PRAGWORLD: A Benchmark Evaluating LLMs' *Local World Model* under Minimal Linguistic Alterations and Conversational Dynamics

Sachin Vashistha[1], **Aryan Bibhuti**[1], **Atharva Naik**[2], **Martin Tutek**[3], **Somak Aditya**[1]

[1]Indian Institute of Technology Kharagpur, India
[2] LTI, Carnegie Mellon University
[3] University of Zagreb

Abstract

Real-world conversations are rich with pragmatic elements, such as entity mentions, references, and implicatures. Understanding such nuances is a requirement for successful natural communication, and often requires building a local *world model* which encodes such elements and captures the dynamics of their evolving states. However, it is not well-understood whether language models (LMs) construct or maintain a robust implicit representation of conversations. In this work, we evaluate the ability of LMs to encode and update their internal world model in dyadic conversations and test their *malleability* under linguistic alterations. To facilitate this, we apply seven minimal linguistic alterations to conversations sourced from popular conversational QA datasets and construct a benchmark with two variants (i.e., Manual and Synthetic) comprising yes-no questions. We evaluate nine open and one closed source LMs and observe that they struggle to maintain robust accuracy. Our analysis unveils that LMs struggle to memorize crucial details, such as tracking entities under linguistic alterations to conversations. We then propose a dual-perspective interpretability framework which identifies transformer layers that are *useful* or *harmful* and highlights linguistic alterations most influenced by harmful layers, typically due to encoding spurious signals or relying on shortcuts. Inspired by these insights, we propose two layer-regularization based fine-tuning strategies that suppress the effect of the harmful layers.

Code, Data, and Extended version — https://github.com/SachinVashisth/PRAGWORLD

1 Introduction

The human ability to comprehend natural language is often owed to our innate skills to utilize relevant world knowledge, the ability to map words (symbols) to meaningful concepts (abstract or concrete entities) in the world – thus creating a mental state of the environment. Forming such a mental model of the world further endows us to suitably calibrate our responses or actions. On the other hand, Transformer-based language models (LMs) achieve impressive language comprehension ability solely by learning from vast amounts of unstructured text. Thus, NLP researchers face a looming question: *Do LMs construct an implicit world model of the environment described in the input* (Bender and Koller 2020)?

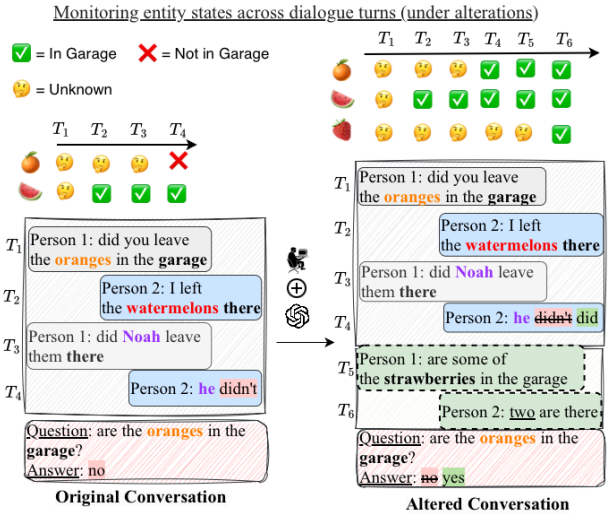


Figure 1: Example alteration applied to a conversation from GRICE (Zheng et al. 2021). Linguistic alterations modify objects or their states and introduce or eliminate new agents or entities. $T_i \ \forall i \in \{1, 2, \ldots, n\}$ denotes the i^{th}utterance.

Despite being trained only on text, a number of works present evidence that LMs implicitly encode nuanced information about the world such as color (Abdou et al. 2021), gender (Bolukbasi et al. 2016) or space (Gurnee and Tegmark 2023), even going so far as representing the board state of chess (Toshniwal et al. 2022; Li et al. 2022; Kuo, Hsueh, and Tsai 2023; Karvonen 2024). Such findings suggest that LMs are able to develop latent representations which encode a *world model*, a capability which greatly exceeds the pretraining task of next-token prediction. However, Vafa et al. (2024) shows that such latent world models of LMs are fragile in tasks such as game playing, logic puzzles and navigation.

Therefore, we proceed with benchmarking the aspect of *malleability* of latent world states – a notion complementary to fidelity and expressiveness – by investigating whether an LM is able to adjust accordingly to new information. To model such dynamics, we opt for a *conversational* setting and probe about how knowledge about concrete entities is updated throughout the conversation. We source seed conver-

sations from existing datasets: GRICE (Zheng et al. 2021) and CICERO (Ghosal et al. 2022). These datasets contain conversations riddled with pragmatic nuances, paired with questions. We introduce various label-preserving or label-altering linguistic alterations to those conversations. In Figure 1, we show how a simple linguistic alteration can *inadvertently* change the outcome of the question about the entity "orange".

Previous benchmarks for entity tracking primarily present the context as a sequence of unambiguously stated facts (Kim and Schuster 2023; Tandon et al. 2020). In this work, we take a step further. We blend in conversational dynamics and evaluate entity (or state) tracking ability under the presence of well-defined minimal linguistic alterations: negation, variable swap, quantity change, variable substitution, quantifier change, logical connective change, and injecting local knowledge. We then leverage manual and semi-automatic methods to create alterations of the seed conversations and create 2614 distinct instances, constituting the manual and synthetic splits of our dataset: PRAGWORLD. We then benchmark a wide range of open and closed-source LMs on PRAGWORLD and show that the models are not *robustly accurate* — i.e., accurate on both the original instances and their altered variants – indicating gaps in reading comprehension or memorization of LMs. In order to understand where LMs fail, we design a dual-perspective interpretability framework using direct effect patching and MLP zero-out ablation. Using this framework, we trace performance issues to fragility in entity state tracking by identifying harmful model layers. Finally, we design regularization techniques that help reduce the unwanted effect of harmful layers, in turn improving robustness towards proposed linguistic alterations.

Concretely, our **contributions** are as follows:

- We apply seven minimal linguistic alterations to seed conversations from the GRICE and CICERO datasets, creating our benchmark, PRAGWORLD, which evaluates *malleability* of LLMs' internal representations.
- We evaluate a wide range of open and closed source LMs on PRAGWORLD and show they are not *robustly accurate* under linguistic alteration, indicating brittleness of the LM's latent world models.
- We propose a dual-perspective interpretability framework using **Direct Effect Patching** and **MLP zero-out Ablation** to pinpoint layers that encode *useful*, or *harmful* reasoning patterns.
- Based on insights of the previous framework, we design and evaluate two regularization techniques: a) *Useful Layer Amplification*, and b) *Harmful Layer Suppression*, which help suppress the effect of harmful layers.

2 Related Work

Do LMs Encode World Models? A number of works explore whether the internal representations of language models recoverably encode something akin to a world model *isomorphic* to the real one (Merrill et al. 2021; Patel and Pavlick 2022; Vafa et al. 2024). Works show that attributes such as colors, (Abdou et al. 2021), gender (Bolukbasi et al. 2016) and directions (Gurnee and Tegmark 2023) can be faithfully recovered from the internal model representations. In this work, we take these analyses a step further and evaluate whether LMs are able to precisely detect, encode, and update local world states described in the course of dyadic conversations. In order to keep track of objects, locations, and people in conversations, plan ahead in games, or reason counterfactually, LMs need to create and frequently dynamically update a local world model. We evaluate the precision of this encoding through minimal linguistic alterations applied to conversations. We compare the complexity evaluated within our benchmark to related works in Supplementary Table 14.

Complexities of the Conversational Format. Conversational dynamics represent the complexity of the real world more accurately compared to static settings by introducing pragmatic elements, implicatures and commonsense inference intertwined with the core task (Sap et al. 2019; Zheng et al. 2021; Ghosal et al. 2022; Li, Zhu, and Zheng 2023). We source our benchmark based on the following two conversational datasets. GRICE (Zheng et al. 2021) is an automatically generated dialogue dataset which evaluates the capabilities of LMs for pragmatic reasoning (Goodman and Frank 2016), resolution of implicatures (Grice 1975; Borg 2009) and coreferences as well as entity tracking. CICERO (Ghosal et al. 2022) is a conversational dataset necessitating a broad number of inference types and commonsense reasoning.

Tracking States within Conversations. We opt for GRICE and CICERO as the conversation structure they encode, contains frequent references to objects, events or people which change locations or states. The capability of LMs to keep track of complex states has been a subject of a number of works (Prakash et al. 2024; Puerto et al. 2024; Kim, Schuster, and Toshniwal 2024), which have revealed that most models struggle with memorizing, and updating, states of entities mentioned in context. Within our work, we evaluate conversational elements as well as entity tracking by manually and automatically introducing local alterations to conversational contexts which alter its semantics. In this way, we evaluate the robustness of LMs performance on such tasks, providing stronger guarantees for their practical usage.

3 Robustness of LM Representations Under Lexically Minimal Alterations

In our work, we evaluate the effect of minimal alterations to dialogue-based contexts on the precision and stability of LMs latent representations, and in turn on their capability to accurately answer questions. Here, we outline the key properties of LMs targeted by our alteration-based benchmark.

Problem Definition. Let $x \in \mathcal{X}$ represent a conversational context consisting of a sequence of utterance pairs u_i between two speakers, $q \in \mathcal{Q}$ be a question posed about this conversation. We study the capability of the LM as a function $f : \mathcal{X} \times \mathcal{Q} \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ is the space of possible answers. In our work, we focus on yes-no questions, so $\mathcal{Y} = \{\text{yes}, \text{no}\}$. Let us, without loss of generality, separate f into two components: the encoding of q, x into the latent space $h = e(q, x)$ and answer generation based on the latent representation $y = d(h)$ (decoding). In our work, we consider how each Transformer layer encodes the conversation, $h = e(q, x) = [h_l]_{\ell=1}^{L}$.

Consider an alteration function $\delta : \mathcal{X} \times \mathcal{Q} \rightarrow \mathcal{X} \times \mathcal{Q}$, which introduces minimal changes to the original conversation and question. The altered input is then given as $\hat{x}, \hat{q} = \delta(x, q)$ and its corresponding latent space encoding and answers are $\hat{h} = [\hat{h}_l]_{l=1}^{L}$ and $\hat{y}$. We design the alterations to differ minimally in terms of context tokens, while still causing a large impact on conversation semantics.

Injectivity. Formally, a function f is injective if it maps distinct inputs to distinct outputs $f(x) \neq f(\hat{x})$ when $x \neq \hat{x}$. Given the high dimensionality of the hidden representations of modern LMs, the encoding $e(.)$ is likely to be injective. In our work, we investigate a *soft* notion of injectivity through representation similarity of the latent conversation encoding h. Since the same question can be asked for minimally varying contexts, it *should* encode the full conversation context in h. Therefore, we expect cases where the encoding fails to capture nuances introduced by alterations to have high similarity with the base conversation. We measure this by calculating the direct effect, i.e., the change *in confidence* through a causal intervention on the original token stream by replacing all "aligned" token representations from the altered token stream. But, the effect of the "altered tokens" will be present as we are injecting all of the downstream, previous layer's self-attention effects. Formally, let $(x{+}q, y^{\text{gold}}),\ (\hat{x}{+}\hat{q}, \hat{y}^{\text{gold}}) \in \mathcal{D}$ be an original–altered pair in our dataset. Let the confidence corresponding to the **Original Run (OR)** and **Altered Run (AR)** be $P\big(\hat{y}^{\text{gold}} \mid x + q\big)$ and $P\big(\hat{y}^{\text{gold}} \mid \hat{x}{+}\hat{q}\big)$ respectively. Let $R_l^{\mathcal{P}} = \texttt{Residual}_l(\mathcal{P})$ be the layer ℓ residual stream activations produced by a standard forward pass with $\mathcal{P}$ as the input prompt. Following Chattopadhyay et al. (2019), we now define **Direct Effect** at layer ℓ as the change in answer probability obtained by patching the altered residuals in the original run:

$$DE(R_\ell^{\hat{x}} \rightarrow R_\ell^{x}) = P\big(\hat{y}^{\text{gold}} \mid x + q;\, \texttt{patch}_\ell(R_\ell^{\hat{x}})\big) - P\big(\hat{y}^{\text{gold}} \mid x + q\big)$$

where, $\texttt{patch}_\ell()$ is the patching operator that replaces layer ℓ residuals with $R_\ell^{\hat{x}}$, and $R_\ell^{\hat{x}} \rightarrow R_\ell^{x}$ indicates that we patch from the altered run into the original run.

Local World Model Stability. LMs need to keep track of the rich conversational state after each utterance to recover relevant information. Following work that shows LMs decodably encode world states (Toshniwal et al. 2022; Karvonen 2024), we hypothesize the same holds for conversational contexts, where LM layers encode conversation states. We evaluate the capacity of the model to encode altered inputs (through direct effect analysis), decodability of the information from the world model (through robust accuracy), and estimate the stability of the internal layer-wise representation encoded by the model (direct effect analysis and MLP zero-out ablation). Formally, let $P^{(\text{MLP}_\ell \rightarrow 0)}(y \mid x + q)$ denote the model's predicted probability when the MLP submodule at layer ℓ is zeroed out, and $P(y \mid x + q)$ denote the probability when no such intervention is applied. Then the predicted label for the case of MLP zero-out ablation is given by: $\hat{y} = \arg\max_{\text{y} \in \{\texttt{yes},\texttt{no}\}} P^{(\text{MLP}_\ell \rightarrow 0)}\big(y \mid x + q\big)$
Let A_ℓ denotes the accuracy over the entire dataset $\mathcal{D}$ after zeroing out MLP submodule at layer ℓ, and A_0 be the accuracy when no such intervention is applied. We classify layer ℓ as *useful* if $A_\ell < A_0$, and *harmful* if $A_\ell > A_0$.

4 Dataset Curation

We source PRAGWORLD from GRICE (Zheng et al. 2021) and CICERO (Ghosal et al. 2022). These datasets are conversational question answering datasets where the conversation determines a "local" context, or world, in the form of an alternating conversation between two people/agents, followed by a probative yes/no question about the state of entities or agents in this world.[1] To probe if the LM's world models in conversational contexts are malleable, we devise minimal lexical alterations that manipulate entities or perform local semantic changes (e.g., negating actions or statements). If the LM has a sufficiently malleable world model, it should accurately answer questions pertaining to both the original and altered instances.

Selection of Seed Conversational Data. We manually select 44 representative seed conversations from GRICE and 33 from CICERO, respectively. We show an example of seed conversation from both the GRICE and CICERO datasets in the Supplementary material Table 7.

Linguistic Alteration Categories. We devise the alterations in such a way that they introduce *minimal* lexical changes which strongly affect the local world model and the semantics of the utterances. Such alterations will evaluate the capability of LMs to track and update states of entities throughout the conversation. We manually or semi-automatically apply the following 7 lexical changes when possible to the seed conversations.

1) Negation (Neg). We negate auxiliary verbs such as "are", "aren't", "did", etc., in a way that changes the truth value of a proposition in the conversational context. E.g., if the original conversation has a pair of turns: "A: Did Noah leave oranges in the garage, B: he didn't" ⇒ 'A: Did Noah leave oranges in the garage, B: he did", making it so that at least some oranges can be found in the garage.

2) Variable Substitution (V-Su). We substitute an entity or object in a conversational context with another one. E.g. "A: Did you leave oranges in the garage?, B: I left the watermelons there." ⇒ "A: Did you leave the oranges in the garage?, B: I left the oranges there.".

3) Quantity change (Qty-Ch). We change the quantity of countable noun entities. E.g., "A: Are some strawberries in the backyard?, B: Two are there." ⇒ "A: Are some strawberries in the backyard?, B: Three are there.".

4) Variable Swap (V-Sw). We swap an entity or object in a conversational context with another one. E.g. "A: Did you leave oranges in the garage?, B: I left the watermelons there." ⇒ "A: Did you leave the watermelons in the garage?, B: I left the oranges there."

5) Quantifier change (Qtf-Ch). We manipulate a quantifier determiner, modifying an entity in a way that alters the implied quantity of the entity. E.g., "B: All grapefruits

[1] We specifically filter the instances in both of these datasets to only include questions with yes/no answers.

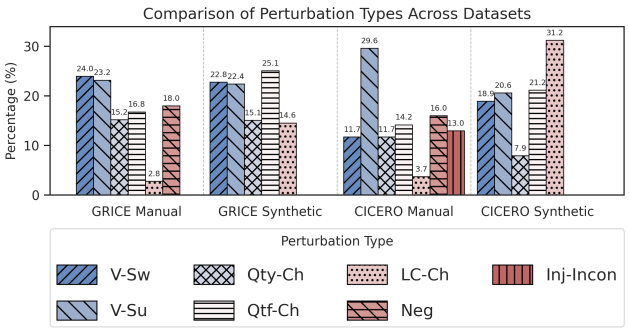


Figure 2: Percentage of conversations with each linguistic alteration category for both splits of PRAGWORLD.

are in the playroom." ⇒ "B: Some grapefruits are in the playroom.". Importantly, note that the altered turn is technically implied by the original turn; however, due to scalar implicature (Carston 1998) "some" is interpreted as "not all" assuming that the speakers in the context are following the cooperative principle.

6) Logical Connective Change (LC-Ch). We manipulate a conjunction combining multiple propositions. E.g., "B: Jack put apples and oranges in the backyard" ⇒ "B: Jack put apples or oranges in the backyard.". This alteration changes the conversational turn from an assertion of two propositions to an uncertain expression of either one of them being (exclusively; Jennings 1994) true.

7) Injecting Inconsistent Data (Inj-Incon). To ascertain whether models utilize their commonsense knowledge, we inject plausible information that may contradict with common sense. E.g., " B: And a big birthday cake too, with fifty candles.." ⇒ "And a birthday cake that changes color every time someone claps, with fifty candles".

Manual Dataset Construction. We manually apply these linguistic alterations to create 500 (300 from GRICE and 200 from CICERO) conversations from the 77 seed conversations. We refer to this as the *manual* split of the PRAGWORLD dataset. Each example is annotated by one author and then reviewed by another author to check whether (a) the alterations are lexically minimal and meaningful, (b) ground truth answers remain logically consistent with the altered context and (c) ambiguities and inconsistencies do not create nonsensical inputs. We include an example in the final version of our benchmarks only if both authors agree.

Synthetic Dataset Generation. We generate the synthetic dataset split using a semi-automatic pipeline with the following steps. First, we take a set of 104 initial seed conversation (49 from GRICE, and 55 from CICERO). Using these, we prompt `GPT-4-turbo-2024-04-09` (OpenAI et al. 2024) to generate novel conversations similar to the seed conversations. The prompts used for the `GPT-4-turbo-2024-04-09` are shown in Supplementary Section F. Secondly, we apply minimal lexical changes to these newly generated conversations using deterministic algorithms (Algorithms 1-5 in Supplementary Section A.1) for each alteration. In the third step, we create predefined questions automatically using templates (Supplementary Section A.2) for each conversation. We consider three types of questions: a) **Quantity Questions**: Ask about a specific quantity of an object. e.g. "Did Lucas place **six** apples in the kitchen?", b) **Universal Quantifier questions**: Ask whether a statement applies to all members of a group. e.g. "Did Lucas place **all** apples in the kitchen?", c) **Existential Quantifier questions**: Ask whether a statement applies to at least one member of a group. e.g. "Did Lucas place **some** apples in the kitchen?". The answer to all the categories of questions can either be **Yes** or **No**. Finally, we manually annotate answers to questions created in the previous step to ensure correctness.

Dataset Statistics The manual split of the PRAGWORLD consists of a total of 500 altered conversations, while the synthetic split consists of 2114 such conversations. Among the latter, 1074 instances have the answer "Yes" to the question, while 1040 instances have the answer "No". Figure 2, and Supplementary Figure 7 shows the distribution of conversations with respect to the alteration categories and context length, respectively.

5 Robustness of Internal Representations

In this section, we report the performance of various open- and closed-source LMs on both splits of PRAGWORLD and perform various analyses showcasing which issues present in LMs our benchmark targets.

Baselines and Performance Metrics. We evaluate both closed-source models, such as `GPT-3.5-Turbo`, and open-source models, including `Deepseek-coder-instruct` (16B), the `Phi-3` series, the `Llama-3` series, and the `Qwen2.5` series. We provide details of the models we use in Supplementary Table 9 and the prompts used in Supplementary Table 10. Throughout our experiments, we report a number of individual and aggregate metrics. **Robust accuracy** deems a model response to be correct if and only if it correctly answers both the original conversation and all of its altered instances. We additionally report accuracy across different alteration subsets: **Flip Accuracy** (accuracy on alterations designed to change the answer), **Invariant Accuracy** (accuracy on alterations designed to maintain the answer), **Original Accuracy** (accuracy on unaltered conversations), and **Altered Accuracy** (accuracy on altered conversations). We also report individual accuracies for each gold label – "Yes" (**Yes Accuracy**) or "No" (**No Accuracy**).

5.1 Performance on PRAGWORLD

We report the performance of analyzed models on the manual and synthetic splits of PRAGWORLD in Table 1. Overall, the `Phi` series models are the best performing ones according to robust accuracy, surpassing even larger models like `GPT-3.5`. Most models report a large gap between **Yes** and **No** accuracy on both splits, indicating their preference towards one of the answer options. This disparity is especially pronounced in smaller models like `Llama-3.2-1B-Ins.`, `Llama-3.2-3B-Ins.`, and `Qwen-2.5-1.5B-Ins.` `GPT-3.5` and `DeepSeek-Inst.` also exhibit this bias, more frequently answering **No**.

	#Params	Context	Robust Acc	Yes Acc	No Acc	Original Acc	Altered Acc	Flip Acc	Invariant Acc
PRAGWORLD (manual)									
GPT-3.5	-	16k	42.86	52.71	93.72	71.43	70.90	71.94	69.23
DeepSeek-Inst	16B (2.4B active)	128k	46.94	77.26	70.85	75.51	74.13	74.82	73.63
Phi-3-mini-4k-ins.	3.8B	4k	47.96	55.60	96.86	74.49	73.88	74.82	72.53
Phi-3.5-mini-ins.	3.8B (active)	128k	48.98	66.06	86.10	78.57	74.13	74.10	74.36
Llama-3.2-1B-Ins.	1B	128k	14.29	10.83	95.96	48.98	48.76	55.4	45.05
Llama-3.2-3B-Ins.	3B	128k	20.41	16.97	**98.65**	54.08	53.23	63.31	48.35
Llama-3.1-8B-Ins.	8B	128k	48.98	54.87	94.62	74.49	72.14	74.82	70.33
Qwen2.5-0.5B-Ins.	500M	128k	19.39	57.04	54.71	48.98	57.71	69.78	50.92
Qwen2.5-1.5B-Ins.	1.5B	128k	22.45	**93.50**	28.25	65.31	64.18	56.12	68.50
Qwen2.5-7B-Ins.	7B	128k	37.76	47.65	95.96	68.37	69.40	72.66	66.67
PRAGWORLD (synthetic)									
GPT-3.5	-	16k	67.21	80.63	91.44	**87.04**	**83.35**	**82.85**	**87.38**
DeepSeek-Inst	16B (2.4B active)	128k	60.93	93.95	64.13	80.97	76.68	76.25	81.00
Phi-3-mini-4k-ins.	3.8B	4k	**64.78**	79.14	87.02	84.62	80.35	80.15	84.11
Phi-3.5-mini-ins.	3.8B (active)	128k	63.97	90.69	70.38	83.60	77.70	78.10	81.31
Llama-3.2-1B-Ins.	1B	128k	47.77	22.16	95.10	58.91	56.25	57.12	56.23
Llama-3.2-3B-Ins.	3B	128k	47.98	24.39	**96.63**	61.13	57.99	58.58	59.03
Llama-3.1-8B-Ins.	8B	128k	60.93	63.69	94.90	80.77	76.44	76.45	79.75
Qwen2.5-0.5B-Ins.	500M	128k	58.50	70.30	80.96	75.91	73.44	72.89	76.64
Qwen2.5-1.5B-Ins.	1.5B	128k	55.87	**96.28**	45.67	74.09	68.69	68.93	72.27
Qwen2.5-7B-Ins.	7B	128k	60.73	70.39	90.96	82.19	77.88	78.30	80.22

Table 1: Model performance on PRAGWORLD. Deepseek-Inst: Deepseek-coder-V2-Lite-Instruct, GPT-3.5: GPT-3.5-Turbo (Dec 2023). Best and second best performances are represented by *bold*, and *underline* respectively.

	#Params	Context Size	Robust Acc	Yes Acc	No Acc	Original Acc	Altered Acc	Flip Acc	Invariant Acc
PRAGWORLD (manual)									
Phi-3-mini-4k-ins.	3.8B	4k	50.00(+2.04)	66.43	**91.03**	78.57	77.11	75.54	77.29
Phi-3.5-mini-ins.	3.8B (active)	128k	52.04 (+3.06)	70.76	86.55	81.63	76.87	76.98	76.56
Llama-3.2-1B-Ins.	1B	128k	32.65 (+18.36)	59.21	78.92	67.35	68.16	68.35	68.50
Llama-3.2-3B-Ins.	3B	128k	48.98(+28.57)	63.9	87.44	77.55	73.63	79.14	70.70
Llama-3.1-8B-Ins.	8B	128k	**59.18**(+10.2)	**79.06**	84.3	**87.76**	**79.85**	76.98	**81.68**
Qwen2.5-0.5B-Ins.	500M	128k	22.45(+3.06)	44.77	73.09	60.20	56.72	66.91	52.01
Qwen2.5-1.5B-Ins.	1.5B	128k	47.96(+25.51)	71.12	76.23	74.49	73.13	77.70	71.06
Qwen2.5-7B-Ins.	7B	128k	55.10(+17.34)	70.04	**91.03**	80.61	79.10	**82.01**	76.92

Table 2: Performance of best best-performing LMs after fine-tuning on the synthetic split. The relative improvement in robust accuracy over the non-finetuned counterparts given in **gray**.

Fine-tuning on Synthetic Data. To improve robust accuracy and reduce the performance discrepancies between Yes and No accuracy, we fine-tune the models on the synthetic data split and evaluate them on the manual split. We report performance of fine-tuned models on the manual split of PRAGWORLD in Table 2. All tuned models show improvements in Robust Accuracy, with notable gains for the `Llama-3` and `Qwen-2.5` series. Smaller models like `Llama-3.2-1B-Ins.` and `Qwen-2.5-1.5B-Ins.` also show large relative improvements in Robust Accuracy, indicating that fine-tuning on synthetic data is particularly beneficial for smaller models. These results suggest that the synthetic split encodes valuable training signal that helps models better handle input alterations. Despite these improvements, the persistent gap between Yes and No Accuracy in some models suggests that fine-tuning alone does not mitigate inherent biases in LMs.

5.2 Benchmarking Entity Tracking Ability of LMs

To evaluate whether LMs are able to accurately keep track of and update the local world state throughout the conversation, we conduct the following experiment. We segment the original conversations (from the manual split of PRAGWORLD) after some number of utterances, generating progressively longer alterations or parts of the original conversation. We select 15 original conversations and partition them into three or four groups based on the dataset they were seeded from. For each of these alterations, we create an equal proportion of Yes/No questions regarding entities and people mentioned in the conversation. We then probe the models with these questions in order to determine whether they accurately memorize information from the conversational context. The final dataset

PragWorld (manual)						
	Base Model			Fine-tuned Model		
	Yes Acc.	No Acc.	Total Acc.	Yes Acc.	No Acc.	Total Acc.
GPT-3.5	79.12	92.85	**85.98**	-	-	-
Deepseek-inst	81.86	79.12	80.49	-	-	-
Phi-3-mini-4k-ins.	55.49	96.70	76.09	71.42	**95.05**	83.24
Phi-3.5-mini-ins.	78.57	84.61	81.59	70.87	94.50	82.69
Llama-3.2-1B-Ins.	6.04	**98.35**	52.19	53.29	90.65	71.97
Llama-3.2-3B-Ins.	39.01	93.40	66.20	70.87	90.65	80.76
Llama-3.1-8B-Ins.	81.86	88.46	85.16	**84.61**	91.20	**87.91**
Qwen2.5-0.5B-Ins.	**100.00**	3.29	51.64	56.04	80.21	68.13
Qwen2.5-1.5B-Ins.	80.21	65.38	72.80	74.17	85.71	79.94
Qwen2.5-7B-Ins.	63.73	93.95	78.84	73.07	92.85	82.96

Table 3: Results of the probing experiment regarding the capability of LMs to memorize information from conversations. Generally, models are able to recall most of the information presented in conversational context. We find that this capability also improves with scale.

for the entity tracking experiment contains 364 instances with a balanced proportion of "Yes" and "No" answers.

We report model performance on the entity tracking experiment in Table 3. We see that fine-tuning consistently improves entity tracking performance across all models, especially models of smaller sizes like `Llama-3.2-1B-Ins.`, and `Qwen2.5-0.5B-Ins.` Larger base models like `Llama-3.1-8B-Ins.`, and `Qwen2.5-7B-Ins.` initially perform well, indicating that the capacity for entity tracking is present within the models. Overall, our findings suggest that fine-tuning and model scale play a crucial role in enhancing entity tracking.

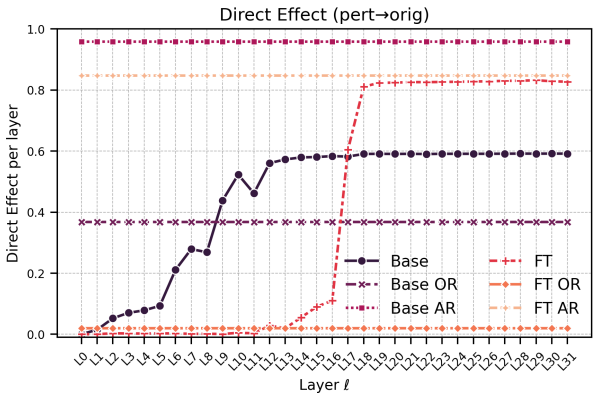
5.3 Effect of Conversation Length and Answer Type

We observe that the performance gap between Yes and No accuracy is much more pronounced in the manual split compared to the synthetic one. Supplementary Table 5 shows the distribution of Yes and No accuracy of `GPT-3.5-Turbo` w.r.t alteration categories for both splits of PragWorld. One possible explanation is that the manual split contains a greater variety of questions compared to the synthetic one. We find that `GPT-3.5-Turbo` particularly struggles with two specific question varieties that are not present in the synthetic split due to the difficulty of automatically generating such variations, as seen in Supplementary Table 8.

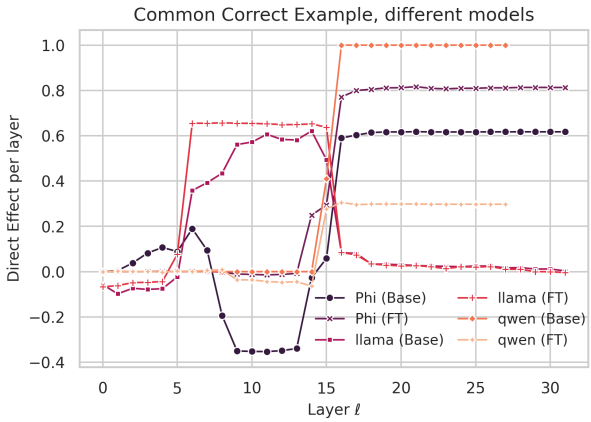
We also analyze the effect of conversation length, in the number of utterances, on accuracy. We bucket conversations into three categories: Short, with less than 11 utterances, Medium, between 11 to 15 utterances, and Long, with conversations longer than 15 utterances. Supplementary Table 6 shows that the accuracy of `GPT-3.5-Turbo` decreases from Short to Long conversations only in case of "No" accuracy in Synthetic split. It shows that accuracy is affected by many factors such as length, type of questions, model type, and type of perturbation. We investigate the confounding effects of pragmatic phenomena in Supplementary Section J.

6 Insights from a Dual-Perspective Mechanistic Interpretability Framework

To analyze internal representations, we perform *causal interventions* at the residual stream, and MLP sublayers, resulting in two techniques: 1) Direct Effect Patching, 2) MLP Zero-Out ablation (introduced before). We follow Joshi et al. (2025) and Geva et al. (2023) respectively, but explain the differences in Supplementary section G.



(a)



(b)

Figure 3: (a) Comparison of the Direct Effect Probing the Base and Fine-tuned (**FT**) versions of the `Phi-3.5-mini-instruct` for an example where both the base and fine-tuned versions are correct on both the original and altered conversation. (b) DE patching for three models on a common example, where the Base and Fine-tuned (**FT**) models are correct on original and altered instances.

6.1 Results of Direct Effect Patching

We now study the base and Fine-tuned variants of the `Phi-3.5-mini-instruct` model. We first choose an original example and a corresponding altered variant, where both the base and finetuned models predicts the label(s) *correctly*. Fig. 3a shows that for the chosen *correct* example, fine-tuned model confidence for the correct altered label increases sharply. The fine-tuned model is robust to patching up to layer

16, but its confidence changes dramatically as we move from layer 16 to 17. In addition, the fine-tuned version shows a low **OR**, and high **AR** confidence as compared to the base version. We observe a similar effect for a sample where the base model is incorrect, while the fine-tuned version is correct on both the original and altered conversation, shown in Figure 8 of Supplementary. We expand our analysis to also include `Llama-3.1-8B-Ins`, and `Qwen2.5-7B-Ins` and compare direct effect patching for the same instance where all models are correct in Figure 3b. The base and fine-tuned versions of Phi and Qwen show an increase in the correct altered label confidence. Their fine-tuned versions also show robustness towards alterations in early layers, and then a sudden increase in confidence between layers 14 and 15. For both Llama versions, we observe an opposite trend where the initial layers show a sudden increase in confidence (at layer 5), which decreases drastically moving from layer 15 to 16.

6.2 Results of MLP zero-out Ablation

We report the results of *MLP zero-out ablation* for Phi-3.5 (Base) in Figure 5. We find that *MLP zero-out* at layers $2, 9$, and 16 (*blue line*) leads to a decrease in accuracy indicating that these layers are *useful*. Ablating layers $5, 6, 7, 11, 13, 17$, and 31 results in improved accuracy, suggesting that these are *harmful* containing spurious signals or shortcut patterns. We investigate the effect of MLP zero-out ablation on different linguistic alterations. As shown in Figure 4, Logical Connective Change (LC-Ch) is most impacted by harmful layers, while Variable Swap (V-Sw) is the least affected. We also find that these two cases are less influenced by harmful layers in the fine-tuned (FT) version of the same model in the same figure. LC-Ch is less affected in the early layers of the fine-tuned version compared to the base version, while the accuracy decreases in the final layers after ablating. The latter also holds for V-Sw. Hence, we find fine-tuning helps suppress the effect of the harmful layers of the Phi model.

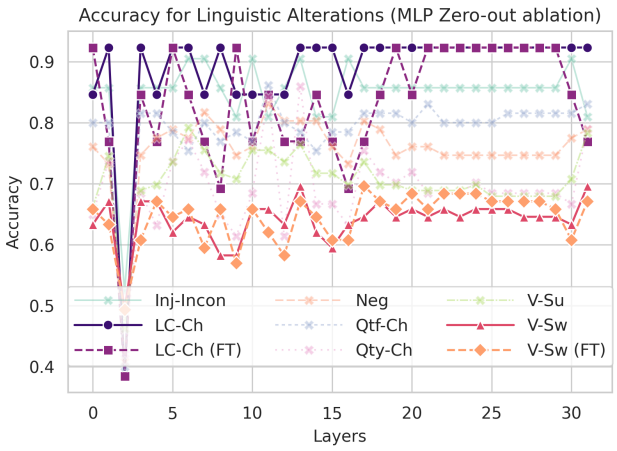


Figure 4: Effect of MLP zero-out ablation on the accuracies (dataset-wide) for different alteration types of base `Phi-3.5-mini-ins`. For LC-Ch, and V-Sw, we also report results for the fine-tuned (FT) model. Other alteration types are transparent for emphasis.

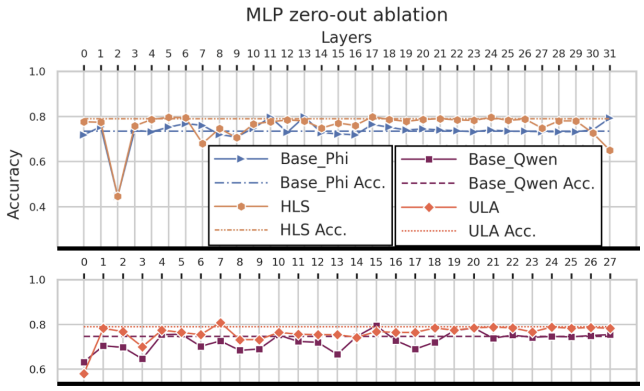


Figure 5: Results of MLP zero-out ablation on `Phi-3.5-mini-ins`, and `Qwen2.5-7B-Ins` fine-tuned using HLS and ULA. Base accuracy denotes the base model accuracy (Acc.) on the entire dataset, and HLS/ULA accuracy is the accuracy after finetuning using HLS/ULA.

6.3 Effect of additional Layer Regularization

We design two regularization strategies based on previous observations: a) *Useful Layer Amplification*, and b) *Harmful Layer Suppression*. In *Useful Layer Amplification* (ULA), we choose a set of useful layers and attach a small two-layer classification head on top of every layer's MLP output. The ULA loss is the mean of classification losses in all useful layers, which we add to the next-token prediction loss, weighted by a *useful weight* $\alpha \in [0, 1]$. In *Harmful Layer Suppression* (HLS), we choose a set of harmful layers, and add an L_2 penalty on MLP's output (the residual output after the feed-forward submodule) of each such layer. The final loss is the mean squared norm of those activations, which is then added to the model's next-token prediction loss, weighted by a *harmful weight* $\beta \in [0, 1]$ Figure 5 reports the effect of HLS (*orange line*) (for harmful layers $7, 11, 13, 17$, and 31 visible in Base_Phi with *harmful weight* = $1.0e$-3) on `Phi-3.5-mini-instruct`. The same figure reports the harmful and useful layers for the base version of the model `Qwen2.5-7B-Instruct`, and the effect of ULA strategy (*red line*) (useful layers 0, 3, and 13 visible in Base_Qwen with *useful weight* = $1.0e$-3), where accuracy decreases for both useful layers and harmful layers. All layers show accuracy lower than *ULA accuracy*.

7 Conclusion

We introduced seven types of linguistic alterations to two dyadic conversation datasets (GRICE and CICERO), creating the PRAGWORLD benchmark with the goal to assess the malleability of LLMs' *implicit* world model on these minimal alterations through the lens of robust accuracy. We benchmark open and closed source models, and observe that they are not robust to linguistic alterations, exhibiting substantial memorization errors. Using mechanistic interpretability techniques, we pinpoint *useful* and *harmful* layers, and highlight linguistic alterations most affected by the harmful layers. Inspired by these observations, we propose a two regularization techniques that help suppress the effect of harmful layers.

References

Abdou, M.; Kulmizev, A.; Hershcovich, D.; Frank, S.; Pavlick, E.; and Søgaard, A. 2021. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Online: Association for Computational Linguistics.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Borg, E. 2009. On three theories of implicature: Default theory, relevance theory and minimalism. *International Review of Pragmatics*, 1(1): 63–83.

Carston, R. 1998. Informativeness, relevance and scalar implicature. *Pragmatics And Beyond New Series*, 179–238.

Chattopadhyay, A.; Manupriya, P.; Sarkar, A.; and Balasubramanian, V. N. 2019. Neural Network Attributions: A Causal Perspective. arXiv:1902.02302.

Geva, M.; Bastings, J.; Filippova, K.; and Globerson, A. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. arXiv:2304.14767.

Ghosal, D.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2022. CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5010–5028. Dublin, Ireland: Association for Computational Linguistics.

Goodman, N. D.; and Frank, M. C. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11): 818–829.

Grice, H. P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill.

Gurnee, W.; and Tegmark, M. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Jennings, R. E. 1994. *The genealogy of disjunction*. Oxford University Press.

Joshi, A.; Ahmad, A.; Shukla, D.; and Modi, A. 2025. Towards Quantifying Commonsense Reasoning with Mechanistic Insights. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 9633–9660. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Karvonen, A. 2024. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:2403.15498*.

Kim, N.; and Schuster, S. 2023. Entity Tracking in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3835–3855.

Kim, N.; Schuster, S.; and Toshniwal, S. 2024. Code Pretraining Improves Entity Tracking Abilities of Language Models. *ArXiv*, abs/2405.21068.

Kuo, M.-T.; Hsueh, C.-C.; and Tsai, R. T.-H. 2023. Large Language Models on the Chessboard: A Study on ChatGPT's Formal Language Comprehension and Complex Reasoning Skills. *arXiv preprint arXiv:2308.15118*.

Li, H.; Zhu, S.-C.; and Zheng, Z. 2023. DiPlomat: a dialogue dataset for situated pragmatic reasoning. *Advances in Neural Information Processing Systems*, 36: 46856–46884.

Li, K.; Hopkins, A. K.; Bau, D.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Merrill, W.; Goldberg, Y.; Schwartz, R.; and Smith, N. A. 2021. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9: 1047–1060.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Patel, R.; and Pavlick, E. 2022. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Prakash, N.; Shaham, T. R.; Haklay, T.; Belinkov, Y.; and Bau, D. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*.

Puerto, H.; Tutek, M.; Aditya, S.; Zhu, X.; and Gurevych, I. 2024. Code Prompting Elicits Conditional Reasoning Abilities in Text+ Code LLMs. *arXiv preprint arXiv:2401.10065*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035.

Tandon, N.; Sakaguchi, K.; Dalvi, B.; Rajagopal, D.; Clark, P.; Guerquin, M.; Richardson, K.; and Hovy, E. 2020. A Dataset for Tracking Entities in Open Domain Procedural Text. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6408–6417. Online: Association for Computational Linguistics.

Toshniwal, S.; Wiseman, S.; Livescu, K.; and Gimpel, K. 2022. Chess as a testbed for language model state tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11385–11393.

Vafa, K.; Chen, J. Y.; Kleinberg, J.; Mullainathan, S.; and Rambachan, A. 2024. Evaluating the World Model Implicit in a Generative Model. *arXiv preprint arXiv:2406.03689*.

Zheng, Z.; Qiu, S.; Fan, L.; Zhu, Y.; and Zhu, S.-C. 2021. GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational rEasoning. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2074–2085. Online: Association for Computational Linguistics.