

Does Sensitivity to Pragmatic Norms Emerge with More Parameters, More Data, or More Conversations?

Raha Askari^{1,2}, Judith Sieker¹, Valerio Basile², Sina Zarrieß¹

¹ Bielefeld University

² University of Turin

Correspondence: raha.askari@uni-bielefeld.de

Abstract

While previous work has examined the ability of language models (LMs) to capture pragmatic patterns, it remains understudied to what extent these abilities are shaped by training conditions. To address this gap, we test several models using an existing evaluation paradigm targeting LMs' sensitivity to conversational norms (Askari et al., 2025). Across three experimental settings, we investigate the role of model size, training dynamics, and data composition in shaping pragmatic competence. Our results show that different pragmatic abilities exhibit distinct learning dynamics across models with varying sizes. While size or data scaling lead to inconsistent gains, exposure to conversational data causes moderate improvements in a modest training setting. Together, our findings suggest that pragmatic competence in LMs is heterogeneous and highlight the importance of targeted training conditions over scale alone for developing pragmatically sensitive models.

1 Introduction

Human conversation is filled with implicit meanings that are conveyed and understood from context and shared conversational expectations rather than relying on literal words. These shared expectations that shape interlocutors' conversational behaviors have been formalized by Grice (1975) as Maxims that speakers are generally expected to follow unless they wish to give rise to pragmatic inferences. For instance, consider the question *Are you seeing someone currently?* While utterances are normally expected to be relevant to the ongoing topic and truthful with respect to the state of the world, an irrelevant answer such as *It is so warm today.* implicitly signals the speaker's unwillingness to answer, while an obviously false expression such as *Sure, I'm seeing a THOUSAND people.* can convey that the question itself is perceived as absurd.

The idea that interlocutors expect adherence to Gricean Maxims is well established in empir-

ical psycholinguistic research, including studies investigating children's and adults' sensitivity to violations of these conversational norms (Ackerman, 1981; Surian et al., 1996; Eskritt et al., 2008; Okanda et al., 2015; Panzeri and Foppolo, 2021; Diez-Itza et al., 2022). This so called sensitivity to the violations of conversational Maxims, which we will refer to as *pragmatic sensitivity* hereafter, signals listeners to seek pragmatic meanings and is therefore necessary for a seamless communication.

The increasing involvement of language models (LM) in everyday interactions necessitates the analysis of their conversational abilities that align with human expectations. While in human-human interaction such context-based adjustments happen intuitively, LMs often fail to seek clarification or establish common ground, rely on literal interpretations, and tend to generate redundantly detailed responses insensitive to the contextual needs (Lysova et al., 2025). This can make human-AI communication frustrating, as the cognitive load of adjusting and repairing intent is leaned towards users.

Pragmatic abilities in language models have been examined from diverse angles (for example, Lipkin et al. (2023); Jian and Siddharth (2024); Mahowald et al. (2024); Atwell et al. (2025); Qiu et al. (2025); Sieker et al. (2025); Lachenmaier et al. (2025); Ali et al. (2026); Sieker and Zarrieß (2026)), yet the contribution of training conditions to pragmatic competence remains comparatively understudied.

Whereas a significant part of current efforts in the overall LM improvement is accumulated in model and data scaling, the exhaustive computational resources for training and maintaining them have destructive impacts on the environment (Bender et al., 2021), and the financial requirements attributed to them monopolize technological advances in this domain (Warstadt et al., 2023). Furthermore, scaling has been linked to performance degradation in certain tasks (Michaelov and Bergen,

Spontaneous Speech
S1: I talked to Doug. S2: oh, how's he doing? S1: Well, I guess, Do-, the thing is, Doug's kind of annoying me a little bit. S2: oh why? S1: Well, first of all, I think he sort of knew I didn't want him to be here. S2: yeah. S1: But still he – I want some place to bring people.
Staged Dialogues
S1: I like to play board games and make my own clothes. S2: That sounds creative. I like to try new things and travel. S1: I've been to a few countries, but I'd love to travel more. S2: I've been to 7 countries in Europe and 3 in Asia. S1: Wow, that's amazing! I'd love to visit Asia one day. S2: It's beautiful, but it's a long flight.

Figure 1: Examples of Spontaneous Speech (Canavan et al., 1997) and Staged Dialogues (Zhang et al., 2018) from the training data of **Experiment III**. S1/S2 denote Speaker 1/Speaker 2.

2023), suggesting that scaling alone cannot guarantee desired behaviors. When it comes to pragmatics, a separate but equally important question is whether pragmatic competence is only a robust outcome of a model’s structural capacity, or whether it can be traced to properties of the data it is exposed to. While conversational data has been used to improve general model performance, little attention has been paid to *what type* of conversations best support pragmatic representations in a text-only setting, for instance, between spontaneous speech, which mirrors the naturalistic dynamics of real interaction but is less structured, or staged written dialogue, which is more structured but less authentic. Together, these considerations call for a comprehensive analysis of whether, and to what extent, increases in model size, greater data exposure, and different data types are better followed by gains in the pragmatics domain.

With a specific focus on human-like pragmatic competence, we measure the shortcomings of LMs through their ability to discriminate between utterances that either adhere to or violate Grice (1975)’s conversational Maxims. Extending previous work, we conduct three controlled analyses that contribute novel findings to human-centric model evaluation. In Experiment I, we evaluate six models of different sizes trained on identical data to assess the effect of parameter scaling on pragmatic performance. In Experiment II, we look closely at the learning curves of these six models by tracing their pragmatic understanding across 18 training steps. In Experiment III, we distinguish between two types of conversational data (Figure 1) and iso-

late the role of training data using four matched datasets varying in type and amount of conversational content. To the best of our knowledge, we are the first (1) to analyze pragmatic understanding through a developmental lens, i.e., by tracking its emergence during training of several sizes, and (2) to investigate what type of conversation constitutes a more effective choice for modeling pragmatic sensitivity¹.

2 Background

Despite the inherently situated, multimodal, and dynamic nature of human interaction, a growing body of work has investigated pragmatic behavior in Language Models (LMs) through controlled text-based evaluation settings. A recent systematic review by Ma et al. (2025), for instance, highlights the broad range of pragmatic phenomena studied in LMs. In the following, we focus specifically on work examining pragmatic behavior under different training conditions and discuss the limitations motivating our study.

Sravanthi et al. (2024) evaluated Llama-2 (Touvron et al., 2023) across three sizes (7B, 13B, and 70B parameters) and their chat fine-tuned counterparts on phenomena such as implicature and presupposition. In their study, benefits of scaling only became apparent at 70B. Moreover, fine-tuning on chat data yielded mixed results that were inconsistent across phenomena. Their analysis was restricted to large models, leaving room for a more comprehensive size-based investigation.

Hu et al. (2023) evaluated the performance of several models compared to humans on pragmatic items such as humor, irony and metaphor. Despite their overall accuracy scores falling far behind human performance, the larger architectures Tk-Instruct (Wang et al., 2022) (3B, 11B parameters) and Flan-T5 (Chung et al., 2022) (3B) scored higher than GPT-2 (Radford et al., 2019) (117M), showing that large scaling has significant effects on pragmatic behavior. However, the differences of the tested models in their experiments were in domains other than their size alone, including modeling architecture, training data and data quantity, leaving the question unanswered whether improvement was solely due to scaling.

Chen et al. (2025) evaluated interactional phenom-

¹Training, evaluation, and training-data extraction scripts, together with evaluation results, are publicly available on GitHub: <https://github.com/rahasgit/pragmatic-sensitivity>.

ena such as *naturalness* and *interestingness* across several Pythia model sizes (Biderman et al., 2023). Inversely, they found that while scaling yielded only mild improvements, fine-tuning on conversational data led to substantially better performance compared to base models. Their analysis, however, did not trace performance across training steps, leaving pragmatic learning trajectories understudied. Furthermore, many of the evaluated properties they evaluated are subjective in nature, making their interpretability fragile.

Building on psycholinguistic paradigms originally developed for children, Askari et al. (2025) proposed an evaluation framework probing model behavior across five Gricean Maxims, applied to both autoregressive and masked BabyLMs (Warstadt et al., 2023), i.e., small language models trained on substantially limited data. Their setup enabled a controlled architectural comparison, revealing a performance advantage for autoregressive models, namely GPT-2 (Radford et al., 2019) and Baby Llama (Timiryasov and Tastet, 2023). However, other factors potentially shaping pragmatic sensitivity, including model scale, training dynamics, and data composition, remained unexplored.

Most recently, Yu et al. (2026) explored the effect of scaling across three model families, namely OLMo-2 (Walsh et al., 2025) (7B, 13B, and 32B), LLaMA-3.1 (Grattafiori et al., 2024) (8B and 70B) and Qwen-3 (Yang et al., 2025) (0.6B, 1.7B, 4B, and 8B). By probing the models’ ability to pragmatically infer indirect answers, they found that increases in size and/or data scale contributed to enhancements in pragmatic understanding. However, their study did not offer a controlled analysis of data quality, as the evaluated models differed in both architecture and training data simultaneously.

Taken together, previous work suggests that pragmatic competence may depend on multiple training factors, including scale, architecture, and conversational exposure. However, existing studies often confound these variables or investigate them in isolation. In contrast, our work provides a controlled analysis of pragmatic sensitivity across three complementary dimensions: parameter scaling, training dynamics, and training data composition. To this end, we build on the controlled Gricean evaluation framework of Askari et al. (2025), extending it beyond architectural comparisons.

Quantity I
Q: What did you receive for Christmas? F: A toy train. V: A gift.
Quantity II
Q: Who is your teacher? F: My teacher is Miss Green. She reads us stories. V: My teacher is Miss Green. She wears shoes.
Quality
Q: Why don’t you come outside? F: Because I’m helping mom bake. V: Because I’m on a spaceship.
Relation
Q: What did you do at the park? F: I went on the swings. V: My hair is short.
Politeness
Q: What do you think of my new haircut? F: It looks nice. V: It looks awful.

Figure 2: Examples of evaluation items from Askari et al. (2025) based on five conversational Maxims (Grice, 1975); each item consists of a question (Q), a Maxim-following answer (F), and a Maxim-violating alternative (V).

3 Experiments

To investigate how pragmatic sensitivity is shaped by training conditions, we conduct three experiments targeting parameter scaling (Experiment I), training dynamics (Experiment II), and training data composition (Experiment III). Across all experiments, we build on the evaluation paradigm of Askari et al. (2025). Their dataset comprises 2,250 items, corresponding to two-turn conversations each containing a question and two candidate answers: one conforming to a conversational Maxim and one violating it. The Maxims represented in the dataset are Quantity I (*be informative*), Quantity II (*avoid redundant information*), Quality (*be truthful*), Relation (*be relevant*), and Politeness; see examples in Figure 2. The underlying hypothesis is that a model sensitive to a conversational norm should find norm-conforming answers more predictable. Similar to their paradigm, we compute sentence probabilities of both answers conditioned on the question, and measure accuracy per Maxim as the proportion of items for which the model’s probability favors the norm-conforming response.

Across all experimental setups, reported values are accuracy scores from a single evaluation pass over the full dataset (more details in Appendix A.1).

We adopt this evaluation framework as it provides a controlled setting for probing pragmatic sensitivity across multiple conversational norms. In particular, its binary setting, mirroring human experimental paradigms in which Maxim-conforming and Maxim-violating responses are contrasted within the same conversational context (Surian et al., 1996), enables direct comparisons between pragmatically appropriate and inappropriate continuations while minimizing confounds associated with open-ended generation or subjective dialogue evaluation.

Across all experiments, we rely on the autoregressive Pythia suite (Biderman et al., 2023) as our model family of choice. We select the open-source Pythia models as they are released across a wide range of sizes and training checkpoints. All models are trained on the same 300B tokens from the curated English portion of the Pile dataset (Gao et al., 2020), in identical training order, over 143,000 steps with a batch size of approximately 2M tokens. These properties enable comparisons across scale and training stages, and allow us to pre-train new models on custom data while maintaining architectural consistency across experiments.

3.1 Experiment I: Model Scale

In this experiment, we aim to establish whether increasing the number of parameters results in better pragmatic performance, by keeping model architecture and training data constant.

Models. We evaluate six fully trained Pythia models with 70M, 160M, 410M, 1B, 1.4B, and 2.8B parameters.

Results As shown in Figure 3, pragmatic sensitivity varies across Maxims and model sizes, with scaling yielding inconsistent gains.

Quantity I, measuring preference for informative over under-informative responses, appears to be the most challenging norm overall. Models struggle to penalize under-informativeness that is signaled by vague or generic answers, likely because such responses contain highly frequent vocabulary compared to more information-dense alternatives. Although performance generally improves with scale, the drop observed for the largest 2.8B model suggests that larger capacity does not uniformly benefit

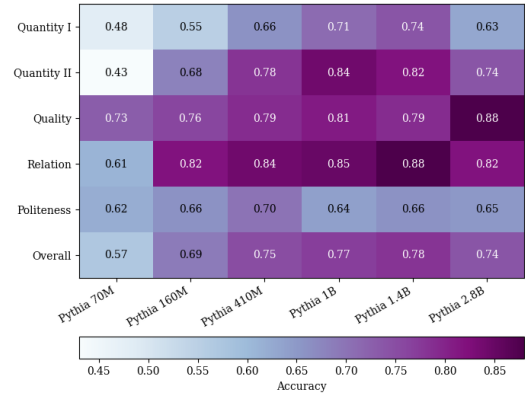


Figure 3: **Experiment I:** Accuracy scores of the final checkpoints of Pythia models across five Maxims; all models were pre-trained on identical data.

Maxim	r_s	p value
Quality	0.90	0.01
Quantity II	0.71	0.11
Quantity I	0.66	0.16
Relation	0.55	0.26
Politeness	0.17	0.74

Table 1: **Experiment I:** Spearman’s correlation (r_s) between model size (rank) and accuracy (rank).

sensitivity to informativeness.

For Quantity II, which evaluates the models’ sensitivity to disfavor redundant responses, scale-based gains are less consistent, with $\geq 1B$ models under-performing their smaller counterparts. When balancing conciseness and brevity, the increased capacity of larger models may bias lean toward favoring more elaborate responses, making them more tolerant of conversational redundancy.

When testing models on their ability to discriminate between truthful and untruthful responses with the Maxim of Quality, model size is almost well aligned with performance gains, although even the smallest models perform considerably above chance (0.50). Overall, all sizes, especially larger ones, seem decently equipped to capture cues associated with factuality and truthfulness beyond surface-level lexical distributions. This may reflect the relative rarity of implausible statements like those in the evaluation dataset outside fictional training contexts, making them easier to distinguish through distributional learning.

Evaluating judgments of relevance with the Maxim of Relation yields the highest overall score across models. Accuracy increases gradually with scale, followed by a slight dip again for the 2.8B model. This pattern suggests that, up to a certain

point, relevance benefits greatly from increases in capacity, likely because modeling topical coherence requires richer contextual representations, but that very large models may diffuse relevance by integrating too many loosely related associations, especially when trained on highly diverse data.

Finally, the models’ sensitivity to politeness cues is assessed with the Politeness Maxim. Although gains build up mildly up to 140M-parameter model, scores do not vary substantially as the number of parameters increases exponentially. This stability may reflect the relatively explicit nature of politeness markers, which can be learned to some degree but do not require large representational capacity.

To quantify the relationship between model scaling and pragmatic sensitivity for each Maxim, we additionally compute Spearman’s rank correlations between model size and accuracy (Table 1). Quality shows the strongest correlation with scaling ($r_s = 0.90$, $p = 0.01$), indicating consistent performance gains with model size. The two Quantity Maxims and Relation show moderate correlations ($r_s = 0.71, 0.66$, and 0.55 respectively), though none reach statistical significance ($p > 0.05$), likely due to the small number of models evaluated. Politeness exhibits the weakest relationship ($r_s = 0.17$, $p = 0.74$), indicating that performance on this Maxim is largely robust to model scaling.

3.2 Experiment II: Training Dynamics

In this experiment, we trace the development of pragmatic sensitivity across 18 training checkpoints for the six models from Experiment I to examine how different conversational norms emerge over the course of learning.

Models. We evaluate 18 checkpoints of the same six Pythia models used in Experiment I, sampled at approximately logarithmic intervals, starting with ten initial steps, i.e. 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and eight steps belonging to later stages of training, i.e. 1k, 2k, 4k, 8k, 16k, 32k, 64k, 128k. The first checkpoint corresponds to exposure to approximately 2M tokens, while the final checkpoint reflects roughly 256B seen tokens.

Results. Figure 4 shows the learning dynamics of pragmatic sensitivity across training checkpoints for all model sizes and Maxims. A closer analysis of these trajectories reveals that increased training exposure does not uniformly benefit the acquisition of all pragmatic norms.

For Quantity I, all models undergo a sharp performance decline around 256M to 512M seen tokens, with recovery beginning only after approximately 4B tokens. Beyond this point, smaller models fluctuate while larger models show modest gains. This suggests that the preference towards sufficiently informative statements remains challenging during training, with larger models benefiting only modestly from increased data exposure.

In Quantity II, larger models exhibit a notably steep learning curve during later training stages, particularly after 2B tokens, with little differentiation across mid-sized and large models. This suggests that, beyond a certain threshold, sensitivity to conversational redundancy is driven more by data exposure than by model capacity. In contrast, the smallest models remain near chance throughout training, indicating that learning to penalize redundancy is highly fragile at small scales and cannot be achieved through data exposure alone.

For the Maxim of Quality, performance differences across model sizes are modest even in early training, and while an upward trend is visible across all models, improvement is slow relative to token exposure. After 256M tokens, gaps between even the smallest and largest models narrow. Overall, representations supporting truthfulness appear relatively robust across training, with all models responding faintly positive to data scaling.

For the Maxim of Relation, larger models acquire relevance cues faster, though their gains saturate after an 8B token budget. While the 160M model remains a steady upward spike across later training stages, the smallest model gives up after 1B tokens. This suggests that larger models stabilize relevance judgments earlier, while smaller models can approach comparable performance levels given sufficient data exposure.

In the Politeness Maxim, despite notable initial differences tied to model size, performance across models intertwines relatively fast, with accuracy oscillating mildly after around 32M tokens and showing little improvement thereafter. This pattern suggests that politeness cues may be more consistently distributed across training data, allowing them to be acquired early, around 32M to 64M tokens, and remain largely robust to further training.

Overall, these trajectories confirm that pragmatic learning is neither uniform across Maxims nor strictly proportional to data exposure, with each norm exhibiting a distinct developmental pattern that interacts differently with model size and train-

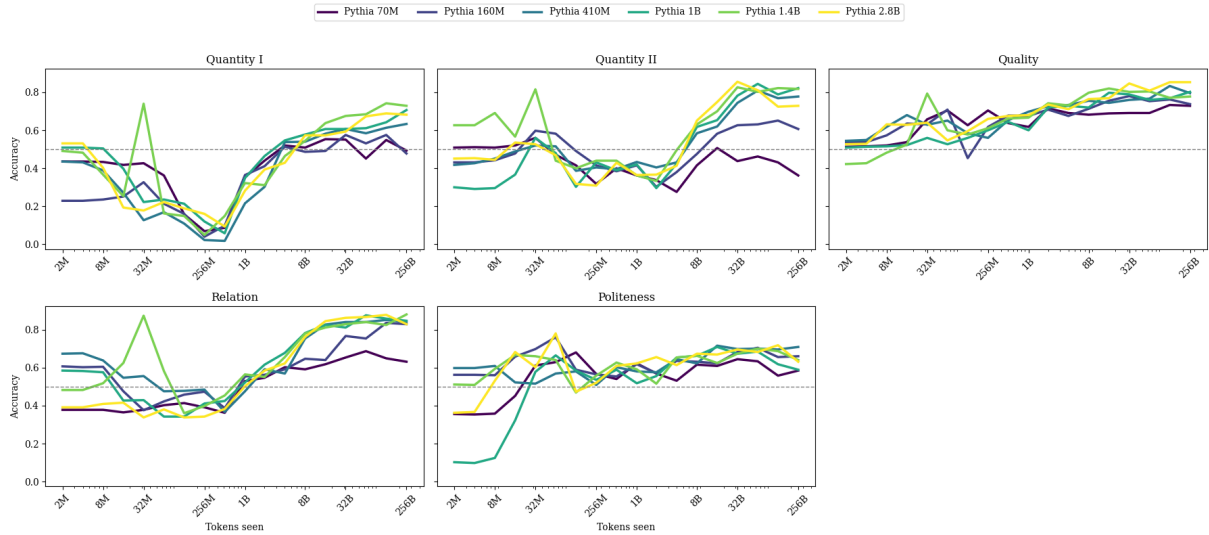


Figure 4: **Experiment II:** Accuracy scores of Pythia models across several checkpoints (shown with the number of tokens at each step). Lines are drawn to guide the eye, and scores are only defined at discrete training steps; dashed line shows chance level.

ing progress.

3.3 Experiment III: The Role of Training Data

In this experiment, we examine the role of data composition by training Pythia BabyLMs with the same model architecture on four size-matched datasets varying in the type and amount of conversational content ².

Training Data. We compose four distinct datasets varying in the amount and type of conversational content: two conversational datasets, *Spontaneous Speech* and *Staged Dialogues*, one non-conversational dataset, *No Dialogue*, and a combined condition, *Mixed*. All datasets were aggregated from existing corpora spanning diverse topics and registers and matched in size at approximately 10M words. This comparatively small scale was motivated by the limited availability of certain data types, particularly *Spontaneous Speech*, as well as our interest in computationally efficient training settings.

Spontaneous Speech consists of transcribed spoken interactions collected under minimally controlled conditions, reflecting natural conversational dynamics and including phenomena such as repetitions and self-repairs. It draws from CHILDES (MacWhinney, 2000), BNC Spoken³ (BNC Con-

sortium, 2007), Switchboard (Stolcke et al., 2000), CallHome (Canavan et al., 1997), and CallFriend (Canavan and Zipperlen, 1996a,b). In contrast, *Staged Dialogues* contains written or scripted interactions collected under more controlled conditions, including persona-based chats and topic-guided conversations, resulting in more structured conversational data. It is composed of Topical Chat (Gopalakrishnan et al., 2023), Persona Chat (Zhang et al., 2018), and Daily Dialogue (Li et al., 2017). *No-dialogue* serves as a non-conversational control condition, allowing us to assess whether pragmatic sensitivity can emerge independently of conversational training signals. It comprises FineWeb-Edu (Penedo et al., 2024), Simple Wikipedia⁴, and KidLM (Nayeem and Rafiei, 2024). Finally, *Mixed* combines all three data types to approximate contemporary language model training mixtures. Detailed corpus composition is provided in Appendix A.2 and Table 5.

Models. All models were based on the Pythia 160M architecture, which comprises 12 transformer layers and 12 attention heads. We report results from the final training checkpoint, while information about training procedure, and the models’ performances on common evaluation benchmarks are provided in Appendix A.3.

Results. Table 2 presents the accuracies of the models trained on the four data conditions for each

²Our trained models are available on Hugging Face: <https://huggingface.co/collections/rahaaskari/conversational-pythia-models>.

³<https://www.natcorp.ox.ac.uk/>

⁴<https://dumps.wikimedia.org/simplewiki/>

Data	Quan. I	Quan. II	Qual.	Rel.	Pol.	Overall	Comparison	Acc. Diff.	<i>p</i>
Staged	0.48	0.64	0.78	0.59	0.74	0.65	Staged vs No-dialogue	+7.8%	< .001
Spontaneous	0.40	0.66	0.84	0.57	0.65	0.62	Spontaneous vs No-dialogue	+5.7%	< .001
Mixed	0.48	0.61	0.76	0.56	0.65	0.61	Spontaneous vs Staged	-2.1%	= .019
Non-dialogue	0.56	0.46	0.65	0.52	0.66	0.57	Mixed vs No-dialogue	+4.2%	< .001

Table 2: **Experiment III:** (Left) Accuracy of Pythia 160M models trained on different datasets; highest scores per Maxim and overall are bolded. (Right) Pairwise McNemar comparisons of Accuracy Differences across all Maxims combined; positive differences indicate higher accuracy for the first model. *p*-values are Holm-corrected.

individual Maxim, along with pairwise McNemar comparisons across all Maxims combined. Since all models are evaluated on the same items, McNemar tests allow us to determine whether the observed accuracy differences between training conditions reflect systematic and statistically reliable item-level behavioral difference. We apply Holm-Bonferroni correction to account for multiple pairwise comparisons and report Maxim-specific McNemar results in Table 8 in the Appendix A.4.

The overall McNemar comparisons show that all dialogue-trained models significantly outperform the *No-dialogue* condition. The strongest effect is observed for *Staged Dialogues* compared to *No-dialogue* ($p < .001$), followed by *Spontaneous Speech* ($p < .001$) and *Mixed* training ($p < .001$). In addition, *Staged Dialogues* significantly outperform *Spontaneous Speech* ($p = .019$). These results suggest that conversational training data improves pragmatic sensitivity overall. However, as shown by the accuracy patterns (Table 2) and the Maxim-specific McNemar results (Table 8), this effect is strongly Maxim-dependent: Conversational training data does benefit some pragmatic abilities, but not uniformly across all pragmatic dimensions.

Notably, the two Quantity dimensions exhibit opposing patterns. While dialogue-trained models show large improvements for detecting redundant information (Quantity II), the *No-dialogue* model consistently performs best on detecting underinformative response (Quantity I). This inferiority of the dialogue-trained models in Quantity I is likely because the *No-dialogue* training data predominantly consists of explanatory and highly specific text, making distinctions between informative and under-informative responses more salient; especially compared to *Spontaneous Speech*, where information density is highly distorted by filler expressions. The slightly stronger performance of *Staged Dialogues* may result from its more structured written form. These results indicate that mod-

eling sufficient informativeness in interaction remains highly challenging and is not aided by conversational examples. In contrast, however, conversational training data does appear beneficial for Quantity II, where the *Spontaneous Speech* model performs best, followed by the *Staged Dialogue* model, where over-informativeness may be artificially inflated at times. The decreased performance of *Mixed* and the below-chance score of the *No-dialogue* model suggest that non-conversational text does not reflect the amount of information typically exchanged in dialogue and biases models toward overly elaborate responses. This supports the idea that sensitivity to conversational redundancy is best acquired from dialogue examples.

For the Maxim of Quality, the conversational models clearly outperform the *No-dialogue* condition despite the latter being trained on factual text. This may reflect that the evaluation items rely on simple everyday world knowledge, such as the implausibility of seeing a thousand people, that occurs more naturally in conversational interaction than in encyclopedic sources. The strong performance of *Spontaneous Speech* may further stem from its broad coverage of interactions across diverse age groups and communicative settings, exposing models to world knowledge at varying levels of complexity.

For the Maxim of Relevance, all models perform near chance, though *Staged Dialogues* achieves the highest accuracy, plausibly due to the explicit topic constraints in its source data. This is in contrast with *Spontaneous Speech* interactions, where topic delimitations are looser and transitions are frequent. The lowest performance by the *No-dialogue* model is likely because the text frequently spans multiple subtopics within a single passage. Overall, this suggests that conversational relevance goes beyond simple semantic similarity, making it harder to capture through language modeling alone.

On the Maxim of Politeness, the *Staged Dia-*

logues model is the only one to show a clear benefit from its training data, consistent with the explicit politeness conventions present in staged interactions. This is unlike the spontaneous examples among familiar interlocutors where politeness is expressed more implicitly. The *No-dialogue* model, while unlikely to have encountered impolite instances in its training data, also lacks reliable politeness markers that the model can consistently detect and learn from, resulting in performance comparable to the other non-staged models.

Overall, structured interactions appear particularly beneficial for pragmatic learning, likely because they provide clearer pragmatic signals than either spontaneous dialogue or non-conversational text. The reduced performance of the *Mixed* model further suggests that incorporating non-dialogue text may dilute conversational learning signals.

4 Error Analysis

To better understand the underlying what drives model failures in pragmatic awareness, we examine how often the models from Experiment I fail on each Maxim (i.e., Maxim difficulty), and how it correlates with semantic similarity and the token length difference between the two candidate answers of the corresponding evaluation items (more details regarding these metrics are provided in Appendix A.5).

Looking at Table 3, Quantity I and Politeness yield the most difficult items on average. Quantity II items show the highest semantic similarity between answer pairs; for instance, *I ate yogurt.* vs. *I ate yogurt that comes in a small plastic pot.* differ in redundancy while having slightly similar content. By contrast, Relation items show the lowest similarity, which is in line with the nature of irrelevant answers that are semantically far from the topic. Looking at token length differences, the most defined difference belongs to Politeness, with the Maxim-conforming answers on average 2.04 tokens more than the Maxim-violating alternatives, consistent with the fact that politeness is often hinted through additional words such as “*Sorry*” and “*Thanks*” compared to less polite alternatives.

When looking at correlations with difficulty, semantic similarity reaches significance only for Quantity II, meaning that only for this Maxim, items where the two answers are more semantically similar tend to be slightly harder, while for all other Maxims, semantic similarity does not cor-

relate with difficulty. Token length difference, on the other hand, is significantly and negatively correlated with difficulty across all Maxims; i.e., when the correct answer is longer, pragmatic judgements tend to be easier than when the incorrect answer contains more tokens. The effect is strongest for Quantity I, where models tend to only succeed on items where the uninformative answer is shorter, e.g. preferring “*Tomato soup*” over “*A dish*” to “*What did you have for dinner?*” Overall, we find that semantic closeness of the answer pairs does not seem to have a large effect on models’ pragmatic judgments. Instead, token length asymmetry is a consistent predictor of difficulty, indicating a length bias in models towards longer answers regardless of conversational needs. Figure 5 in the Appendix shows examples with varying difficulty levels across the pretrained models of Exp. I.

5 Discussion and Conclusion

Interlocutors convey explicit and implicit messages through adherence to or deliberate violation of conversational norms formalized as Maxims (Grice, 1975). For LMs to recover such meanings, they must have the sensitivity to rationalize the speaker’s choice over Maxim violations in order to arrive at a pragmatic inference. The present work examines how this pragmatic sensitivity is shaped by model size, training dynamics, and data composition, using a binary task of assigning higher probabilities to Maxim-conforming answers.

Although the Gricean framework has faced criticism for being vague, incomplete, or culturally specific (Frederking, 1996), it nonetheless provides a principled basis for identifying conversational behaviors that humans acquire from a very young age. As such, it offers a robust, if partial, lens for evaluating whether language models exhibit behavior analogous to human conversational expectations. Our results should be interpreted within this scope: as a targeted probe of specific, well-defined pragmatic competences rather than a comprehensive assessment of conversational ability.

Taking the findings of all three experiments together, our analyses show that models are able to learn certain pragmatic patterns from training data, and that these representations are reinforced to some degree by both parameter scaling and increased data exposure. However, as illustrated in (a) and (b) in Table 4, these gains are modest relative to the resources they require. By contrast, as

Maxim	Diffic.	Sem. Sim.	r (sim)	Len. Diff.	r (len)
Quantity I	0.372	0.711	-0.089	+0.68	-0.695***
Quantity II	0.285	0.838	+0.153**	-0.20	-0.311***
Quality	0.207	0.655	+0.050	-0.52	-0.205***
Relation	0.196	0.572	-0.019	+0.68	-0.394***
Politeness	0.345	0.605	+0.072	+2.04	-0.358***

Table 3: Per-Maxim difficulty (**Diffic.**) for models in Experiment I and its Pearson correlations with average semantic similarity (**Sem. Sim.**) and average length difference (**Len. Diff.**) between candidate answers of the evaluation items. Higher difficulty values indicate greater difficulty. Higher semantic similarity (0–1) means more similar. Positive length differences indicate longer Maxim-conforming answers. ** $p < .01$; *** $p < .001$.

Param. increase	GPU hrs	Δ	Checkpoint	Tokens	Δ	Model	Δ 8M	Δ 16M
70M→160M	+520	+12%	1→8	+14M	+4%	8M-checkpoint	baseline	—
160M→410M	+1,510	+6%	8→512	+998M	+2%	16M-checkpoint	+3%	baseline
410M→1B	+2,290	+2%	512→16k	+31B	+16%	Staged	+17%	+14%
1B→1.4B	+2,290	+1%	16k→128k	+224B	-3%	Spontaneous	+14%	+11%
1.4B→2.8B	+7,120	-4%	(b) Data exposure (Exp. II)			(c) Training data type (Exp. III)		
(a) Parameter scaling (Exp. I)								

Table 4: Changes in overall accuracy (average across Maxims) across three dimensions: (a) parameter scaling with corresponding GPU hour increases (as reported by [Biderman et al. \(2023\)](#)); (b) token exposure across Pythia 160M checkpoints; (c) overall accuracy of models trained on ~ 12 M conversational tokens compared to pretrained Pythia 160M checkpoints at 8M and 16M tokens.

shown in (c) of Table 4, training on approximately 12M targeted conversational tokens produces significant gains over a Pythia 160M checkpoint trained on comparable numbers of tokens from a general corpus, suggesting that data composition can yield meaningful improvements in a small training budget. We present these comparisons without presuming to rank the three approaches directly, as their costs and objectives are not strictly comparable, but the contrasts are nonetheless informative.

Although the literature surrounding the development of current large models is transparent about required training hours, it is rarely discussed how these exponential costs pay off with regards to accuracy improvement. Hence, alongside the already encouraged transparency regarding reporting the energy costs of model training ([Henderson et al., 2020](#)), defining what should be considered a significant improvement relative to these costs will ease some of the pressure on parameter and data scaling as the sole methods of model development.

Zooming into the findings of Experiment III, naturally occurring conversations, although valuable for studying authentic interactional behavior, proved less effective for modeling pragmatic competence than staged dialogues. The repairs, repetitions, and syntactic fragmentation inherent to spontaneous speech may lead models to internalize irregular patterns misaligned with pragmatic appropriateness.

This raises an important design question: whether the goal of conversational modeling is to build models that faithfully reproduce human speech, or models that are optimized for communicative clarity and coherence, despite falling short on naturalness at times.

On a general note, existing efforts target dimensions that overlap individually with the conversational principles assessed in this work. For instance, Reinforcement Learning from Human Feedback ([Ouyang et al., 2022](#)) encourages socially appropriate conversational behavior, among which brevity and politeness; similarly, Retrieval-Augmented Generation ([Lewis et al., 2020](#)) can mitigate violations of Quality by allowing models to ground their responses in retrieved evidence instead of hallucinating. Other examples include teaching models to express uncertainty ([Lin et al., 2022](#)) or to ask for clarification ([Andukuri et al., 2024](#)), allowing them to avoid giving irrelevant responses due to misunderstanding. Our results, however, point to a promising direction: shifting the primary developmental focus toward interaction itself. Such an objective entails expanding the availability of conversational datasets enriched with layered contextual information, such as social settings, power dynamics, and demographic variation, to equip models with the kind of context-sensitive flexibility that characterizes human pragmatic behavior.

Limitations

The scale of the experiments was constrained by the available computational resources and conversational data, which in turn limited the generalizability of our findings in the following ways.

First, our work lacks multilingualism. The availability of multiple model sizes and training checkpoints within the Pythia suite enabled controlled analyses of scaling effects and learning dynamics, but the monolingual nature of the models and our training data restricted our experiments to English.

Second, as discussed in Askari et al. (2025), while a probability-based evaluation is a common practice, the scores that models assign to the evaluation items can be biased by factors such as lexical or length variations of utterances rather than reflecting learned pragmatic representations.

Third, our evaluation of pretrained models did not extend to very large architectures. It is plausible that very large models are more successful in extracting pragmatic patterns from the same data. Additionally, despite seeing positive results in small models trained on conversational data, whether gains attributed to conversational examples persist when training larger models remains unanswered in the current study.

Fourth, this work focuses exclusively on the pre-training stage and does not address the effect of post-training alignment or instruction tuning on pragmatic sensitivity, although such procedures, given their explicit optimization for human preferences, may conceivably yield further improvements on this task.

Finally, although conversational data improved the model performance, inter-corpora differences (especially between the two conversational datasets) and their effect on model performance were not explored in the current work. We acknowledge that multiple characteristics of the datasets may contribute to the observed differences, and our results should be interpreted as a comparison between representative categories of conversational data rather than an exhaustive characterization of all possible sources.

Our findings should therefore be considered within these constraints, and broader cross-linguistic, cross-corpora or cross-architecture conclusions require further investigation.

Ethical Considerations

This work uses publicly available datasets, pre-trained models, and training corpora in accordance with their intended purpose. While we do not plan to redistribute any of the datasets used, we share our scripts for corpus extraction and filtering alongside the original sources of each corpus. As no new data was collected and no human participants were involved, no anonymization or consent procedures were required. No sensitive or personal information was gathered beyond what may exist in the public sources used.

The environmental impact is limited, as experiments involve only small models and a constrained training setup. The models trained in our experiments were developed solely to address our research questions; since their generative competence was not the focus of this work, we did not perform data detoxification prior to training and their outputs were not evaluated for undesirable behavior. As such, we warn against any potential bias or harmful language in their generation, which reflects patterns in the training data and not the views of the authors. We release our trained models for research purposes only and do not intend for them to be used commercially.

Finally, the exclusive focus on English was determined by resource availability rather than intent to exclude other languages or communities.

References

- Brian P Ackerman. 1981. When is a question not answered? the understanding of young children of utterances violating or conforming to the rules of conversational sequencing. *Journal of Experimental Child Psychology*, 31(3):487–507.
- Manar Ali, Judith Sieker, Sina Zarrieß, and Hendrik Buschmeier. 2026. [Reference games as a testbed for the alignment of model uncertainty and clarification requests](#). In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 990–998, San Diego, California, USA. Association for Computational Linguistics.
- Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.
- 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.

- Katherine Atwell, Mandy Simons, and Malihe Alikhani. 2025. [Measuring bias and agreement in large language model presupposition judgments](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2096–2107, Vienna, Austria. Association for Computational Linguistics.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). *Preprint*, arXiv:2304.01373.
- BNC Consortium. 2007. [The british national corpus, xml edition](#). Distributed by the University of Oxford on behalf of the BNC Consortium.
- Alexandra Canavan, David Graff, and George Zipperlen. 1997. CALLHOME American English Speech LDC97S42.
- Alexandra Canavan and George Zipperlen. 1996a. CALLFRIEND American English-Non-Southern Dialect LDC96S46.
- Alexandra Canavan and George Zipperlen. 1996b. CALLFRIEND American English-Southern Dialect LDC97S47.
- Zixun Chen, Petr Babkin, Akshat Gupta, Gopala Anumanchipalli, and Xiaomo Liu. 2025. A multi-aspect evaluation of dialogue in pythia. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. BabyLM turns 4 and goes multilingual: Call for papers for the 2026 babyLM workshop. *arXiv preprint arXiv:2602.20092*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- Andrea Gregor De Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56(5):5190–5213.
- Eliseo Diez-Itza, Aitana Viejo, and Maite Fernández-Urquiza. 2022. [Pragmatic profiles of adults with fragile x syndrome and williams syndrome](#). *Brain Sciences*, 12(3).
- Michelle Eskritt, Juanita Whalen, and Kang Lee. 2008. Preschoolers can recognize violations of the gricean maxims. *British Journal of Developmental Psychology*, 26(3):435–443.
- R Frederking. 1996. Grice’s maxims: do the right thing. *Frederking, RE*.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
- Karthik Gopalakrishnan, Behnam Hedayatnia, Qinqiang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.
- 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.
- Herbert P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press.
- Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of machine learning research*, 21(248):1–43.
- 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.
- Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi U Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, and 1 others. 2025. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Transactions of the Association for Computational Linguistics*, 13:1245–1270.
- Mingyue Jian and N. Siddharth. 2024. [Are llms good pragmatic speakers?](#) *Preprint*, arXiv:2411.01562.

- Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855.
- 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.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [DailyDialog: A manually labelled multi-turn dialogue dataset](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
- Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*.
- Iuliia Lysova, Lara Ahmed, Bethan Cunningham, Yonghua Huang, and Martina Wiltschko. 2025. Do chatbots dream of ai sheep? a semantic–pragmatic investigation of "naturalness" in human–ai interaction. *AI & SOCIETY*, pages 1–10.
- 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.
- Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.
- Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.
- James A. Michaelov and Benjamin K. Bergen. 2023. [Emergent inabilities? inverse scaling over the course of pretraining](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14607–14615, Singapore. Association for Computational Linguistics.
- Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.
- Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. [COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.
- Mir Tafseer Nayeem and Davood Rafiei. 2024. [KidLM: Advancing language models for children – early insights and future directions](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4813–4836, Miami, Florida, USA. Association for Computational Linguistics.
- Mako Okanda, Kosuke Asada, Yusuke Moriguchi, and Shoji Itakura. 2015. Understanding violations of gricean maxims in preschoolers and adults. *Frontiers in psychology*, 6:901.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Francesca Panzeri and Francesca Foppolo. 2021. Children’s and adults’ sensitivity to gricean maxims and to the maximize presupposition principle. *Frontiers in Psychology*, 12:624628.
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.
- 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 19913–19935, Suzhou, China. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *OpenAI*. Accessed: 2024-11-15.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical*

- Methods in Natural Language Processing*. Association for Computational Linguistics.
- Judith Sieker, Clara Lachenmaier, and Sina Zarri . 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 Zarri . 2026. [How hypocritical is your LLM judge? listener-speaker asymmetries in the pragmatic competence of large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 7551–7565, San Diego, California, United States. Association for Computational Linguistics.
- 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.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *Computational Linguistics*, 26(3):339–374.
- Luca Surian, Simon Baron-Cohen, and Heather van der Lely. 1996. [Are children with autism deaf to gricean maxims?](#) *Cognitive neuropsychiatry*, 1:55–72.
- Inar Timiryasov and Jean-Loup Tastet. 2023. [Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 279–289, Singapore. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafford, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. [2 OLMo 2 furious \(COLM’s version\)](#). In *Second Conference on Language Modeling*.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Gianis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, and 16 others. 2022. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. [Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pi ric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. 2026. [The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 192–213, Rabat, Morocco. Association for Computational Linguistics.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. [Personalizing dialogue agents: I have a dog, do you have pets too?](#) In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Appendix

A.1 Evaluation Paradigm

To compute conditional probability for each item, the question is prepended to each candidate answer, and token-level log probabilities are extracted starting from the answer tokens, conditioned on all preceding tokens including the question. The sentence probability is the normalized sum of these answer token log probabilities. Scores are obtained through Minicons library (Misra, 2022), and the number of times the model assigns a higher probability to the Maxim-conforming answer are counted as its accuracy. We obtain the evaluation paradigm and dataset from Askari et al. (2025)⁵. Examples from the evaluation items can be seen in Figure 2.

A.2 Training Data Details and Composition

Dataset	Source corpus	Tokens (M)
Spontaneous	CHILDES	8
	BNC Spoken	2.2
	Switchboard	1.9
	CallHome	0.32
	CallFriend	0.07
	Total	12.5
Staged	Topical Chat	6.4
	Persona Chat	4.3
	Daily Dialogue	1.8
	Total	12.5
No-dialogue	FineWeb-Edu	5
	Simple Wiki	5
	KidLM	2.5
	Total	12.5
Mixed	Spontaneous	4.2
	Staged	4.2
	No-dialogue	4.2
	Total	12.6

Table 5: Composition of the four training datasets.

All datasets were constructed from pre-existing corpora selected to ensure diversity in topic, register, and interaction type, and split into train, validation, and test sets following an 80-10-10 distribution. Spontaneous speech examples underwent basic curation to remove transcription markers, and across all dialogue items, missing speaker tags were automatically replaced with randomly selected English names⁶ for consistency across dialogue datasets.

⁵https://huggingface.co/datasets/rahaaskari/gricean_baby

⁶<https://github.com/KarlAmort/firstname-database>

Spontaneous Speech. The *Spontaneous Speech* dataset consists of recorded spoken dialogues that were subsequently transcribed. These interactions span a variety of topics and involve diverse participant configurations, including family members, friends, strangers, and child-directed speech. Speakers were given no or minimal instructions, resulting in highly natural exchanges that reflect authentic conversational dynamics. As a consequence of their spontaneous nature, the dialogues also exhibit interactional phenomena including repetitions, scrambled structures and self-repairs. Although originally collected for psycholinguistic research on human conversational behavior and language acquisition, these corpora have since been incorporated into LM training data for due to their value in documenting authentic language use.

Staged Dialogues. The *Staged Dialogues* dataset, by contrast, comprises conversations whose original modality is written language. These include chat interactions between adult strangers who were assigned specific topics, instructed to convey particular information, or given personas to adopt, as well as scripted dialogues authored by humans to cover a range of topics. Due to their written form, these conversations exhibit fewer of the interactional phenomena of spontaneous speech, but their highly controlled production conditions deviates them considerably from naturalistic interaction. Unlike spontaneous speech corpora, these datasets were collected specifically for the purpose of model training.

No-dialogue. The *No-dialogue* dataset serves as a non-conversational control condition and contains expository and non-interactional text. This condition allows us to assess whether pragmatic sensitivity can emerge independently of conversational training signals.

Mixed. The *Mixed* dataset is intended to reflect current trends in LM training, where conversational and non-conversational data are typically combined. Including it allows us to compare models trained on diverse data mixtures against those exposed exclusively to interactions.

Table 5 provides the detailed composition of the datasets used in Experiment III. The used portions from CHILDES, BNC Spoken and Switchboard were distributed by the BabyLM challenge⁷ as part

⁷<https://huggingface.co/BabyLM-community>

of the training data of the challenge, while CallHome and CallFriend were taken from TalkBank⁸.

A.3 Training Information and Benchmark Evaluation

Model	Training	Validation	Test
Spontaneous	2.69	2.54	2.50
Staged	2.65	2.54	2.54
No-dialogue	3.38	3.51	3.54
Mixed	3.22	3.16	3.16

Table 6: The losses of Pythia models trained on varied training data.

For models of Experiment III, training was performed using HuggingFace’s causal language modeling pipeline (Wolf et al., 2020) for 2 epochs, with a maximum sequence length of 2048 tokens, a learning rate of 1e-4, and an effective batch size of 4. To ensure that performance differences across conditions reflect training data alone rather than tokenization choices, the same pretrained Pythia BPE tokenizer was used for all four models. Training was conducted on a Windows machine with a 24-thread Ryzen 9 7900X CPU and 32 GB RAM, taking approx. 11.5 hours per model. No validation-based model selection was performed.

A summary of training, validation, and test losses for all models is provided in Table 6, while Table 7 shows the models’ scores on the standard benchmarks using the BabyLM zero-shot evaluation Pipeline (Choshen et al., 2026).

Task	Spontan.	Staged	Mixed	No-dial.
BLiMP	71.62	66.82	69.03	68.00
Suppl.	60.26	60.62	55.22	51.75
EWoK	57.27	53.69	54.92	54.67
Entity Track.	29.81	37.75	33.42	32.82
COMPS	55.32	53.52	55.27	55.38
Eye Track.	0.22	0.52	0.38	0.64
Self-Paced.	0.00	0.01	0.11	0.34

Table 7: BabyLM 2026 evaluation results across training conditions. Higher is better for all metrics; highest score in each benchmark is bolded.

BLiMP (Warstadt et al., 2020) and BLiMP supplement (Warstadt et al., 2023) measure syntactic competence; EWoK (Ivanova et al., 2025) tests world knowledge; Entity Tracking (Kim and Schuster, 2023) evaluates whether models can track changes in the states of entities; COMPS measures

⁸<https://talkbank.org/ca/access/CallFriend/> and <https://talkbank.org/ca/access/CallHome/>

Maxim	Comparison	Acc. Diff	p
Quantity I	Spont. vs No-dialogue	-15.3%	< .001
	Staged vs No-dialogue	-8.0%	= .004
	Spont. vs Staged	-7.3%	= .070
	Mixed vs No-dialogue	-7.6%	= .002
Quantity II	Spont. vs No-dialogue	+20.4%	< .001
	Staged vs No-dialogue	+18.7%	< .001
	Spont. vs Staged	+1.8%	= 1.000
	Mixed vs No-dialogue	+14.9%	< .001
Quality	Spont. vs No-dialogue	+18.9%	< .001
	Staged vs No-dialogue	+13.3%	< .001
	Spont. vs Staged	+5.6%	< .001
	Mixed vs No-dialogue	+10.7%	< .001
Relation	Staged vs No-dialogue	+6.7%	= .011
	Spont. vs No-dialogue	+5.1%	= .381
	Spont. vs Staged	-1.6%	= 1.000
	Mixed vs No-dialogue	+4.0%	= .491
Politeness	Spont. vs Staged	-9.1%	< .001
	Staged vs No-dialogue	+8.4%	= .004
	Spont. vs No-dialogue	-0.7%	= 1.000
	Mixed vs No-dialogue	-0.9%	= 1.000

Table 8: Pairwise McNemar comparisons per Maxim. Positive differences indicate higher accuracy for the first model in the comparison. Reported p-values are Holm-corrected. **Spont.** denotes Spontaneous.

the LM’s property knowledge (Misra et al., 2023); Eye Tracking and Self-paced Reading (De Varda et al., 2024) align’s LM’s predictions with human eye tracking and reading times. Higher is better for all metrics.

A.4 McNemar Comparisons Per Maxim

Table 8 shows pairwise Maxim-specific McNemar results for Experiment III.

A.5 Details of Error Analysis

To quantify Maxim difficulty for the models of Experiment I, we count the number of models that fail for each evaluation item, and divide it by the number of models (N=6). We then define Difficulty by deducting this value it from 1, which returns a value from 0 to 1, with 0 meaning that all models succeeded in identifying the Maxim-conforming answer for each item. The Maxim-difficulty is then the average Difficulty value of the corresponding items for each Maxim (N=450). Figure 5 shows two examples for each Maxim with varying difficulties.

Following that, we calculate semantic similarity between the two candidate answers as cosine similarity using all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), where higher values indicate more semantically similar and potentially harder

to distinguish answer pairs. Furthermore, items were tokenized using the Pythia tokenizer and token length difference was computed by subtracting the token count of the Maxim-violating answer from the conforming answer; positive values therefore indicate that the correct answer is longer.

Quantity I	
d=1	Q: How do you prefer your cereal? F: With cold milk. V: In a bowl.
d=0	Q: What did you have for dinner? F: Spaghetti. V: Some food.
Quantity II	
d=1	Q: Where did you go last weekend? F: I went to grandma's and baked cookies. V: I went to grandma's and didn't stay in my room.
d=0	Q: What games do you like? F: I like tag and hide-and-seek which are fun to play. V: I like tag and hide-and-seek which are games.
Quality	
d=1	Q: Why don't you come outside? F: Because I'm helping mom bake. V: Because I'm on a spaceship.
d=0	Q: Where do you eat dinner? F: I eat dinner in the kitchen. V: I eat dinner on a cloud.
Relation	
d=1	Q: What's your favourite thing to draw? F: I like drawing animals. V: I like eating pizza.
d=0	Q: Which TV programme do you like most? F: My favourite is cartoons. V: My favourite is sandwiches.
Politeness	
d=1	Q: Is it okay if I use your crayons? F: No, sorry, I forgot them at home. V: No, you're not allowed.
d=0	Q: Want to try a bit of my cookies? F: No, I'm full, but thank you. V: No, they look gross.

Figure 5: Evaluation items at two difficulty levels per Maxim. **Q** = question, **F** = Maxim-conforming answer, **V** = Maxim-violating answer. d=1: all six pretrained models failed; d=0: all models succeeded.

A.6 Use of AI assistants

Throughout the present work, AI assistants were deployed as tools for refining drafts, spell checking, and some of the visualizations. The ideas, analyses and discussions were all authentic contributions of authors.