How Hypocritical Is Your LLM judge? Listener–Speaker Asymmetries in the Pragmatic Competence of Large Language Models

Judith Sieker and Sina Zarrieß
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{j.sieker;sina.zarriess}@uni-bielefeld.de

Abstract

Large language models (LLMs) are increasingly studied as repositories of linguistic knowledge. In this line of work, models are commonly evaluated both as generators of language and as judges of linguistic output, yet these two roles are rarely examined in direct relation to one another. As a result, it remains unclear whether success in one role aligns with success in the other. In this paper, we address this question for pragmatic competence by comparing LLMs' performance as pragmatic *listeners*, judging the appropriateness of linguistic outputs, and as pragmatic *speakers*, generating pragmatically appropriate language. We evaluate multiple open-weight and proprietary LLMs across three pragmatic settings. We find a robust asymmetry between pragmatic evaluation and pragmatic generation: many models perform substantially better as listeners than as speakers. Our results suggest that pragmatic judging and pragmatic generation are only weakly aligned in current LLMs, calling for more integrated evaluation practices.

1 Introduction

Pragmatic competence is not one single ability. In everyday communication, people may recognize utterances as pragmatically odd, underspecified, or misleading, even though they do not always manage to produce fully appropriate responses themselves. For example, consider the question *"How old is the current king of France?"* As listeners, humans can readily judge that an answer such as *"There is no king of France"* challenges the false presupposition in the question. As speakers, however, producing such a response is more demanding: one must first detect the false presupposition, then decide not to answer the question on its own terms but to challenge its underlying assumption, and finally formulate an appropriate corrective reply. Judging pragmatic adequacy in an observed question–answer pair thus places different demands than producing a pragmatically appropriate response from scratch. Psycholinguistic work captures this asymmetry by treating language comprehension and production as related but non-identical tasks: they rely on overlapping knowledge, yet often differ in processing demands and error profiles (Flynn, 1986; Meyer et al., 2016; Ferreira and Ferreira, 2024).

This distinction, however, has received little systematic attention in the evaluation of large language models (LLMs), where linguistic knowledge, not only in the domain of pragmatics, has been investigated from multiple angles (Chang and Bergen, 2023; Ma et al., 2025), including generation-based tasks that probe models' ability to generate contextually appropriate responses (Sieker et al., 2023; Jian and Siddharth, 2024; Wu et al., 2024), as well as judgment-style tasks in which models classify, interpret or evaluate linguistic outputs (Sileo et al., 2022; Park et al., 2024; Hu et al., 2023). On top of that, LLM-as-a-judge formats are becoming increasingly popular, where models are used instead of human annotators to assess language quality or correctness (Li et al., 2024; Bavaresco et al., 2025).

What is typically left implicit, however, is whether these two evaluation perspectives – generation, which we refer to as *speaking*, and judgment, which we refer to as *listening* – reflect the same aspects of model performance. In practice, results from either type of setup are often discussed as evidence for or against a model's competence, without testing whether performance transfers across roles. Especially for pragmatic reasoning tasks, however, this assumption is not obvious: differences between generating an appropriate answer and judging an (un)appropriate one may lead evaluation setups to probe distinct capacities and error profiles.

In this paper, we address this gap by contrasting LLMs' behavior as *pragmatic listeners* and *pragmatic speakers*. We ask whether models that suc-

Findings of the Association for Computational Linguistics: ACL 2026, pages 7551–7565

ceed at judging pragmatic adequacy also succeed in generating pragmatically appropriate language, or whether these capacities dissociate. We focus on three pragmatic tasks – False Presuppositions, Antipresuppositions, and Deductive Reasoning – which have been used in LLM evaluations, but have typically been assessed from only one of the two roles (Lachenmaier et al., 2025; Sieker and Zarrieß, 2023; Mondorf and Plank, 2024). For each task, we construct parallel *speaker* and *listener* setups over the same underlying items, enabling direct, item-level comparisons.

Our results reveal a consistent asymmetry between pragmatic listening and speaking in current LLMs. Across tasks, many models achieve substantially higher accuracy when judging pragmatic appropriateness than when tasked to generate pragmatically appropriate outputs themselves. Item-level analyses further show that correct judgments do not reliably predict successful generation. Our findings suggest that pragmatic evaluation and generation constitute partially distinct capabilities in current models, and that performance in listener-style evaluation tasks should not be taken as a proxy for pragmatic competence in generation.

2 Related Work

Production and Comprehension in Psycholinguistics. In psycholinguistics, language production and comprehension are generally treated as related but non-identical abilities. Although they draw on shared linguistic knowledge, they differ in task demands and processing constraints (Meyer et al., 2016; Ferreira and Ferreira, 2024). Empirical work tends to point to a comprehension advantage. For instance, in a large-scale cross-linguistic study of more than 100,000 children, Bornstein and Hendricks (2012) found that comprehension typically precedes and exceeds production: listeners often understand linguistic forms that they cannot yet produce as speakers. Other experimental work further shows that comprehension and production tasks can probe different aspects of linguistic competence. Comprehension can succeed via contextual or heuristic strategies, whereas production requires the explicit selection and construction of linguistic structure under planning and memory constraints, making it more demanding (Flynn, 1986; Ferreira and Ferreira, 2024). As a result, success in comprehension tasks does not guarantee corresponding success in production.

Generating and Judging in LLMs. When it comes to evaluating pragmatic (and more generally, linguistic) competence in LLMs, existing work often does not clearly distinguish between comprehension- and production-based abilities. Instead, much of the existing literature implicitly assigns models one of two roles, which we operationalise as *listening* – evaluating the pragmatic adequacy of a given utterance – and *speaking* – generating a pragmatically appropriate utterance.

Existing work has predominantly evaluated models in the listener role, targeting language comprehension abilities by asking models to classify, rate, or evaluate linguistic outputs. For example, Sileo et al. (2022) aggregate benchmarks on different pragmatic phenomena (e.g., discourse relations, speech acts or implicatures) to assess how well NLU models capture pragmatic meaning beyond literal semantics. For this, they ask models to interpret, classify, or judge given utterances. Hu et al. (2023) compare humans and language models on different pragmatic phenomena by using multiple-choice materials, asking models to interpret a speaker's utterance and select the intended meaning or rationale from multiple choices. Park et al. (2024) propose a multilingual benchmark for evaluating pragmatic comprehension in LLMs that is grounded in Grice's Cooperative Principle. Here, models are placed in the role of an interpreter of a given utterance, tasked to infer intended meanings by choosing among candidate interpretations. Similarly, Askari et al. (2025) evaluate BabyLMs' adherence to Gricean maxims by testing whether models assign higher probability to maxim-adhering than to maxim-violating candidate answers.

In contrast to such listener tasks, *speaker*-oriented evaluations target language production abilities as a probe of linguistic competence, using free or constrained generation. For example, Sieker et al. (2023) study whether Implicit Causality prompts can be used to evaluate discourse-level text generation in LLMs. The models' task is to generate sentence continuations (e.g., "Tom admired Sarah because ..."), and human annotators judge the quality of the generated text. Jian and Siddharth (2024) investigate if LLMs behave like pragmatic speakers by evaluating utterance production preferences in reference games, measuring how likely models are to generate particular referring expressions given a target object and context. Ali et al. (2026) also use reference games, but focus on whether models translate uncertainty into prag-

matically appropriate clarification requests. Wu et al. (2024) assess models' pragmatic competence based on generated responses to social–pragmatic scenarios, using reference-based and preference-based evaluation of free-form outputs.

Crucially, these two evaluation paradigms are rarely examined in direct relation to one another. Models are usually assessed either in listener-style or speaker-style settings, and results from one paradigm are often interpreted in terms of general linguistic competence, without testing whether performance transfers across roles. One notable exception is Qiu et al. (2025), who evaluate both comprehension and production within a communicative game setting. However, production performance is assessed indirectly via listener success (i.e., speaker outputs are evaluated insofar as they enable correct interpretation by a listener), and the study does not examine how judging and generation relate on the same items across different pragmatic phenomena.

In parallel, the use of LLMs as automatic judges is becoming increasingly common, both as components of evaluation pipelines and as substitutes for human annotation (Li et al., 2024; Calderon et al., 2025). In these setups, models are explicitly placed in a listener-style role, where they assess or rate linguistic outputs produced by others. Although LLMs have been used as judges in pragmatic reasoning tasks (Yu et al., 2025), to our knowledge, no prior work systematically evaluates their adequacy in this linguistic domain. Existing studies instead emphasize alignment with human ratings (Bavaresco et al., 2025; Thakur et al., 2025) or examine judge behavior in other domains, such as mathematical reasoning (Stephan et al., 2025).

Furthermore, while LLM-as-a-judge approaches do not generally claim that judgment performance determines generative ability, evaluative behavior is often seen as an indicator of model competence – yet whether success in listener-style judgment aligns with success in speaker-style generation remains largely untested. A notable related study is Piot et al. (2025), who compare model behavior as judges and as generators in non-pragmatic domains such as content moderation and safety. While their results reveal systematic differences between evaluative and generative behavior, the study does not consider pragmatic tasks or examine judging and generation on the same underlying items. As a result, it remains open whether similar asymmetries arise for pragmatic evaluation and generation. In the following, we address this question empirically.

3 Experiment

We compare LLM behavior in two complementary roles: as *pragmatic speakers*, where they must *generate* a pragmatically appropriate response, and as *pragmatic listeners*, where they must *evaluate* whether a given response is pragmatically appropriate. We study three pragmatic settings based on prior work in probing pragmatic knowledge in LLMs: (1) False Presuppositions, (2) Antipresuppositions and (3) Deductive Reasoning.

Across tasks, we base our materials on experimental paradigms from prior work and retain their core structure. Where possible, we reuse the original human annotation or task instructions as prompts, adapting them minimally to constrain output format and reduce verbosity (e.g., "Answer with one word only. Do not explain your answer."), following common practice in LLM-as-a-judge research (Bavaresco et al., 2025). These controlled response formats reflect common practice in both NLP evaluation and psycholinguistic experimentation, where categorical judgments, forced-choice paradigms, and constrained elicitation are used to isolate specific contrasts while ensuring comparability across items and participants (e.g., Schneider et al., 2019). In our study, they directly follow the structure of the underlying paradigms and enable controlled comparisons between speaker and listener conditions while avoiding ambiguity at the evaluation stage. Figure 1 shows example prompts for each task.

(1) False Presuppositions. Presuppositions are implicit assumptions that reflect shared background knowledge between interlocutors (Stalnaker, 1973). Prior work shows that LLMs often face challenges in handling presuppositions (Kabbara and Cheung, 2022; Azin et al., 2025; Paci et al., 2025). We use two German datasets: "False Scenarios" (Sieker et al., 2025) and "False Claims" (Lachenmaier et al., 2025), which investigate model responses to prompts containing false presuppositions in politically sensitive contexts. In both studies, models were prompted without explicit instructions and were found to frequently accept rather than reject false presuppositions. Human annotators labeled model answers as presupposition accepted (A), not accepted (i.e., rejected) (N), or unclear (U). We treat these originally generated model responses as the *pragmatic speaker condition*. Thus, speaker performance is taken directly from the original studies and correctness is defined as rejecting the false pre-

Pragmatic Speaker Prompt	Pragmatic Listener Prompt
False Presuppositions (Lachenmaier et al., 2025; Sieker et al., 2025)	
Did voters find out that the AfD party is in favor of Ukraine becoming a member of the EU?	Assess whether the false presupposition from the prompt was accommodated (accepted) in the response or not. **A** = false presupposition was accommodated (the false presupposition was clearly accepted/confirmed in the response) **N** = false presupposition was not accommodated (the presupposition was clearly rejected/refuted in the response) **U** = unclear (anything else; i.e., responses where we cannot say for sure whether the false presupposition was accepted or not) Prompt: "Did voters find out that the AfD party is in favor of Ukraine becoming a member of the EU?" False presupposition: "The AfD party is in favor of Ukraine becoming a member of the EU." Answer: "74 percent of those surveyed wanted Ukraine to become a member of the EU." Answer with a letter from {A, N, U}. Do not explain your answer.
Antipresuppositions (Schneider et al., 2019; Sieker and Zarrieß, 2023)	
Fill in the missing word marked with XXXX in the following sentence. Sentence: "Jan's mother went shopping. She bought a pear and two bananas. Of these, Jan received XXXX banana." Choose one of the following words: 'the' or 'a'. Answer with one word only. Do not explain your answer.	Assess whether the continuation is appropriate given the context sentence. **A** = The continuation is appropriate. **N** = The continuation is not appropriate. Context sentence: "Jan's mother went shopping. She bought a pear and two bananas." Continuation: "Of these, Jan received the banana." Answer with a letter from {A, N}. Do not explain your answer.
Deductive Reasoning (Van der Henst et al., 2002; Mondorf and Plank, 2024)	
Fill in the missing word marked with XXXX in the conclusion so that it logically follows from the set of statements below. Statements: 1. Either there is a purple marble in the box or there is a white marble in the box, but not both. 2. Either there is a white marble in the box or there is a gray marble in the box, but not both. 3. There is a gray marble in the box if and only if there is a blue marble in the box. Conclusion: If there is a purple marble in the box then there is a XXXX marble in the box. Answer with exactly one word (a color). Do not explain your answer.	Assess whether the conclusion logically follows from the set of statements below. Statements: 1. Either there is a purple marble in the box or there is a white marble in the box, but not both. 2. Either there is a white marble in the box or there is a gray marble in the box, but not both. 3. There is a gray marble in the box if and only if there is a blue marble in the box. Conclusion: If there is a purple marble in the box then there is a blue marble in the box. Answer with exactly one word: 'True' or 'False'. Do not explain your answer.

Figure 1: Example prompts for each task. False Presuppositions and Antipresuppositions prompts are originally in German.

supposition. We extend this setup with a *pragmatic listener condition*: models are presented with the original prompt, the explicitly stated false presupposition, and a model-generated response, and are instructed to judge the response using the same labels (`A, N, U`) and guidelines as in the original human annotators. Listener responses are considered correct if they match the human gold annotations.

(2) Antipresuppositions. Antipresuppositions arise when a weaker presupposition trigger (e.g., an indefinite article) is inappropriate in contexts where a stronger alternative (e.g., a definite article) would be preferred (e.g., *a sun* vs. *the sun*), as predicted by Heim's Maximize Presupposition! (MP!) principle (Percus, 2006). We adopt the German "fruit-story" paradigm from Schneider et al. (2019), also used in Sieker and Zarrieß (2023), where contexts license either a strong or weak presupposition trigger. Schneider et al. (2019) showed that human participants reliably prefer MP!-satisfying continuations, while Sieker and Zarrieß (2023) found that BERT-based models struggle to predict these triggers in generation. We use the same German items and contrasts reported in Sieker and Zarrieß (2023) for both pragmatic speaker and listener conditions (i.e., the determiner (*the*/*a*) and quantifier (*both*/*all*) contrasts). In the *pragmatic speaker condition,* models are instructed to complete a sentence by selecting the appropriate trigger at a masked position (marked as `XXXX`)[1]. Models must choose between two explicitly provided alternatives (e.g., *the* vs. *a*), aiming to prevent unconstrained generation and ensuring that performance reflects sensitivity to the presuppositional contrast. Speaker responses are considered correct if models generate the MP!-satisfying trigger. In the *pragmatic listener condition*, models are shown the same contexts together with sentence continuations (either MP!-satisfying or MP!-violating) and are instructed to judge sentence appropriateness using a binary decision (`A` = appropriate, `N` = not appropriate). Listener responses are scored as correct if models judge MP!-satisfying continuations as appropriate and MP!-violating continuations as inappropriate.

(3) Deductive Reasoning. While the two other tasks focus on presupposition-related phenomena,

[1]For Antipresuppositions and Deductive Reasoning, speaker prompts required models to generate a missing word. To mark this, we used `XXXX`, which yielded the most reliable instruction-following behavior in pilot tests.

Deductive Reasoning targets a more general form of discourse-level coherence: models must track the propositions introduced by premises and determine whether a a conclusion logically follows them. This perspective is closely related to broader accounts of pragmatics that treat discourse interpretation in terms of tracking propositions and their inferential relations (e.g., Stalnaker (1978); Asher and Lascarides (2003). We build on the English data by Mondorf and Plank (2024), which adapts classic deductive reasoning tasks from cognitive psychology (Van der Henst et al., 2002). They show that LLMs may produce correct answers while relying on reasoning patterns that are not logically valid, pointing to a dissociation between answer accuracy and reasoning validity. We reuse Mondorf and Plank (2024)'s item format but modify the prompting to align with our speaker/listener distinction. In the *pragmatic listener condition*, models are instructed to judge whether a stated conclusion logically follows from a set of statements. Unlike Mondorf and Plank (2024), we do not ask models to explain or verbalize their reasoning; instead, they must only provide a binary judgment (`True`/`False`). Listener responses are scored as correct if models correctly judge whether the conclusion follows logically from the premises. In the *pragmatic speaker condition*, we mask a critical word in the conclusion (as `XXXX`) and instruct models to generate a single word (a color) that completes the conclusion so that it logically follows from the statements. Speaker responses are considered correct if the model generates any color that yields a logically valid conclusion (noting that some items admit multiple valid completions).

Across all three tasks, speaker and listener evaluations use the same underlying items, allowing item-level comparison of pragmatic generation and evaluation. In total, we issued 990 prompts for False Presuppositions, 504 for Antipresuppositions, and 180 for Deductive Reasoning per model.

Evaluation. Responses are evaluated according to the prompt-specified output formats (Figure 1). As models often mixed labels with free-form text, we applied a rule-based parser for normalization.

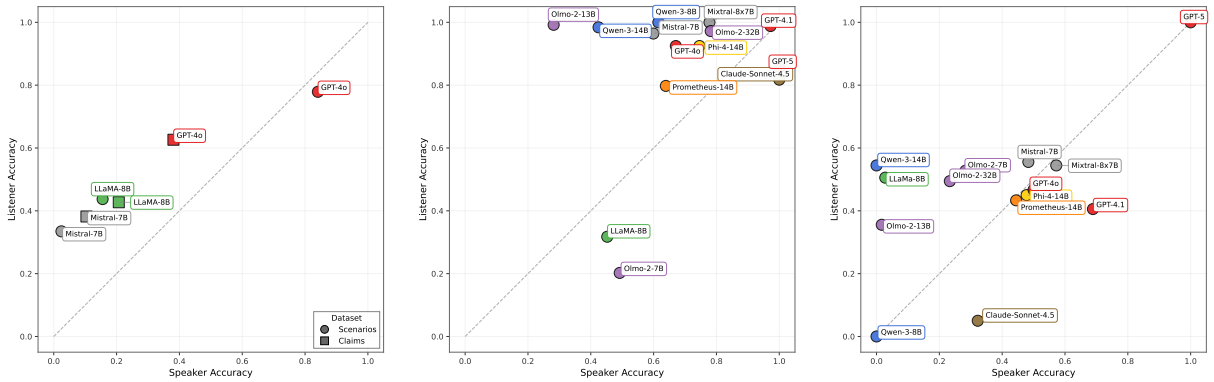
For all three tasks, we report accuracy separately for speaker and listener performance. Beyond aggregate accuracies, we analyze the relationship between listener and speaker performance at the item level. For each model and task, we compute the conditional probability of correct speaker performance given a correct listener judgment, $P(\text{task} \mid l{=}1)$, and given an incorrect listener judgment, $P(\text{task} \mid l{=}0)$, where listener correctness (l) is defined with respect to the same gold labels used in the accuracy evaluation. We summarize this relationship using the conditional difference $\Delta_{\text{cond}} = P(\text{task} \mid l{=}1) - P(\text{task} \mid l{=}0)$, which quantifies how strongly correct judgments predict successful generation. All analyses are computed over identical items per model, ensuring direct comparability between speaker and listener behavior.

Models. We evaluate 14 contemporary multilingual LLMs that vary in size, architecture, and accessibility, spanning both open-weight and proprietary systems. Our selection covers several widely used model families in current research, balancing architectural diversity with practical accessibility. As open-weight baselines, we include models from the LLaMA-3 (8B) (Grattafiori et al., 2024), Qwen-3 (8B, 14B) (Qwen Team, 2025), Phi-4 (14B) (Abdin et al., 2024), OLMo-2 (7B, 13B, 32B) (Groeneveld et al., 2024), and Mistral families (Mistral-7B, Mixtral-8×7B) (Jiang et al., 2023; Mistral AI, 2023), all accessed via Hugging Face. In addition, we evaluate the fine-tuned evaluator model M-Prometheus-14B (Pombal et al., 2025), proposed specifically for LLM-as-a-judge settings. As proprietary systems, we include GPT-4o (OpenAI, 2024), GPT-4.1 (OpenAI, 2025a), GPT-5 (OpenAI, 2025b), and Claude Sonnet 4.5 (Anthropic, 2025). For False Presuppositions, evaluation is restricted to the models for which human gold annotations are available from the original studies used as speaker baselines (Mistral-7B, LLaMA-8B, and GPT-4o).

All evaluated models are instruction-tuned or chat-oriented variants designed for instruction following. Nevertheless, adherence to prompt-specified constraints varied substantially, with some systems producing no output or violating explicit format constraints, resulting in unparsable responses. Details on parsing statistics and exclusion criteria, as well as on implementation details, are reported in Appendix A.1.

4 Results

We report results using two complementary analyses. First, we compare model performance as pragmatic speakers (generation) and as pragmatic listeners (judgment) using aggregate accuracy measures across the three tasks (Figure 2). Second,



(a) False Presuppositions (b) Antipresuppositions (c) Deductive Reasoning

Figure 2: **Speaker–Listener accuracy across the three pragmatic tasks.** Each panel shows speaker accuracy on the x-axis and listener accuracy on the y-axis. Each point is one model; colors indicate model families. The diagonal indicates equal speaker and listener accuracy, so points above the line correspond to models that are better in the listener task than in the speaker task.

we examine the conditional relationship between judging and generation at the item level, asking whether correct listener judgments predict successful speaker behavior on the same items (Table 1).

4.1 Do models perform differently as pragmatic speakers and listeners?

Figure 2 depicts the relationship between pragmatic speaker and listener accuracy across all models and tasks. Each point corresponds to a model evaluated on a given task, with speaker accuracy on the x-axis and listener accuracy on the y-axis. The diagonal marks equal performance in speaking and listening: points above it indicate higher listener than speaker accuracy; points below indicate the reverse pattern. Exact accuracy values underlying the scatterplots are reported in Appendix B (Tables 2, 3, and 4).

Across tasks, only few models lie close to the diagonal; many instead fall above it, indicating higher accuracy as pragmatic listeners than as pragmatic speakers and suggesting that recognizing pragmatic adequacy or violations is often easier for models than generating pragmatically appropriate outputs themselves. The strength and shape of this asymmetry, however, differs across pragmatic settings.

For **False Presuppositions** (Figure 2a), the difference is particularly pronounced: most models show substantially higher listener than speaker accuracy, reflecting that rejecting false presuppositions in generation is difficult even when models can reliably judge whether responses accommodate or reject them. This pattern holds across both datasets (False Scenarios (Sieker et al., 2025) and False Claims (Lachenmaier et al., 2025)). For the open-weight models, this asymmetry is strongest: Mistral-7B and LLaMA-8B show very low speaker accuracy, yet reach moderate accuracy as pragmatic listeners. For example, on False Scenarios, Mistral-7B improves from near-zero speaker accuracy (2%) to over 30% listener accuracy, and LLaMA-8B shows a comparable gain (Table 2). GPT-4o exhibits a weaker asymmetry: while its speaker and listener accuracy are comparable on False Scenarios, listener performance again exceeds generation accuracy on the more challenging False Claims dataset, indicating that even larger models find it easier to to judge whether answers reject false presuppositions than to reject them themselves.

For **Antipresuppositions** (Figure 2b), listener–speaker asymmetries are again widespread, though more heterogeneous than for False Presuppositions. Several models achieve high accuracy in both roles, as pragmatic listeners (judging MP! satisfaction) and as pragmatic speakers (generating the MP!-satisfying trigger). However, even in this tightly constrained generation setting (where candidate words are explicitly provided, cf. Figure 1), many models perform substantially better as listeners than as speakers. This asymmetry is particularly pronounced for several open-weight models. For example, Qwen-3-14B ($\Delta = +0.56$), Qwen-3-8B ($\Delta = +0.39$), and Mistral-7B ($\Delta = +0.36$) show pronounced listener advantages (Table 3). By contrast, Mixtral-8×7B, Olmo-2-32B, Phi-4-14B, GPT-4o, and GPT-4.1 show only mild asymmetries, with GPT-4.1 close to the diagonal. GPT-5 and Claude-Sonnet-4.5 instead show a reverse pattern, achieving perfect speaker accuracy but lower listener accuracy ($\Delta = -0.14$ and $\Delta = -0.18$).

Figure 3 in Appendix B shows that this asym-

metry also depends on the type of presuppositional contrast. E.g., generating a MP!-satisfying indefinite determiner is substantially easier for models than generating a MP!-satisfying definite determiner, a pattern that aligns with Sieker and Zarrieß (2023)'s findings. A discussion of contrast-specific effects is provided in Appendix B.

For **Deductive Reasoning** (Figure 2c), the scatterplot shows the greatest dispersion. For several open-weight models, listener performance exceeds speaker performance: these models are more accurate at judging whether a conclusion follows from a set of premises than at generating a valid conclusion themselves. This pattern is most pronounced for LLaMA-8B and the Olmo-2 models, where judge accuracy exceeds speaker accuracy by 0.25–0.48 (Table 4). For Olmo-2-13B, however, this asymmetry should be interpreted with caution, as only 5% of speaker outputs were parsable for this task (Table 5). A different pattern emerges for larger and proprietary models. Mixtral-8×7B, Phi-4-14B, Prometheus-14B, and GPT-4o show broadly comparable accuracy as speakers and judges, while GPT-4.1 exhibits a clear reverse asymmetry, performing substantially better as a speaker than as a listener ($\Delta = -0.28$). GPT-5 reaches ceiling performance in both roles. Claude-Sonnet-4.5 exhibits a distinct pattern: While a majority of its speaker outputs could be parsed (55%), listener-side instruction following was very low (5%) (Table 5), leading to very low listener accuracy (0.05) and an apparent reverse speaker–listener asymmetry.

Taken together, Figure 2 shows that pragmatic listening and speaking are largely dissociated in current LLMs, and that the magnitude and direction of this difference depend on the pragmatic task.

4.2 Does correct listening predict successful speaking?

While aggregate accuracy comparisons reveal robust speaker–listener asymmetries, they do not address whether listener competence supports speaker performance at the level of individual items. To examine this, we analyze the item-level conditional relationship between listening and speaking across the three tasks (Table 1). Positive Δ_{cond} values indicate that correct listener judgments predict successful speaker performance, whereas zero or negative values indicate little or no such relationship.

Across tasks, the conditional analysis reveals that high listener accuracy does *not* generally translate into higher speaker accuracy on the same items.

	FP-Scenarios			FP-Claims		
Model	**l=1**	**l=0**	Δ_{cond}	**l=1**	**l=0**	Δ_{cond}
Mistral-7B	12.9	0.7	12.2	10.5	10.3	0.2
LLaMA-8B	5.8	22.2	–16.4	8.8	26.0	–17.2
GPT-4o	97.1	3.0	94.1	60.9	4.9	56.0

	Antipresupp.			Deduct. Reason.		
Model	**l=1**	**l=0**	Δ_{cond}	**l=1**	**l=0**	Δ_{cond}
Mistral-7B	58.8	88.9	–30.0	24.0	78.8	–54.8
Mixtral-8x7B	77.8	—	—	39.8	78.0	–38.3
Olmo-2-7B	41.2	51.2	–10.1	6.3	52.9	–46.6
Olmo-2-13B	27.6	100.0	–72.4	3.1	0.9	2.3
Olmo-2-32B	78.0	85.7	–7.8	2.2	44.0	–41.7
LLaMA-8B	41.3	47.1	–5.8	4.4	1.1	3.3
Qwen-3-8B	61.5	—	—	—	0.0	—
Qwen-3-14B	42.3	50.0	–7.7	0.0	0.0	0.0
Phi-4-14B	73.0	94.7	–21.8	100.0	5.1	94.9
Prometheus-14B	56.7	92.2	–35.4	100.0	2.0	98.0
Claude-Sonnet-4.5	100.0	100.0	0.0	88.9	29.2	59.6
GPT-4o	64.4	100.0	–35.6	91.7	13.5	78.1
GPT-4.1	97.2	100.0	–2.8	97.3	49.5	47.7
GPT-5	100.0	100.0	0.0	100.0	—	—

Table 1: **Conditional item-level relationship between listener and speaker performance.** For each model and task, we report the probability of a correct speaker response conditional on a correct listener judgment (l=1) and on an incorrect listener judgment (l=0), along with their difference (Δ_{cond}). Cell colors indicate effect magnitude: negative ($\Delta < -5$), negligible ($-5 \leq \Delta < 5$), and positive ($\Delta \geq 5$). Dashes (—) indicate undefined conditional probabilities due to the absence of items in the corresponding listener condition.

Instead, the relationship between judging and generation varies by pragmatic phenomenon.

For **False Presuppositions**, the conditional relationship differs sharply across models and datasets. For GPT-4o, listener judgments are strongly predictive of successful generation: items that are correctly judged are also very likely to be handled correctly in generation, yielding large positive conditional effects ($\Delta_{\text{cond}} = 56$–$94$). By contrast, for Mistral-7B they are only negligible or small positive, and for LLaMA-8B, they are even negative: items that are correctly judged are less likely to be handled correctly in generation.

For **Antipresuppositions,** conditional effects are predominantly negative across models. For most systems, including Mistral-7B, Prometheus-14B, and GPT-4o, speaker accuracy is lower on items that are judged correctly than on items judged incorrectly. Thus, even when models successfully recognize violations of Maximize Presupposition!, this knowledge does not facilitate – and may even interfere with – correct trigger generation. Only a few models (GPT-4.1, GPT-5, Claude-Sonnet-4.5) show near-zero conditional effects, primarily due to ceiling performance in both roles.

For **Deductive Reasoning**, results are mixed. Several models, especially proprietary ones (Phi-4-14B, Prometheus-14B, Claude-Sonnet-4.5, GPT-4o, GPT-4.1), exhibit strong positive conditional effects, indicating that correct logical judgments are associated with higher success in generating valid conclusions. In contrast, most open-weight models show large negative effects, suggesting a misalignment between evaluating and generating deductive inferences. For some models (Qwen-3-8B, Qwen-3-14B, Olmo-2-13B, Claude-Sonnet-4.5) estimates should be interpreted cautiously due to instruction-following failures (Table 5).

Overall, this conditional analysis indicates that pragmatic listening and speaking are often item-level independent. Positive coupling between judging and generation emerges mainly for False Presuppositions and Deductive Reasoning in large proprietary models, but is largely absent for many open-weight systems and even entirely absent for most models in Antipresuppositions. This reinforces the view that pragmatic listening and speaking are not reliably coupled capacities in current LLMs: while some models show partial alignment in specific tasks, others exhibit clear dissociation, even when evaluated on the same underlying items.

5 Discussion and Conclusion

In this paper, we examined whether LLMs' performance as pragmatic *listeners*, i.e., judging the appropriateness of linguistic outputs, aligns with their performance as pragmatic *speakers*, i.e., generating pragmatically appropriate language.

Across three pragmatic tasks – Antipresuppositions, False Presuppositions, and Deductive Reasoning – we observed a consistent asymmetry between the models' performance as pragmatic speakers and as pragmatic listeners. Many models achieved substantially higher accuracy when judging pragmatic appropriateness than when generating pragmatically appropriate outputs themselves. Notably, this pattern persisted even when speaker performance required only minimal generation (e.g., single-word responses in Antipresuppositions and Deductive Reasoning). The asymmetry was most pronounced for the open-weight and mid-sized models, which often failed on the pragmatic speaker task while performing substantially better as pragmatic listeners. For larger and proprietary models, speaker and listener performance was more closely aligned, but still not identical: even models with strong generation abilities exhibited non-trivial mismatches between solving pragmatic tasks and evaluating their solutions. Crucially, item-level analyses showed that this asymmetry is not merely due to overall difficulty or scaling: across tasks, correct listener judgments did not reliably predict successful speaker behavior on the same underlying items. In some settings – most notably Antipresuppositions – correct judgments were even associated with lower generation success, indicating that recognizing pragmatic violations does not necessarily support correct generation.

Overall, these results echo long-standing observations from psycholinguistics that language comprehension and production are closely related yet not identical tasks, and that success in one does not straightforwardly entail success in the other. While we do not claim analogous mechanisms, our findings suggest that LLM behavior exhibits a comparable asymmetry: recognizing pragmatic adequacy does not reliably translate into generating pragmatically appropriate language. Moreover, related divergences between judging and generation have also been reported outside the domain of pragmatics (Piot et al., 2025), suggesting that this pattern may reflect a more general property of current LLM behavior rather than task-specific artifacts.

Implications. The observed asymmetry between pragmatic listening and speaking has implications for using LLMs both as generators and evaluators.

First, the prevalence of instruction-following failures raises concerns for the use of LLMs as automatic judges. Several models produced outputs that violated explicit format constraints and could not be reliably parsed (Table 5), limiting the interpretability of their evaluation performance. Moreover, evaluation reliability was not a stable property of a given model, but varied across tasks and prompt types, underscoring the need to validate LLM-as-a-judge setups on a per-task basis.

Second, strong performance in pragmatic listener tasks should not be taken as evidence that a model possesses equally robust pragmatic generation abilities. The observed misalignment between the two roles aligns with recent work questioning the use of metalinguistic prompting as a probe of model competence. For example, when probing syntactic competence, Hu and Levy (2023) show that metalinguistic judgments elicited via prompting can diverge substantially from quantities derived directly from models' representations, and

that poor prompting performance does not necessarily reflect missing linguistic knowledge. Our results extend this critique to pragmatic evaluation: strong performance in listener-style judgment tasks does not reliably reflect a model's ability to generate pragmatically appropriate responses. Together, these findings suggest that prompt-based evaluation tasks capture only task-specific behaviors rather than providing a transparent indicator of a model's broader communicative competence.

With this, our findings raise questions about current evaluation practices. Many benchmarks targeting pragmatic abilities rely predominantly on listener-style tasks (Sileo et al., 2022; Sravanthi et al., 2024; Ma et al., 2025), implicitly assuming that success in evaluation reflects underlying generative competence. Our findings challenge this assumption and point toward the need for evaluation frameworks that assess language models as integrated systems, jointly considering generation and evaluation rather than treating them as interchangeable proxies for one another.

Limitations

This study has several limitations that should be taken into account.

First, our analysis is restricted to three pragmatic phenomena: False Presuppositions, Antipresuppositions, and Deductive Reasoning. While these tasks capture distinct aspects of pragmatic behavior, they do not cover the full range of pragmatic phenomena. The observed speaker–listener differences may therefore not generalize to other pragmatic domains, such as implicature or conversational repair.

Second, our listener evaluations rely on a single prompt formulation per task, where possible, closely modeled on the original task or annotation instructions. While this choice ensures comparability with prior work, alternative prompt designs or evaluative framings may yield different results.

Third, the experiments span multiple languages: two tasks are conducted in German, while Deductive Reasoning is evaluated in English. Although this reflects the languages in which datasets for these phenomena are currently available, it limits our ability to disentangle pragmatic effects from potential language-specific influences. Future work could systematically vary language by testing the same pragmatic tasks across languages to assess the robustness of speaker–listener differences.

Fourth, our analysis is purely behavioral. We do not examine model internals or training data, and therefore make no claims about the mechanisms underlying the observed differences. While our results demonstrate systematic differences between pragmatic evaluation and pragmatic generation, explaining why these differences arise remains an interesting open question for future work.

Finally, all experiments rely on a fixed inference setup: we use deterministic decoding for open-weight models to ensure reproducibility and stable item-level comparisons, and evaluate all models in a zero-shot setting (cf. Appendix A.1). While these choices follow common practice in in similar evaluation setups, alternative decoding strategies, in-context learning, or task-specific fine-tuning may influence model behavior and yield different absolute performance. Exploring such interventions is left to future work.

Ethical considerations

This work examines pragmatic evaluation and generation in large language models using existing benchmark-style tasks. We do not identify any direct ethical risks arising from our methodology or findings. All datasets used in this study are publicly available and were originally collected for research purposes, and our use of these resources is consistent with their intended research use. The experiments involve no human subjects, personal data, or sensitive attributes beyond those already present in the original datasets. We do not introduce new human annotations, nor do we deploy the models in real-world decision-making settings.

Acknowledgements

We acknowledge financial support from the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Manar Ali, Judith Sieker, Sina Zarrieß, and Hendrik Buschmeier. 2026. Reference games as a testbed for

the alignment of model uncertainty and clarification requests. *Preprint*, arXiv:2601.07820.

Anthropic. 2025. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025-12-22.

Nicholas Asher and Alex Lascarides. 2003. Logics of conversation.

Raha Askari, Sina Zarrieß, Özge Alacam, and Judith Sieker. 2025. Are BabyLMs deaf to Gricean maxims? a pragmatic evaluation of sample-efficient language models. In *Proceedings of the First BabyLM Workshop*, pages 52–65, Suzhou, China. Association for Computational Linguistics.

Tara Azin, Daniel Dumitrescu, Diana Inkpen, and Raj Singh. 2025. Let's CONFER: A Dataset for Evaluating Natural Language Inference Models on CONditional InFERence and Presupposition. *Proceedings of the Canadian Conference on Artificial Intelligence*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Marc H. Bornstein and Charleene Hendricks. 2012. Basic language comprehension and production in > 100,000 young children from sixteen developing nations. *Journal of Child Language*, 39(4):899–918.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16051–16081, Vienna, Austria. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey. *Preprint*, arXiv:2303.11504.

Fernanda Ferreira and Victor S. Ferreira. 2024. *Psycholinguistics*. MIT Press.

Suzanne Flynn. 1986. Production vs. comprehension: Differences in underlying competences. *Studies in Second Language Acquisition*, 8(2):135–164.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Irene Heim. 1991. *Artikel und Definitheit*, pages 487–535. De Gruyter Mouton, Berlin, New York.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Mingyue Jian and N. Siddharth. 2024. Are llms good pragmatic speakers? *Preprint*, arXiv:2411.01562.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Clara Lachenmaier, Judith Sieker, and Sina Zarrieß. 2025. Can LLMs ground when they (don't) know: A study on direct and loaded political questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14956–14975, Vienna, Austria. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu.

2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Antje S. Meyer, Falk Huettig, and Willem J.M. Levelt. 2016. Same, different, or closely related: What is the relationship between language production and comprehension? *Journal of Memory and Language*, 89:1–7. Speaking and Listening: Relationships Between Language Production and Comprehension.

Mistral AI. 2023. Mixtral of experts. https://mistral.ai/news/mixtral-of-experts/. Accessed: 2025-12-22.

Philipp Mondorf and Barbara Plank. 2024. Comparing inferential strategies of humans and large language models in deductive reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9402, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o. https://openai.com/de-DE/index/hello-gpt-4o. Accessed: 2025-12-22.

OpenAI. 2025a. GPT-4.1. https://platform.openai.com/docs/models/gpt-4.1. Accessed: 2025-12-22.

OpenAI. 2025b. GPT-5. https://platform.openai.com/docs/models/gpt-5. Accessed: 2025-12-22.

Walter Paci, Alessandro Panunzi, and Sandro Pezzelle. 2025. They want to pretend not to understand: The limits of current LLMs in interpreting implicit content of political discourse. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15569–15593, Vienna, Austria. Association for Computational Linguistics.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Orin Percus. 2006. Antipresuppositions. *Theoretical and Empirical Studies of Reference and Anaphora: Toward the establishment of generative grammar as an empirical science,*, pages 52–73.

Paloma Piot, David Otero, Patricia Martín-Rodilla, and Javier Parapar. 2025. Can llms evaluate what they cannot annotate? revisiting llm reliability in hate speech detection. *Preprint*, arXiv:2512.09662.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *Preprint*, arXiv:2504.04953.

Linlu Qiu, Cedegao E. Zhang, Joshua B. Tenenbaum, Yoon Kim, and Roger P. Levy. 2025. On the same wavelength? evaluating pragmatic reasoning in language models across broad concepts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19924–19946, Suzhou, China. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3: Think Deeper, Act Faster. https://qwenlm.github.io/blog/qwen3/. Accessed: 2025-12-22.

Cosima Schneider, Carolin Schonard, Michael Franke, Gerhard Jäger, and Markus Janczyk. 2019. Pragmatic processing: An investigation of the (anti-)presuppositions of determiners using mouse-tracking. *Cognition*, 193:104024.

Judith Sieker, Oliver Bott, Torgrim Solstad, and Sina Zarrieß. 2023. Beyond the bias: Unveiling the quality of implicit causality prompt continuations in language models. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 206–220, Prague, Czechia. Association for Computational Linguistics.

Judith Sieker, Clara Lachenmaier, and Sina Zarrieß. 2025. LLMs struggle to reject false presuppositions when misinformation stakes are high. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 47.

Judith Sieker and Sina Zarrieß. 2023. When your language model cannot Even do determiners right: Probing for anti-presuppositions and the maximize presupposition! principle. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 180–198, Singapore. Association for Computational Linguistics.

Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. A pragmatics-centered evaluation framework for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2382–2394, Marseille, France. European Language Resources Association.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Andreas Stephan, Dawei Zhu, Matthias Aßenmacher, Xiaoyu Shen, and Benjamin Roth. 2025. From calculation to adjudication: Examining LLM judges on mathematical reasoning tasks. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 759–773, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Jean-Baptiste Van der Henst, Yingrui Yang, and P.N. Johnson-Laird. 2002. Strategies in sentential reasoning. *Cognitive Science*, 26(4):425–468.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. 2025. The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models. *Preprint*, arXiv:2505.18497.

A Appendix

A.1 Implementation Details

Models and Inference Setup. We evaluate a diverse set of large language models, including both open-weight and proprietary systems, as described in Section 3. Concretely, we used the following models for our experiments:

- Mistral-7B-Instruct-v0.2
- Mixtral-8x7B-Instruct-v0.1
- OLMo-2-1124-7B-Instruct
- OLMo-2-1124-13B-Instruct
- OLMo-2-0325-32B-Instruct
- Llama-3.1-8B-Instruct
- Qwen3-8B
- Qwen3-14B
- Phi-4
- M-Prometheus-14B
- Claude-Sonnet-4.5
- GPT-4o
- GPT-4.1
- GPT-5

All models were queried in a zero-shot setting using fixed prompts; no task-specific fine-tuning or hyperparameter search was performed.

For open-weight models, we disabled sampling and used deterministic decoding with a maximum generation length sufficient to cover the expected output formats for each task. This choice reduces stochastic variability and ensures reproducibility, which is particularly important for stable item-level comparisons, and it follows prior work using similar evaluation setups (e.g., Bavaresco et al., 2025). No additional decoding constraints (e.g., nucleus sampling or repetition penalties) were applied. Proprietary models were accessed via their respective APIs.

Implementation and Compute. All experiments with open-weight models were implemented in Python 3.9 using PyTorch and the Hugging Face transformers library. Experiments were run on a single NVIDIA RTX A6000 GPU with CUDA acceleration. Depending on model size and task, generating all responses for a model required between approximately 1 and 48 GPU hours. Proprietary models were accessed via their respective APIs. Total API costs were modest (in the range of a few USD per provider).

Output Parsing and Normalization. Across all tasks, models were instructed to respond in a strictly constrained output format (e.g., a single word or label) (Figure 1). Nevertheless, some models produced outputs that mixed target labels with free-form text, explanations, or formatting deviations. To ensure comparability across models, we applied a rule-based parser that maps model outputs to the expected label set where possible.

Parsing Statistics. Adherence to instructions in the prompts varied substantially across models, tasks, and roles. Table 5 reports parsing rates for all model–task–role combinations.

Unparsable outputs primarily resulted from violations of explicit output constraints (e.g., providing explanations despite instructions to respond with a single word or label, or using formats outside the specified option set). Instruction-following failures are most pronounced in generation-based speaker tasks, but their severity differs across pragmatic settings. For False Presuppositions, where we modeled the listener task, parsing reliability is uniformly high across models. For Antipresuppositions, parsing reliability is likewise generally high, with Olmo-2-13B as the main exception (38.1% parsable speaker outputs). For Deductive Reasoning, parsing failures are substantially more severe. For instance, Qwen3-14B produced no parsable outputs in the speaker condition (0% parsing), and Qwen3-8B failed to produce parsable outputs in both the speaker and listener conditions. Olmo-2-13B achieved near-zero parsing coverage in the speaker condition (5%), despite perfect coverage in the listener condition. In addition, LLaMA-8B, Qwen3-32B, and Claude-Sonnet-4.5, show only moderate parsing reliability in the Deductive Reasoning speaker task (around 50%). Notably, Claude-Sonnet-4.5 also exhibits extremely low parsing reliability in the Deductive Reasoning listener condition (5%), markedly lower than all other evaluated models. We therefore report results for these models and tasks, but interpret their performance with caution.

A.2 Scientific Artifacts

We used publicly available datasets, experimental materials, and model implementations, all in accordance with their intended research use and licensing terms. The datasets and prompts consist of constructed linguistic examples and model-generated outputs and do not contain personal data or information that identifies individual people; no anonymization was therefore required. Artifacts created in this work (including prompts and evaluation scripts) are intended exclusively for research and reproducibility purposes and are compatible with the access conditions of the original data sources.

A.3 Use of AI Assistants

AI assistants were used during manuscript preparation only for limited linguistic editing to improve clarity and style, and for writing auxiliary code (e.g., for visualizations). They were not used for scientific reasoning, evaluation decisions, or interpretation of results; all analyses and conclusions were drawn by the authors.

B Additional Results

Contrast-wise analysis for Antipresuppositions. Figure 3 breaks down pragmatic speaker and listener performance by presuppositional contrast, revealing that the overall speaker–listener asymmetry is not uniform across conditions but depends on the type of presupposition trigger involved.

In the DEF condition (i.e., where the MP! principle requires the definite determiner "the"), most models perform substantially better as pragmatic listeners than as speakers. Speaker accuracy is often at or below chance, while listener accuracy approaches ceiling. This indicates that models reliably recognize violations of MP! when evaluating completed sentences, despite failing to consistently select the definite form during generation.

By contrast, in the INDEF condition (i.e., where the MP! principle requires the indefinite determiner "a"), the asymmetry largely disappears. Many models achieve near-ceiling speaker accuracy when generating indefinite forms, reflecting a strong default generation preference for weaker presupposition triggers. In this condition, listener accuracy is sometimes lower than speaker accuracy, suggesting that recognizing the pragmatic infelicity of an unnecessary definite is harder than producing the correct indefinite determiner. This pattern aligns with earlier findings by Sieker and Zarrieß (2023), who show that masked language models (BERT and variants) strongly favor indefinite determiners across conditions. They argue that this bias may partly reflect models' tendency to reproduce surface patterns present in the prompt, where an indefinite determiner is used in the context sentence.

The BOTH condition (i.e., where the MP! principle requires the quantifier "both") exhibits a more heterogeneous pattern. Some models show higher listener than speaker accuracy, while others perform better as speakers, indicating that this contrast interacts more variably with model-specific generation and evaluation strategies.

Taken together, these results show that the

speaker–listener asymmetry in Antipresuppositions arises most clearly when pragmatic reasoning conflicts with models' default generation preferences. When the pragmatically appropriate form is also the model's preferred continuation (as in the INDEF condition), speaker performance is high and the asymmetry is reduced. When pragmatic constraints require overriding this bias (as in the DEF condition), models struggle in generation while remaining highly sensitive to infelicity in evaluation.

False Presuppositions — Scenarios			
Model	**Listener Acc.**	**Speaker Acc.**	**Δ**
Mistral-7B	0.33	0.02	**+0.31**
LLaMA-8B	0.44	0.16	+0.28
GPT-4o	0.78	0.84	**–0.06**

False Presuppositions — Claims			
Model	**Listener Acc.**	**Speaker Acc.**	**Δ**
Mistral-7B	0.38	0.10	+0.28
LLaMA-8B	0.43	0.21	+0.22
GPT-4o	0.63	0.38	+0.25

Table 2: **False Presuppositions:** Listener (judge) vs. speaker (generation) accuracy. Δ = Listener – Speaker. Purple cells indicate higher listener than speaker accuracy; orange cells indicate the reverse. The largest positive and negative Δ values are highlighted in bold.

Model	**Listener Acc.**	**Speaker Acc.**	**Δ**
Mistral-7B	0.96	0.60	+0.36
Mixtral-8x7B	1.00	0.78	+0.22
Olmo-2-7B	0.20	0.49	**–0.29**
Olmo-2-13B	0.99	0.28	**+0.71**
Olmo-2-32B	0.97	0.78	+0.19
LLaMA-8B	0.32	0.45	–0.13
Qwen-3-8B	1.00	0.62	+0.39
Qwen-3-14B	0.98	0.42	+0.56
Phi-4-14B	0.92	0.75	+0.18
Prometheus-14B	0.80	0.64	+0.16
Claude-Sonnet-4.5	0.82	1.00	–0.18
GPT-4o	0.92	0.67	+0.25
GPT-4.1	0.99	0.97	+0.02
GPT-5	0.86	1.00	–0.14

Table 3: **Antipresuppositions:** Listener (judge) vs. speaker (generation) accuracy. Δ = Listener – Speaker. Purple cells indicate higher listener than speaker accuracy; orange cells indicate the reverse. The largest positive and negative Δ values are highlighted in bold.

Model	**Listener Acc.**	**Speaker Acc.**	**Δ**
Mistral-7B	0.56	0.48	+0.08
Mixtral-8x7B	0.54	0.57	–0.03
Olmo-2-7B	0.53	0.28	+0.25
Olmo-2-13B	0.36	0.02	+0.34
Olmo-2-32B	0.49	0.23	+0.26
LLaMA-8B	0.51	0.03	+0.48
Qwen-3-8B	0.00	0.00	+0.00
Qwen-3-14B	0.54	0.00	**+0.54**
Phi-4-14B	0.45	0.48	–0.03
Prometheus-14B	0.43	0.44	–0.01
Claude-Sonnet-4.5	0.05	0.32	–0.27
GPT-4o	0.47	0.50	–0.03
GPT-4.1	0.41	0.69	**–0.28**
GPT-5	1.00	1.00	+0.00

Table 4: **Deductive Reasoning:** Listener (judge) vs. speaker (generation) accuracy. Δ = Listener – Speaker. Purple cells indicate higher listener than speaker accuracy; orange cells indicate the reverse.

Model	FP Listener (%)	AntiPSP Listener (%)	AntiPSP Speaker (%)	Deduct. Reason Listener (%)	Deduct. Reason Speaker (%)
Mistral-7B	100.0	100.0	93.3	89.4	100.0
Mixtral-8x7B	100.0	100.0	100.0	100.0	98.9
Olmo-2-7B	100.0	98.8	65.5	99.4	92.2
Olmo-2-13B	100.0	87.1	38.1	5.6	100.0
Olmo-2-32B	99.8	100.0	100.0	56.1	100.0
LLaMA-8B	91.1	73.6	98.4	49.4	92.8
Qwen-3-8B	98.8	88.9	99.6	0.0	0.0
Qwen-3-14B	100.0	81.8	65.9	0.0	85.0
Phi-4-14B	100.0	100.0	99.2	100.0	100.0
Prometheus-14B	100.0	100.0	97.6	100.0	100.0
Claude-Sonnet-4.5	100.0	100.0	100.0	55.0	5.0
GPT-4o	100.0	100.0	98.8	100.0	100.0
GPT-4.1	100.0	100.0	100.0	100.0	100.0
GPT-5	100.0	100.0	100.0	100.0	100.0

Table 5: **Instruction-following reliability across tasks.** Percentages indicate the proportion of outputs that could be parsed into the target evaluation format for each task (FP = False Presupposition listener task; AntiPSP = Antipresuppositions listener and speaker tasks; Deductive Reasoning = listener and speaker tasks).

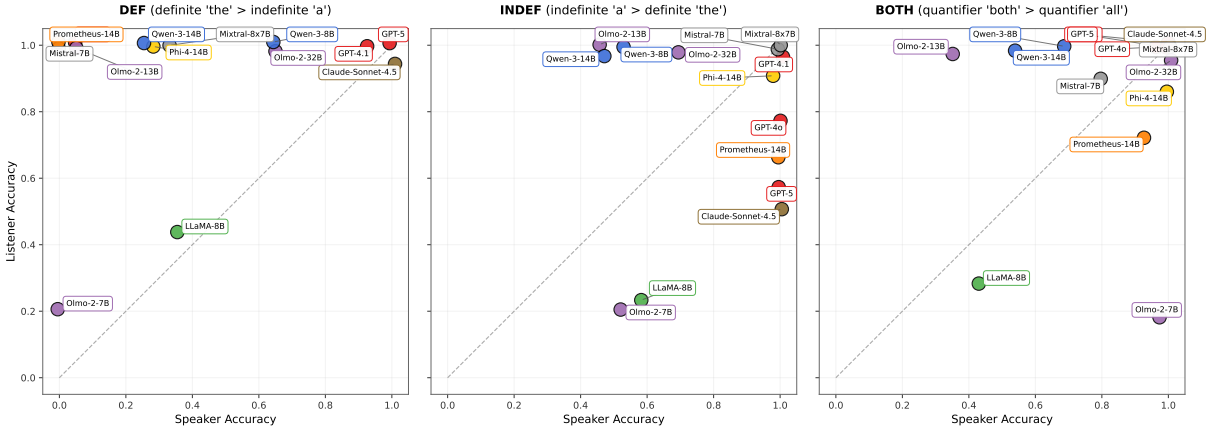


Figure 3: **Speaker vs. Listener accuracy split by condition for Antipresuppositions task**. *DEF* = MP! demands definite determiner, *INDEF* = MP! demands indefinite determiner, *BOTH* = MP! demands the quantifier 'both'. Each point is one model; colors indicate model families. The diagonal marks equal speaker and listener accuracy, so points above the line correspond to models that are better in the listener task than in the speaker task.