4.1 Model evaluation

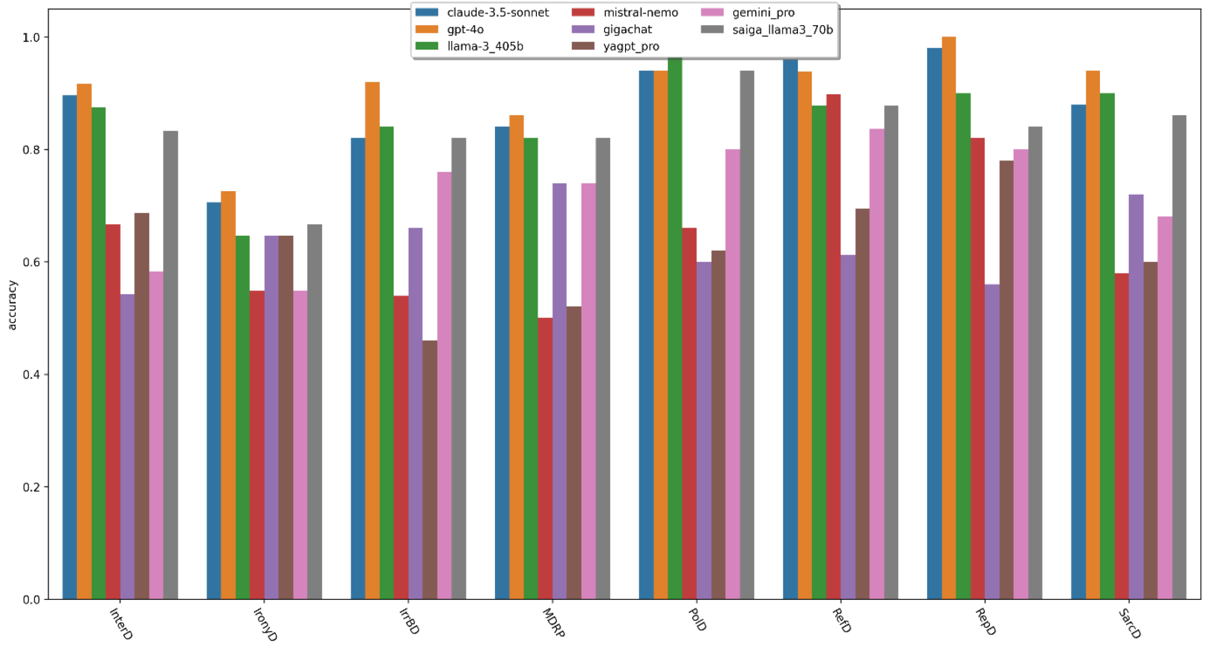
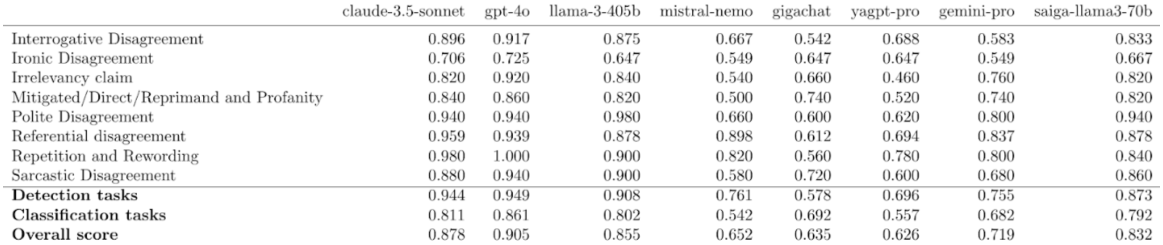


Figure 1: Overall accuracy per task



	claude-3.5-sonnet	gpt-4o	llama-3-405b	mistral-nemo	gigachat	yagpt-pro	gemini-pro	saiga-llama3-70b
Interrogative Disagreement	0.896	0.917	0.875	0.667	0.542	0.688	0.583	0.833
Ironic Disagreement	0.706	0.725	0.647	0.549	0.647	0.647	0.549	0.667
Irrelevancy claim	0.820	0.920	0.840	0.540	0.660	0.460	0.760	0.820
Mitigated/Direct/Reprimand and Profanity	0.840	0.860	0.820	0.500	0.740	0.520	0.740	0.820
Polite Disagreement	0.940	0.940	0.980	0.660	0.600	0.620	0.800	0.940
Referential disagreement	0.959	0.939	0.878	0.898	0.612	0.694	0.837	0.878
Repetition and Rewording	0.980	1.000	0.900	0.820	0.560	0.780	0.800	0.840
Sarcastic Disagreement	0.880	0.940	0.900	0.580	0.720	0.600	0.680	0.860
Detection tasks	0.944	0.949	0.908	0.761	0.578	0.696	0.755	0.873
Classification tasks	0.811	0.861	0.802	0.542	0.692	0.557	0.682	0.792
Overall score	0.878	0.905	0.855	0.652	0.635	0.626	0.719	0.832

Table 1: Accuracy of evaluated models

As previously mentioned, our dataset evaluated two types of skills: 1) binary choice between the presence and absence of disagreement, and 2) classification of disagreement strategies. As shown in Table 1, multilingual LLMs (Claude 3.5, GPT-4o, Llama-3, Gemini-Pro) consistently outperform in solving each class of tasks. Fine-tuning a pre-trained model on Russian data (Saiga-Llama-3) yields similar results. In contrast, a model primarily oriented towards the Russian language (GigaChat) shows significantly better results in the task of disagreement classification and achieves the highest scores in the MDDP task (Figure 1), which is one of the most challenging for other models. Another distinguishing feature of models focused on the Russian language (GigaChat, YaGPT) is their enhanced ability to accurately identify ironic disagreement: they demonstrate comparable results with other tasks. For other models, the performance decline in detecting irony can reach 20-25% below the average across all tasks.

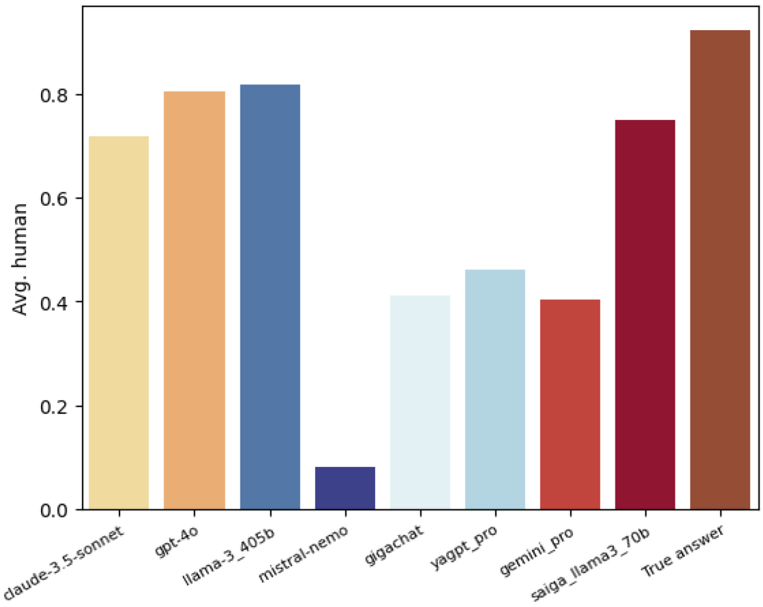


Figure 2: Correlation between model prediction and avg. human judgement

We selected the SarcD task set for comparative evaluation using annotators. It should be noted that sarcasm detection proved to be an easier task for the models than irony detection – all LLMs, except YaGPT, showed higher accuracy in solving this type of task. We aggregated the scores of all experts and used the resulting class to measure the correlation between the average human annotator, our target class, and the model scores (see Figure 2). The results indicate that the accuracy in sarcasm detection, even among leading large language models (Claude 3.5, GPT-4o, Llama-3, Saiga-Llama-3), is significantly lower than the responses of annotators. This is remarkable as it shows that the benchmark is capable of identifying common pragmatic beliefs, which is crucial to provide positive user experience and ensure alignment of these models with human intentions.

4.2 Levenshtein distance evaluation

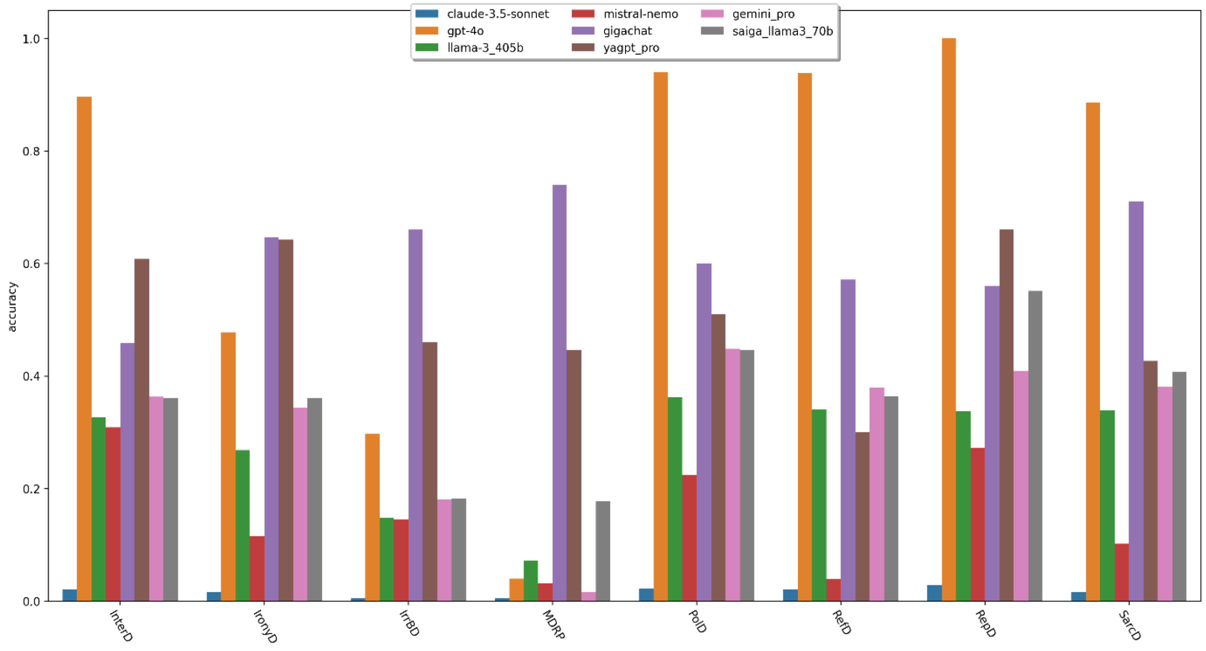


Figure 3: Overall normalized Levenshtein similarity per task

	claude-3.5-sonnet	gpt-4o	llama-3-405b	mistral-nemo	gigachat	yagpt-pro	gemini-pro	saiga-llama3-70b
Interrogative Disagreement	0.020	0.896	0.326	0.309	0.458	0.608	0.363	0.361
Ironic Disagreement	0.016	0.478	0.268	0.115	0.647	0.642	0.343	0.360
Irrelevancy claim	0.005	0.298	0.148	0.145	0.660	0.460	0.181	0.182
Mitigated/Direct/Reprimand and Profanity	0.005	0.040	0.072	0.031	0.740	0.446	0.016	0.177
Polite Disagreement	0.022	0.940	0.362	0.224	0.600	0.510	0.449	0.446
Referential disagreement	0.020	0.939	0.341	0.040	0.571	0.301	0.379	0.364
Repetition and Rewording	0.028	1.	0.337	0.273	0.560	0.661	0.409	0.551
Sarcastic Disagreement	0.016	0.886	0.339	0.102	0.710	0.427	0.380	0.408
Detection tasks	0.023	0.944	0.342	0.212	0.547	0.520	0.400	0.430
Classification tasks	0.010	0.426	0.207	0.098	0.689	0.494	0.230	0.282
Overall score	0.016	0.685	0.274	0.155	0.618	0.507	0.315	0.356

Table 2: Normalized levenshtein similarity of evaluated models

The evaluation based on Levenshtein distance (Fig. 3) exhibited a similar pattern with minor variations. Our aim was to assess whether the model could adhere to instructions directly without requiring additional explanations or arguments. If the results closely align with accuracy, it indicates that the model strictly follows the prompt, as the discrepancy between the expected answer and the model's output is minimal or nonexistent. Conversely, if accuracy is high, but the Levenshtein score is low (particularly evident in models like Claude-3.5, which tend to elaborate on the reasoning process), it suggests that while the model provides the correct answer, it also includes explanations or additional formatting that were removed during our post-processing. If both accuracy and Levenshtein scores are low, it typically indicates a fundamental misunderstanding of the task by the model, as observed with models like Mistral-Nemo.

4.3 Bias Detection

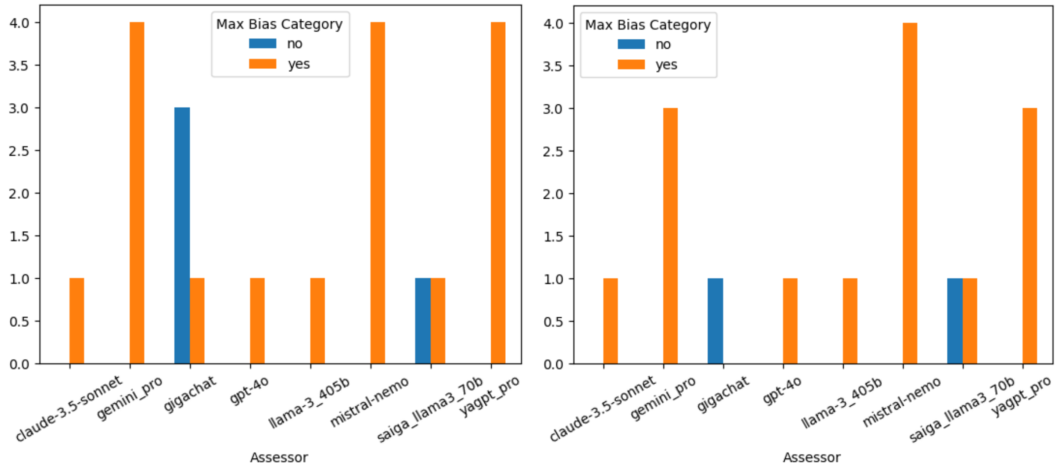


Figure 4: a) Number of tasks where stat. sign. biases were observed
b) Number of tasks where stat. sign. biases were observed when the sample contains negative particles. Max. possible – 6.

We conducted a statistical analysis to test whether the models we evaluated have biases towards detecting or neglecting disagreement. Thus, we constrained the analysis to the subset of tasks involving disagreement detection with a binary answer and conducted a chi-squared test with the initial hypothesis being the uniform distribution as each task contained an equal number of affirmative and negative answers. The graph in Fig. 4a represents the number of tasks where statistically significant biases were detected. Interestingly, most of the models tend to respond to the prompts affirmatively. Moreover, only GigaChat and Saiga models appear to display sensitivity to the task in this manner and display different biases. Overall, all the models display some level of bias, at least on one of the benchmarks.

We also conducted the same test to measure the extent to which negative particles affect the biases of models. Thus, we extracted a subset of samples that included negative particles in the context and applied the test to those. The results are provided in Fig. 4b. Contrary to possible expectations, the outputs for these subsets of samples don't differ much from the whole benchmark tasks and still contain statistically significant deviations towards negative answers only for the same models as in the general setting.

4.4 Error Analysis

The conducted study identified communicative contexts in which the majority of LLMs in our dataset exhibited erroneous results.

Firstly, these responses often express the speaker’s uncertainty. They may include clarifying questions aimed at elucidating the situation, yet they support the speaker's intention and do not contradict it. For example, errors were observed in situations requiring clarification of a bus stop name, with 6 out of 8 language models making mistakes. Similarly, 6 out of 8 LLMs made errors in discussions about the benefits of dairy products. Furthermore, all LLMs failed in clarifying the size of clothing demonstrated by the interlocutor.

Secondly, the responses are often expressive reactions that involve a negative evaluation of the interlocutor’s proposition, yet without an explicit intention to disagree. For example:

- Lately, restaurants have started serving beetles too [V posledneye vremya nachali v restoranakh i zhukov podavat].
- Has exoticism really gotten to this point? [Neuzheli ekzotika i do takogo doshla?] (6 out of 8 LLMs made errors)

In this instance, it can be hypothesized that the perception of disagreement was prompted by the presence of the lexeme ‘neuzheli’ (‘really’), along with the association of emotional communication with negative speech behavior.

Finally, the tasks for classifying direct and ironic disagreement caused difficulties. In most cases, the models showed a tendency to identify irony even in the absence of irony in the context. For example, the response ‘Ya by tak ne skazal’ (‘I wouldn’t say that’) was classified as ironic in 7 out of 8 LLMs. In the study sample, there were three cases in which all models made a mistake, five cases in which seven models made an error, and another five cases in which six models gave an incorrect estimate.

5 Conclusions

The study underscores the significance of assessing pragmatic competence in LLMs, focusing on their ability to detect and classify disagreements in communication. Evaluating LLMs within the framework of politeness theory highlights the way models interpret direct and indirect opposing viewpoints, alongside their potential implications for human interaction. Disagreement in communication manifests itself across a spectrum, ranging from polite and mitigated forms to unmitigated and potentially offensive ones, such as those involving reprimands, profanity, or sarcasm. This study's benchmarking approach therefore aimed to capture these linguistic complexities while examining the models’ capacity to navigate them effectively.

While the models tested – including multilingual models and those fine-tuned for Russian – demonstrated varying levels of success across tasks, several consistent patterns emerged. Multilingual models like Claude 3.5 and GPT-4o, as well as fine-tuned ones like Saiga-Llama-3, performed well overall, but language-specific models such as GigaChat displayed particular strengths in nuanced tasks like detecting ironic disagreement. However, irony detection proved challenging for most models, showcasing their limitations in understanding subtle pragmatic cues.

The observed discrepancies between model performance and human judgments further demonstrate the challenges inherent in teaching artificial systems to interpret elements of pragmatics. For instance, sarcasm detection, although it can be seen as a more straightforward task than irony detection, revealed significant gaps in the models’ capacities when compared to human annotators. Additionally, error analysis highlighted situations where most LLMs faltered, particularly in distinguishing disagreement from mere uncertainty or emotional evaluations. This reflects the complexity of conversational pragmatics, where context, tone, and cultural nuances play key roles – areas in which LLMs often lack nuanced understanding.

Importantly, biases were also identified in models’ responses, such as a tendency to favor affirmative answers or misinterpret context containing negative particles. These biases, while varying across models, highlight the ongoing need for refinement in LLM training, particularly in ensuring balanced output sensitivity and improved detection of linguistic subtleties.

References

[1] Aaron Grattafiori et al. The Llama 3 Herd of Models // arXiv preprint. — 2024. — Vol. arXiv:2407.21783. — Access mode: https://arxiv.org/abs/2407.21783.

[2] Angouri J., Locher M. A. Theorising disagreement // Journal of Pragmatics. — 2012. — Vol. 44, No. 12. — P. 1549–1553. Access mode: https://doi.org/10.1016/j.pragma.2012.06.011.

[3] Benz A., Jasinskaja K. Questions Under Discussion: From Sentence to Discourse // Discourse Processes. — 2017. — Vol. 54, No. 3. — P. 177–186. Access mode: https://doi.org/10.1080/0163853x.2017.1316038.

[4] Bolander B., Locher M. A. Conflictual and consensual disagreement // In: Hoffmann C., Bublitz W. (Eds.). Pragmatics of Social Media. — De Gruyter Mouton, 2017. — P. 607–632. Access mode: https://doi.org/10.1515/9783110431070-022.

[5] Bousfield D. Impoliteness in Interaction. — John Benjamins Publishing Company, 2008.

[6] Brown P., Levinson S. C. Politeness: Some universals in language usage. — Cambridge: Cambridge University Press, 1987. — (Original work published 1978).

[7] Culpeper J. Towards an anatomy of impoliteness // Journal of Pragmatics. — 1996. — Vol. 25, No. 3. — P. 349–367. Access mode: https://doi.org/10.1016/0378-2166(95)00014-3.

[8] Ghosh Aniruddha, Li Guofu, Veale Tony, Rosso Paolo, Shutova Ekaterina, Barnden John, Reyes Antonio. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). — Denver, Colorado, 2015. — P. 470–478.

[9] Gu Yuling, Fu Yao, Pyatkin Valentina, Magnusson Ian, Dalvi Mishra Bhavana, Clark Peter. Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE // Proceedings of the 3rd Workshop on Figurative Language Processing (FLP). — Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022. — P. 84–93.

[10] Kakava C. Opposition in modern Greek discourse: Cultural and contextual constraints // Journal of Pragmatics. — 2002. — Vol. 34, No. 10–11. — P. 1537–1568. Access mode: https://doi.org/10.1016/s0378-2166(02)00075-9.

[11] Kotthoff H. Disagreement and concession in disputes: On the context sensitivity of preference structures // Language in Society. — 1993. — Vol. 22, No. 2. — P. 193–216. Access mode: https://doi.org/10.1017/s0047404500017103.

[12] Langlotz A., Locher M. A. Ways of communicating emotional stance in online disagreements // Journal of Pragmatics. — 2012. — Vol. 44, No. 12. — P. 1591–1606. Access mode: https://doi.org/10.1016/j.pragma.2012.04.002.

[13] Lee C. J., Katz A. N. The differential role of ridicule in sarcasm and irony // Metaphor and Symbol. — 1998. — Vol. 13, No. 1. — P. 1–15.

[14] Leech G. N. Principles of pragmatics. — Longman, 1983.

[15] Liu Emmy, Cui Chenxuan, Zheng Kenneth, Neubig Graham. Testing the Ability of Language Models to Interpret Figurative Language // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Seattle, United States, 2022. — P. 4437–4452.

[16] Locher M. A. Power and politeness in action: Disagreement in oral communication. — Mouton de Gruyter, 2004.

[17] Min B., Ross H., Sulem E., Veyseh A. P. B., Nguyen T. H., Sainz O., Agirre E., Heintz I., Roth D. Recent advances in natural language processing via large pre-trained language models: A survey // ACM Computing Surveys. — 2023. — Vol. 56, No. 2. — P. 2. Access mode: https://doi.org/10.1145/3605943.

[18] Muntigl P., Turnbull W. Conversational structure and facework in arguing // Journal of Pragmatics. — 1998. — Vol. 29, No. 3. — P. 225–256. Access mode: https://doi.org/10.1016/s0378-2166(97)00048-9.

[19] Qi Chengwen, Li Bowen, Hui Binyuan, Wang Bailin, Li Jinyang, Wu Jinwang, Laili Yuanjun. An Investigation of LLMs' Inefficacy in Understanding Converse Relations // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. — Singapore: Association for Computational Linguistics, 2023. — P. 6932–6953.

[20] Reyes A., Rosso P., Veale T. A multidimensional approach for detecting irony in Twitter // Language Resources and Evaluation. — 2013. — Vol. 47, No. 1. — P. 239–268.

[21] Ruis L. et al. The Goldilocks of Pragmatic Understanding: Fine-tuning Strategy Matters for Implicature Resolution by LLMs // Advances in Neural Information Processing Systems. — 2024. — Vol. 36.

[22] Schiffrin D. Jewish argument as sociability // Language in Society. — 1984. — Vol. 13, No. 3. — P. 311–335. Access mode: https://doi.org/10.1017/s0047404500010526.

[23] Shum W., Lee C. (Im)politeness and disagreement in two Hong Kong internet discussion forums // Journal of Pragmatics. — 2013. — Vol. 50, No. 1. — P. 52–83. Access mode: https://doi.org/10.1016/j.pragma.2013.01.010.

[24] 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. — 2024. — P. 12075–12097. Access mode: https://doi.org/10.18653/v1/2024.findings-acl.719.

[25] Song Y., Wang G., Li S, & Lin B. Y. (2024). The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism. *arXiv preprint arXiv: 2407.10457.*

[26] Sykora M., Elayan S., Jackson T. W. A qualitative analysis of sarcasm, irony and related #hashtags on Twitter // Big Data & Society. — 2020. — Vol. 7, No. 2. Access mode: https://doi.org/10.1177/2053951720972735.

[27] Taguchi N. Pragmatic competence in Japanese as a second language: An introduction // In: Taguchi N. (Ed.). Pragmatic Competence. — Mouton de Gruyter, 2009. — P. 1–18.

[28] Tong Xiaoyu, Shutova Ekaterina, Lewis Martha. Recent advances in neural metaphor processing: A linguistic, cognitive, and social perspective // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Online: Association for Computational Linguistics, 2021. — P. 4673–4686.

[29] Uzelgun M. A., Mohammed D., Lewiński M., Castro P. Managing disagreement through yes, but… constructions: An argumentative analysis // Discourse Studies. — 2015. — Vol. 17, No. 4. — P. 467–484. Access mode: https://doi.org/10.1177/1461445615578965.

[30] Witek M. Irony as a speech action // Journal of Pragmatics. — 2022. — Vol. 190. — P. 76–90. Access mode: https://doi.org/10.1016/j.pragma.2022.01.010.

[31] Yao Binwei, Jiang Ming, Bobinac Tara, Yang Diyi, Hu Junjie. Benchmarking Machine Translation with Cultural Awareness // Findings of the Association for Computational Linguistics: EMNLP 2024. — Miami, United States, 2024. — pp. 13078–13096. Access mode:

6 Appendix A. Representative task examples

6.1 Direct disagreement

Example 1

A: Onlajn-obuchenie tak zhe e'ffektivno, kak tradicionnoe obuchenie v klasse, blagodarya svoej gibkosti. (Online learning is just as effective as traditional classroom learning because of the flexibility it offers.)

B: Ya tak ne dumayu. Otsutstvie ochnogo vzaimodejstviya uxudshaet process obucheniya. (I don’t think so. The lack of face-to-face interaction hinders the learning experience.)

Example 2

A: Arenda doma luchshe, potomu chto ona pozvolyaet se'konomit' na remonte i nalogax na imushhestvo. (Renting a house is better because it saves so much money in repairs and property taxes.)

B: No vladenie domom e'konomit bol'she deneg so vremenem blagodarya nakopleniyu kapitala i vozmozhny'm nalogovy'm vy'chetam. (But, owning a house saves more money over time through equity buildup and potential tax deduction.)

Example 3

A: Mne kazhetsya, nado kupit' novyj noutbuk, etot uzhe ele rabotaet. (I think we need to buy a new laptop; this one is barely working.)

B: Nichego podobnogo, etot nout eshchyo paru let protyanet. (No way! This one’s got a couple more years in it.)

Example 4

A: Ty smotrel novyj sezon seriala? (Have you seen the new season of the show?)

B: Net, poka eshche ne dobralsya do nego. (Not yet, haven’t gotten around to it.)

6.2 Polite disagreement

Example 1 (With disagreement)

A: Tvoi reis vyletaet v 2 chasa dnya, ne opozdai (Your flight departs at 2 PM, don't be late).

B: Poleznaya informatsiya, no pokhozhe, chto on uzhe uletel (Useful information, but it seems it has already taken off)

Example 2 (With disagreement)

A: Ya s vami ne soglasen. Domashnie zhivotny'e prinosyat stol'ko radosti i uyuta v dom! Ne mogu predstavit' svoyu zhizn' bez svoego pushistogo druga. (I disagree. Pets bring us so much joy and comfort! I can’t imagine life without my fluffy friend.)

B: Ya ne soglasen s vashim mneniem. Konechno, domashnie zhivotny'e mogut prinosit' radost', no e'to ne dlya vsex. U menya est' allergiya na sherst', poe'tomu ya predpochitayu ne imet' domashnix pitomcev. (I don’t subscribe to your opinion. Pets can make people happier, sure, but they aren’t for everyone. I’m allergic to fur so I wouldn’t want to have a pet.)

Example 3 (With disagreement)

A: Poezd tol'ko v 5 vechera, ty uspevaesh'. (The train is only at 5 PM, you’ve got plenty of time.)

B: A ty uveren, chto 5 vechera, a ne v 7? (Are you sure it’s 5 PM and not 7?)

Example 4 (Without disagreement)

A: Vot vam stakan vody, etu tabletku tak prosto ne proglotit' (Here's a glass of water; this pill isn't easy to swallow)

B: Aga, spasibo bol'shoe. Razlomit' zhe ee tozhe mozhno, da? (Oh, thanks a lot. I can break it in half too, right?)

Example 5 (Without disagreement)

A: Sovsem chto-to kust zasyhaet. Nado ego peresadit' v ten', navernoe... (This bush looks like it's drying out. Maybe we should move it to the shade?)

B: Togda, kak dumaesh', mozhet i elku peresadit'? (In that case, do you think we should move the spruce too?)

6.3 Interrogative Disagreement

Example 1 (With disagreement)

A: Nam nuzhno perestat' ispol'zovat' odnorazovye plastikovye izdeliya, oni zagryaznyayut okean (We need to stop using single-use plastic products; they pollute the ocean)

B: I chto ot etogo izmenitsya? (And what difference will that make?)

Example 2 (With disagreement)

A: Ya ne samy'j bol'shoj puteshestvennik, no kazhdoe puteshestvie prinosit mne neveroyatny'e vpechatleniya i novy'e znakomstva. (I don’t travel that often but each time I do I get absolutely incredible experience and make new friends.)

B: Puteshestviya takzhe mogut by't' neveroyatno stressovy'mi i iznuritel'ny'mi. No pochemu by' ne provesti otpusk v uyute doma, e'konomya pri e'tom i vremya, i den'gi? (Travelling can also be incredibly stressful and tiring. Why wouldn’t you spend your vacation in the comfort of your home, saving both time and money?)

Example 3 (With disagreement)

A: Luchshe vsekh sobak chipirovat', ne tol'ko bol'shih. (All dogs should be chipped, not just big ones.)

B: Ty s kakoj planety? Chto tebe chihuahua mozhet sdelat'? (What planet are you from? What harm could a chihuahua do to you?)

Example 4 (Without disagreement)

A: Ya vzyal nozhi i vilki, no kazhetsya chto-to zabyl… (I took the knives and forks, but I seem to have forgotten something...)

B: A kak zhe tarelki? (What about the plates?)

Example 5 (Without disagreement)

A: Ya 5 let prouchilsya v universitete i ochen' skuchayu po etomu vremeni. (I studied at university for five years and really miss that time.)

B: V kakom imenno? (Which one exactly?)

6.4 Repetitive Disagreement

Example 1 (With disagreement)

A: V takom sluchae stoit snachala razmorozit' kuritsu (In that case, you should defrost the chicken first)

B: V takom sluchae stoit snachala podumat', nuzhen li mne vash sovet. (In that case, you should first consider whether I need your advice)

Example 2 (With disagreement)

A: My vchera ne zakryvali dver', esli ya pravil'no pomnyu. (We didn’t lock the door last night, if I remember correctly.)

B: Aga-aga, vizhu ya, kak vy ne zakryvali dver'. (Yeah, right, I can totally see how you didn’t lock the door.)

Example 3 (Without disagreement)

A: Zavtra ozhidaetsya dozhdlivyy den', ne zabud' vzat' zont (Tomorrow is expected to be a rainy day, don't forget to take an umbrella)

B: Vzat' zont. (Take an umbrella)

Example 4 (Without disagreement)

A: Nado kupit' hleb ili ris na uzhin. (We need to buy either bread or rice for dinner.)

B: Tak hleb ili ris? (So which one—bread or rice?)

6.5 Referential disagreement

Example 1 (With disagreement)

A: Ideya universitetov kazhetsya slishkom uzhe ustarevshey (The idea of universities seems a bit too outdated)

B: Moy papa – professor, on smotrit na eto sovershenno inache. (My dad is a professor; he sees it completely differently).

Example 2 (With disagreement)

A: Plavanie — samoe nelepoe olimpijskoe sobytie. (Swimming is the most ridiculous Olympic event.)

B: Plavanie — eto zdorovo. Ya govoryu eto kak chelovek, kotoryj zanimalsya sportivnym plavaniem bolee 15 let. (Swimming is great. I say this as someone who did competitive swimming for over 15 years.)

Example 3 (Without disagreement)

A: Nichego ne uspel kupit' v podarok. Kak dumaesh', emu ponravitsya kniga? (I didn’t manage to buy a gift. Do you think he’d like a book?)

B: Govoryat, chto kniga – luchshij podarok! (They say a book is the best gift!)

Example 4 (Without disagreement)

A: Ya kupil sol' i perec, no ne mogu najti v paketah. (I bought salt and pepper, but I can't find them in the bags.)

B: V cheke oni est', davaj iskat'. (They’re on the receipt, so let’s keep looking.)

6.6 Irrelevancy claim

Example 1 (Disagreement)

A: Ty pojdesh' na novuyu vystavku? (Are you going to the new exhibition?)

B: Net, eto slishkom daleko ot doma. (No, it’s too far from home.)

Example 2 (Disagreement)

A: Kak ty dumaesh', nam stoit zavesti sobaku? (Do you think we should get a dog?)

B: Mozhet, koshku? (How about a cat?)

Example 3 (Refusal to engage with the topic)

A: Kstati, ya sobirayus' kupit' novye krossovki. (By the way, I'm going to buy new sneakers.)

B: I kak eto pomozhet otremontirovat' mashinu? (And how will that help fix the car?)

Example 4 (Refusal to engage with the topic)

A: Povyshenie minimal'noj zarplaty privedet k rostu bezraboticy. (Raising the minimum wage will lead to higher unemployment.)

B: Vopros bezraboticy ne imeet otnosheniya k obsuzhdeniyu. My govorim o prave rabotnikov na dostojnuyu oplatu truda. (Unemployment isn’t the issue here. We’re talking about workers’ right to fair pay.)

6.7 Mitigated/Direct/Reprimand and Profanity

Example 1 (Mitigated disagreement)

A: Ochen' lyublyu Dzhuliyu Roberts! (I really love Julia Roberts!)

B: Slushaj, ne mogu tut soglasit'sya, mne ona, k sozhaleniyu, ni v odnom fil'me ne ponravilas'. (Honestly, I can’t agree with you here; unfortunately, I haven’t liked her in any movie)

Example 2 (Mitigated disagreement)

A: Davaj pouzhinaem v etom novom restorane, govoryat, tam vkusno. (Let’s have dinner at that new restaurant, I’ve heard the food is good.)

B: V celom ya podderzhivayu, no mozhet, snachala pochitaem otzyvy? (I’m on board, but maybe we should check the reviews first?)

Example 3 (Direct disagreement)

A: Kakoe horoshee utro! (What a lovely morning!)

B: Net, nu kakoe zhe ono horoshee? Chetyre utra. (Lovely? How is it lovely? It’s 4 a.m.)

Example 4 (Direct disagreement)

A: Nado obnovit' garderob, eti veshchi uzhe ne v mode. (I need to update my wardrobe, these clothes are out of fashion.)

B: Ne dumayu, zachem tratit' den'gi. (I don’t think so, why spend money?)

Example 5 (Direct disagreement and Reprimand/Profanity)

A: Kakoye vsyo-taki zamechatel'noye utro! (What a wonderful morning it is!)

B: Slushay, idi ty k chertu. Utro u nego zamechatel'noye... (Listen, go to hell. Wonderful morning, he says...)

Example 6 (Direct disagreement and Reprimand/Profanity)

A: Ya vse v mashine ostavil. Hochesh' – idi sam posmotri. (I left everything in the car. If you want, go check yourself.)

B: Ty menya za duraka derzhish'? Slushaj, ded ty staryj, ya svoimi dvumya glazami videl, kak ty papku v shkaf polozhil! (Are you seriously that dumb? Listen, I saw with my own eyes how you put the folder in the cupboard!)

6.8 Ironic disagreement

Example 1 (Disagreement with irony)

A: Tebe ponravilsya novyy restoran? (Did you like the new restaurant?)

B: O, da, osobenno yego unikal'nyy servis s podachey kholodnoy yedy. (Oh, yes, especially its unique service of serving cold food.)

Example 2 (Disagreement with irony)

A: Mne vse ravno, chto anglijskij – yazyk mezhdunarodnogo obshcheniya, ya ne sobirayus' ego uchit'. (I don't care that English is the international language of communication, I'm not going to learn it.)

B: Chem bol'she takih prekrasnyh mnenij, tem men'she konkurenciya na rynke. Spasibo! (The more of these wonderful opinions, the less competition in the market. Thanks!)

Example 3 (Disagreement with irony)

A: Eta vecherinka byla prosto potryasayushchej, ne tak li? (This party was amazing, wasn’t it?)

B: Absolyutno! Osobenno chast', gde vse utknulis' v svoi telefony. (Absolutely! Especially the part where everyone was glued to their phones.)

Example 4 (Disagreement without irony)

A: Vy ushli ot otveta. (You dodged the question.)

B: Ya? (Me?)

Example 5 (Disagreement without irony)

A: My zapuskaem novyj eko-frendli proekt. (We are launching a new eco-friendly project.)

B: Podobnaya iniciativa ni k chemu ne privedet. (Such an initiative will lead to nothing.)

6.9 Sarcastic disagreement

Example 1 (Disagreement with sarcasm)

A: A pochemu vy protiv? chto on takogo sdelal? (Why are you against? What did he do?)

B: Glupo etogo ne znat' =))) no ya podskazhu unikal'nyj sposob dlya prosveshcheniya: set' INTERNET + poiskovik = otvet na vopros. (It’s silly not to know =))) but here’s a unique way to educate yourself: the INTERNET + a search engine = your answer.)

Example 2 (Disagreement with sarcasm)

A: Pomoch' tebe s domashnej rabotoj po matematike? (Do you need help with your math homework?)

B: Uchitel' goda ob"yavilsya! (Here comes the Teacher of the Year!)

Example 3 (Disagreement with sarcasm)

A: Ya uveren, chto vegetarianstvo – luchshee, chto mozhno pridumat' dlya zdorov'ya. (I'm sure vegetarianism is the best thing for health.)

B: Kak zdorovo, chto ty tak verish' v eti chudesa. Tak derzhat'))) (How wonderful that you believe these miracles. Keep it up!)))

Example 4 (Disagreement without sarcasm)

A: Vy skazali, chto deficit vitamina D vedet k vospaleniyu desen. (You said that a vitamin D deficiency leads to gum inflammation.)

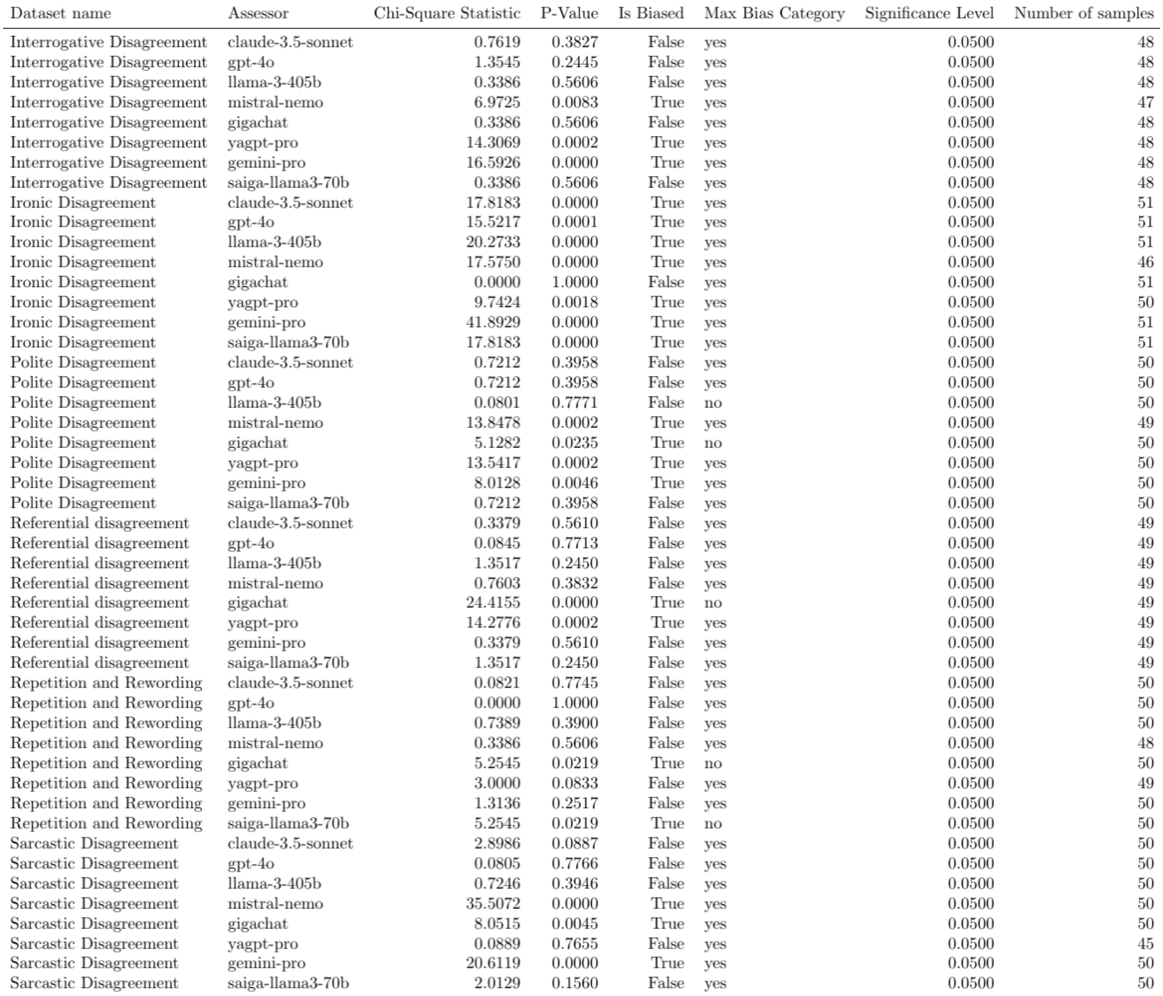
B: Gde ya takoe govoril..? (When did I say that...?)

Example 5 (Disagreement without sarcasm)

A: Vtoraya chast' "Vlastelina kolec" samaya luchshaya. Mneniya? (The second part of "The Lord of the Rings" is the best. Thoughts?)

B: Eeee, chto? Ty, navernoe, hotel skazat' "tret'ya". (Uhh, what? You probably meant "the third.")

7 Appendix B. Significance analysis



Dataset name	Assessor	Chi-Square Statistic	P-Value	Is Biased	Max Bias Category	Significance Level	Number of samples
Interrogative Disagreement	claude-3.5-sonnet	0.7619	0.3827	False	yes	0.0500	48
Interrogative Disagreement	gpt-4o	1.3545	0.2445	False	yes	0.0500	48
Interrogative Disagreement	llama-3-405b	0.3386	0.5606	False	yes	0.0500	48
Interrogative Disagreement	mistral-nemo	6.9725	0.0083	True	yes	0.0500	47
Interrogative Disagreement	gigachat	0.3386	0.5606	False	yes	0.0500	48
Interrogative Disagreement	yagpt-pro	14.3069	0.0002	True	yes	0.0500	48
Interrogative Disagreement	gemini-pro	16.5926	0.0000	True	yes	0.0500	48
Interrogative Disagreement	saiga-llama3-70b	0.3386	0.5606	False	yes	0.0500	48
Ironic Disagreement	claude-3.5-sonnet	17.8183	0.0000	True	yes	0.0500	51
Ironic Disagreement	gpt-4o	15.5217	0.0001	True	yes	0.0500	51
Ironic Disagreement	llama-3-405b	20.2733	0.0000	True	yes	0.0500	51
Ironic Disagreement	mistral-nemo	17.5750	0.0000	True	yes	0.0500	46
Ironic Disagreement	gigachat	0.0000	1.0000	False	yes	0.0500	51
Ironic Disagreement	yagpt-pro	9.7424	0.0018	True	yes	0.0500	50
Ironic Disagreement	gemini-pro	41.8929	0.0000	True	yes	0.0500	51
Ironic Disagreement	saiga-llama3-70b	17.8183	0.0000	True	yes	0.0500	51
Polite Disagreement	claude-3.5-sonnet	0.7212	0.3958	False	yes	0.0500	50
Polite Disagreement	gpt-4o	0.7212	0.3958	False	yes	0.0500	50
Polite Disagreement	llama-3-405b	0.0801	0.7771	False	no	0.0500	50
Polite Disagreement	mistral-nemo	13.8478	0.0002	True	yes	0.0500	49
Polite Disagreement	gigachat	5.1282	0.0235	True	no	0.0500	50
Polite Disagreement	yagpt-pro	13.5417	0.0002	True	yes	0.0500	50
Polite Disagreement	gemini-pro	8.0128	0.0046	True	yes	0.0500	50
Polite Disagreement	saiga-llama3-70b	0.7212	0.3958	False	yes	0.0500	50
Referential disagreement	claude-3.5-sonnet	0.3379	0.5610	False	yes	0.0500	49
Referential disagreement	gpt-4o	0.0845	0.7713	False	yes	0.0500	49
Referential disagreement	llama-3-405b	1.3517	0.2450	False	yes	0.0500	49
Referential disagreement	mistral-nemo	0.7603	0.3832	False	yes	0.0500	49
Referential disagreement	gigachat	24.4155	0.0000	True	no	0.0500	49
Referential disagreement	yagpt-pro	14.2776	0.0002	True	yes	0.0500	49
Referential disagreement	gemini-pro	0.3379	0.5610	False	yes	0.0500	49
Referential disagreement	saiga-llama3-70b	1.3517	0.2450	False	yes	0.0500	49
Repetition and Rewording	claude-3.5-sonnet	0.0821	0.7745	False	yes	0.0500	50
Repetition and Rewording	gpt-4o	0.0000	1.0000	False	yes	0.0500	50
Repetition and Rewording	llama-3-405b	0.7389	0.3900	False	yes	0.0500	50
Repetition and Rewording	mistral-nemo	0.3386	0.5606	False	yes	0.0500	48
Repetition and Rewording	gigachat	5.2545	0.0219	True	no	0.0500	50
Repetition and Rewording	yagpt-pro	3.0000	0.0833	False	yes	0.0500	49
Repetition and Rewording	gemini-pro	1.3136	0.2517	False	yes	0.0500	50
Repetition and Rewording	saiga-llama3-70b	5.2545	0.0219	True	no	0.0500	50
Sarcastic Disagreement	claude-3.5-sonnet	2.8986	0.0887	False	yes	0.0500	50
Sarcastic Disagreement	gpt-4o	0.0805	0.7766	False	yes	0.0500	50
Sarcastic Disagreement	llama-3-405b	0.7246	0.3946	False	yes	0.0500	50
Sarcastic Disagreement	mistral-nemo	35.5072	0.0000	True	yes	0.0500	50
Sarcastic Disagreement	gigachat	8.0515	0.0045	True	yes	0.0500	50
Sarcastic Disagreement	yagpt-pro	0.0889	0.7655	False	yes	0.0500	45
Sarcastic Disagreement	gemini-pro	20.6119	0.0000	True	yes	0.0500	50
Sarcastic Disagreement	saiga-llama3-70b	2.0129	0.1560	False	yes	0.0500	50

Table 3. Significance analysis of bias for the models

Dataset name	Assessor	Chi-Square Statistic	P-Value	Is Biased	Max Bias Category	Significance Level	Number of samples
Interrogative Disagreement	claude-3.5-sonnet	1.3333	0.2482	False	yes	0.0500	16
Interrogative Disagreement	gpt-4o	1.3333	0.2482	False	yes	0.0500	16
Interrogative Disagreement	llama-3-405b	0.0000	1.0000	False	yes	0.0500	16
Interrogative Disagreement	mistral-nemo	5.4545	0.0195	True	yes	0.0500	15
Interrogative Disagreement	gigachat	1.3333	0.2482	False	yes	0.0500	16
Interrogative Disagreement	yagpt-pro	3.0000	0.0833	False	yes	0.0500	16
Interrogative Disagreement	gemini-pro	3.0000	0.0833	False	yes	0.0500	16
Interrogative Disagreement	saiga-llama3-70b	0.0000	1.0000	False	yes	0.0500	16
Ironic Disagreement	claude-3.5-sonnet	4.7009	0.0301	True	yes	0.0500	22
Ironic Disagreement	gpt-4o	6.7692	0.0093	True	yes	0.0500	22
Ironic Disagreement	llama-3-405b	9.2137	0.0024	True	yes	0.0500	22
Ironic Disagreement	mistral-nemo	12.4444	0.0004	True	yes	0.0500	21
Ironic Disagreement	gigachat	0.1880	0.6646	False	yes	0.0500	22
Ironic Disagreement	yagpt-pro	4.7009	0.0301	True	yes	0.0500	22
Ironic Disagreement	gemini-pro	15.2308	0.0001	True	yes	0.0500	22
Ironic Disagreement	saiga-llama3-70b	6.7692	0.0093	True	yes	0.0500	22
Polite Disagreement	claude-3.5-sonnet	0.4706	0.4927	False	yes	0.0500	34
Polite Disagreement	gpt-4o	1.0588	0.3035	False	yes	0.0500	34
Polite Disagreement	llama-3-405b	0.1176	0.7316	False	no	0.0500	34
Polite Disagreement	mistral-nemo	12.1324	0.0005	True	yes	0.0500	33
Polite Disagreement	gigachat	1.0588	0.3035	False	no	0.0500	34
Polite Disagreement	yagpt-pro	7.5294	0.0061	True	yes	0.0500	34
Polite Disagreement	gemini-pro	4.2353	0.0396	True	yes	0.0500	34
Polite Disagreement	saiga-llama3-70b	1.0588	0.3035	False	yes	0.0500	34
Referential disagreement	claude-3.5-sonnet	0.0000	1.0000	False	yes	0.0500	23
Referential disagreement	gpt-4o	0.0000	1.0000	False	yes	0.0500	23
Referential disagreement	llama-3-405b	0.1825	0.6692	False	yes	0.0500	23
Referential disagreement	mistral-nemo	2.9206	0.0875	False	yes	0.0500	23
Referential disagreement	gigachat	8.9444	0.0028	True	no	0.0500	23
Referential disagreement	yagpt-pro	8.9444	0.0028	True	yes	0.0500	23
Referential disagreement	gemini-pro	0.7302	0.3928	False	yes	0.0500	23
Referential disagreement	saiga-llama3-70b	0.1825	0.6692	False	yes	0.0500	23
Repetition and Rewording	claude-3.5-sonnet	0.2017	0.6534	False	yes	0.0500	24
Repetition and Rewording	gpt-4o	0.0000	1.0000	False	yes	0.0500	24
Repetition and Rewording	llama-3-405b	3.2269	0.0724	False	yes	0.0500	24
Repetition and Rewording	mistral-nemo	0.8214	0.3648	False	yes	0.0500	23
Repetition and Rewording	gigachat	3.2269	0.0724	False	yes	0.0500	24
Repetition and Rewording	yagpt-pro	0.8067	0.3691	False	yes	0.0500	24
Repetition and Rewording	gemini-pro	0.2017	0.6534	False	yes	0.0500	24
Repetition and Rewording	saiga-llama3-70b	7.2605	0.0070	True	no	0.0500	24
Sarcastic Disagreement	claude-3.5-sonnet	0.8333	0.3613	False	yes	0.0500	20
Sarcastic Disagreement	gpt-4o	0.0000	1.0000	False	yes	0.0500	20
Sarcastic Disagreement	llama-3-405b	0.2083	0.6481	False	yes	0.0500	20
Sarcastic Disagreement	mistral-nemo	13.3333	0.0003	True	yes	0.0500	20
Sarcastic Disagreement	gigachat	1.8750	0.1709	False	yes	0.0500	20
Sarcastic Disagreement	yagpt-pro	0.2250	0.6353	False	no	0.0500	18
Sarcastic Disagreement	gemini-pro	10.2083	0.0014	True	yes	0.0500	20
Sarcastic Disagreement	saiga-llama3-70b	0.2083	0.6481	False	yes	0.0500	20

Table 4. Significance analysis of bias for the models in samples including negative particles

Appendix C. Model generation parameters

Model	Temperature	Intended generation strategy
Claude-3.5 Sonnet	0.0	Greedy
GPT-4o	0.0	Greedy
Llama-3 405B	0.0	Greedy
Mistral Nemo	0.0	Greedy
Gigachat Pro	0.0	Greedy
Yagpt Pro	0.0	Greedy
Gemini Pro	0.0	Greedy
Saiga Llama3 70B	0.0	Greedy