

SLANG: New Concept Comprehension of Large Language Models

Lingrui Mei^{1,4} Shenghua Liu^{1,4*} Yiwei Wang^{2,3} Baolong Bi^{1,4} Xueqi Cheng^{1,4}

¹CAS Key Laboratory of AI Safety, Institute of Computing Technology, CAS

²University of California, Los Angeles ³University of California, Merced

⁴University of Chinese Academy of Sciences

meilingrui22@mails.ucas.ac.cn wangyw.evan@gmail.com {liushenghua,bibaolong23z,cxq}@ict.ac.cn

Abstract

The dynamic nature of language, particularly evident in the realm of slang and memes on the Internet, poses serious challenges to the adaptability of Large Language Models (LLMs). Traditionally anchored to static datasets, these models often struggle to keep up with the rapid linguistic evolution characteristic of online communities. This research aims to bridge this gap by enhancing LLMs' comprehension of the evolving new concepts on the Internet, without the high cost of continual retraining. In pursuit of this goal, we introduce **SLANG**, a benchmark designed to autonomously integrate novel data and assess LLMs' ability to comprehend emerging concepts, alongside **FOCUS**, an approach uses causal inference to enhance LLMs to understand new phrases and their colloquial context. Our benchmark and approach involves understanding real-world instances of linguistic shifts, serving as contextual beacons, to form more precise and contextually relevant connections between newly emerging expressions and their meanings. The empirical analysis shows that our causal inference-based approach outperforms the baseline methods in terms of precision and relevance in the comprehension of Internet slang and memes. ¹

1 Introduction

Recently, language evolution has been accelerated by the online community, which has introduced new dimensions to linguistic shifts (Varis and van Nuenen, 2017; Firth et al., 2019; Hammarström, 2016). These rapid changes in language pose serious challenges to the Large Language Models (LLMs) on understanding the newly emerging concepts (Yang et al., 2023; Sun et al., 2021).

Generally, LLMs are trained on static data (Brown et al., 2020), which limits their adaptivity to

*Corresponding author.

¹Our code is available at <https://github.com/Meirtz/FocusOnSlang-Toolbox>.

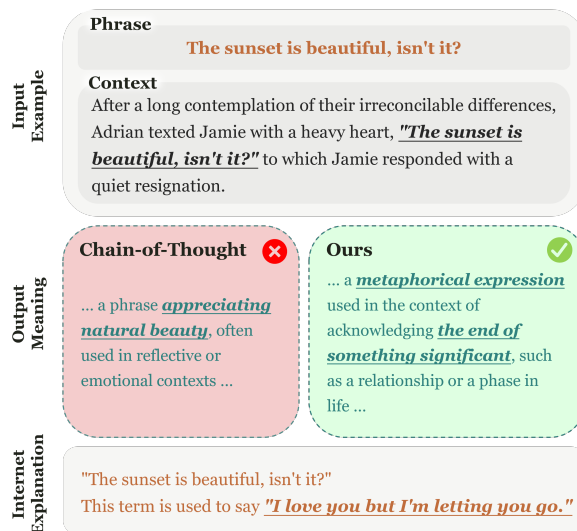


Figure 1: Comparative analysis of LLMs' understanding of new phrases using CoT (Wei et al., 2022) and FOCUS methods. The left side demonstrates the limited understanding through the CoT approach, focusing on the literal interpretation. In contrast, the right side using the FOCUS method shows the model's enhanced capability to grasp metaphors and deeper meanings.

the dynamic and ever-evolving nature of human language. This limitation is particularly pronounced in the context of digital communication, where new forms of expression and concepts emerge at an unprecedented pace (Sun et al., 2021). Hence, it is essential for LLMs to understand linguistic shifts and new concepts without constant updates or external data.

Moreover, LLMs often make decisions based on superficial patterns rather than justified reasons. This can hinder their ability to accurately interpret and follow human instructions, as highlighted in several studies (Tang et al., 2023; Wang et al., 2022, Zhou et al., 2023b, Wang et al., 2023). For example, as depicted in Figure 1, the Chain-of-Thought (CoT) prompting (Wei et al., 2022) simply interprets the phrase *The sunset is beautiful, isn't it?* and misses the deeper, metaphorical meaning, which

could represent an acknowledgment of an ending, like concluding a life phase or a relationship, in a complex conversation. This situation underscores the importance of enhancing and evaluating LLMs in a way that goes beyond their performance metrics. It’s essential to consider the fundamental principles that guide their decision-making processes.

Therefore, we propose **SLANG** (Similarity of Lexical Analysis aNd Grasp), a benchmark to assess language models’ adaptability to linguistic shifts, and **FOCUS** (Factual cOntext CaUsal analysis), an approach based on causal inference for enhancing comprehension of new concepts.

The SLANG benchmark, developed from UrbanDictionary (Urban Dictionary LLC), focuses on assessing the capability of language models to maintain coherence and accuracy in the face of dynamic and unconventional language use, such as slang and idiomatic expressions, thereby evaluating the capability of LLMs in grasping new concepts. We select recent entries after a specific cutoff date and filter out phrases already likely in LLM training data. We utilize user-generated ratings (upvotes and downvotes) to refine the dataset, ensuring its quality and comprehensiveness. The dataset is then standardized into a formal dictionary format, simplifying explanations and examples for universal understanding while retaining original meanings. This preprocessing approach ensures SLANG effectively evaluates LLMs’ adaptability to linguistic shifts.

FOCUS employs causal inference to enhance models’ comprehension of new concepts within evolving linguistic contexts. By analyzing causal relationships in language, FOCUS advances models’ predictive capabilities beyond traditional correlation-based learning. This method allows for a nuanced grasp of language dynamics, improving models’ adaptability and effectiveness in applications requiring deep understanding of language use. FOCUS significantly enhanced performance in language model comprehension, demonstrating superior precision and adaptability. With Claude 3, FOCUS achieved an F_1 score of 0.4596, precision of 0.4452, and recall of 0.4827, alongside an accuracy of 89.7%, outperforming previous methods in comprehension and adaptability.

The codes for the SLANG and FOCUS toolboxes are open-sourced, contributing to the community’s resources for advancing language model development.

2 SLANG

We introduce SLANG in response to rapidly evolving language. SLANG benchmark evaluates LLMs’ capability to interpret the dynamic landscape of user-generated new concepts. It uniquely features factual and counterfactual datasets, each crucial for gauging LLM adaptability. We detail SLANG’s dataset construction, and evaluation metrics in the following subsections.

2.1 Preprocessing

Extraction Our dataset construction involved extracting numerous concepts from UrbanDictionary, a platform known for user-generated content that reflects current language trends and the evolving internet lexicon, making it a unique, constantly updated repository and dynamic forum for new Internet language concepts. Specifically, our approach involved selecting concepts added after a predetermined date to ensure content novelty. We meticulously extract relevant data, including the phrase, its user-provided definitions, usage examples, and user-generated ratings (*upvotes* and *downvotes*). Additionally, the data construction pipeline is set up to automatically include fresh, non-member data for upcoming cutoff dates, ensuring the dataset stays up-to-date and comprehensive.

Filtering The content filtering of our dataset consisted of several steps to ensure the quality and novelty of the concepts:

- **Temporal Filtering:** Many UrbanDictionary phrases are now common and potentially included in LLMs’ training data. To ensure the novelty of our dataset, we leveraged the knowledge cut-off dates of LLMs, strategically selecting phrases that emerged after these dates. For instance, gpt-4-0613 has a knowledge cut-off date of April 2021, as detailed in the OpenAI documentation², and we selected concepts that were added to UrbanDictionary after January 2022. This temporal gap was strategically chosen to include recent phrases that may have gained popularity online after the cut-off date but were not yet recorded in UrbanDictionary.
- **User Rate Filtering:** We analyzed user-generated ratings to refine the dataset. Entries with overwhelmingly negative receptions

²<https://platform.openai.com/docs/models>

(more than 80% downvotes) were excluded. Comparative histograms in Figure 2 illustrate the distribution of *upvotes* (left) and *absolute upvotes* (right) across dataset entries before and after our cleaning process. This stringent data cleaning was crucial in ensuring the quality and reliability of the entries selected for our research. For details on the validation of user-generated votes, see Appendix E.

- **Removal of Inappropriate Content:** Inappropriate content, including NSFW material and hate speech, was removed to preserve academic integrity and ensure quality.
- **Novelty Check:** To ensure that the concepts in our dataset were unknown to the LLMs, we employed gpt-4-0613 for thorough filtration based on the method described by Yin et al., 2023. Additionally, due to the existence of models with different cut-off dates, we applied the "needle in a haystack" test (Kamradt, 2023, see Appendix D) to confirm the novelty of the knowledge. This meticulous validation process involved strategically embedding the selected phrases into extensive corpora and evaluating the models' capability to extract them.

After the temporal filtering step, we started with 7220 concepts. The subsequent steps filtered out 5463, 1328, and 21 samples respectively, resulting in a final dataset of 408 usable new concepts.

Factual dataset Following the above filtering steps, we obtained our factual dataset. Acknowledging the informal nature of the original data, we transformed it into a uniform, formal dictionary format (see Appendix B). This process involved simplifying explanations and examples for universal understanding while preserving their original meanings. We adhered to a custom-designed template for consistency and clarity across all concepts. Each explanation was enriched with four synonym-based variants to capture diverse potential responses.

Counterfactual Dataset To further evaluate the ability of LLMs in understanding new concepts, we created the counterfactual dataset. This dataset is derived from the factual dataset by preserving the original phrases while modifying the contexts and explanations. Given the factual dataset $\mathcal{D}_{\text{fact}}$, the generation process is structured as follows:

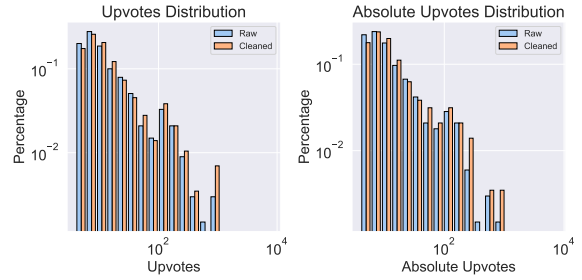


Figure 2: Comparative histograms illustrating the distribution of *upvotes* (left) and *absolute upvotes* (right) across dataset entries. Both histograms are plotted on a logarithmic scale with the vertical axis representing the log percentage of the total dataset and the horizontal axis indicating the log count of *upvotes* or *absolute upvotes*. The blue bars represent the raw data, while the orange bars depict the cleaned data, facilitating a direct comparison of the distributions before and after data cleaning.

- **Entity Extraction:** We extract the entities e from each explanation y in $\mathcal{D}_{\text{fact}}$.
- **Counterfactual Replacement:** For each entity $e_i \in e$, we generate a counterfactual entity e'_i that is conceptually divergent from the original entity e_i while retaining the original phrase structure.
- **Context Construction:** Based on the counterfactual explanations y'_i , we use GPT-4 to generate new contexts x'_i . Consequently, the counterfactual dataset \mathcal{D}_{cf} is composed of pairs: $\mathcal{D}_{\text{cf}} = \{(p_1, x'_1, y'_1), (p_2, x'_2, y'_2), \dots, (p_n, x'_n, y'_n)\}$.

This approach ensures that while the original phrases are preserved, the contexts and explanations are transformed to convey entirely different meanings. Consequently, each entry in the counterfactual dataset introduces novel concepts that is distinct from the original dataset, providing unique challenges for the LLMs to interpret and understand.

2.2 Metrics

For this task, we employed traditional metrics like F_1 score, recall, and precision (Yang et al., 2018), and added BLEU (3-gram) (Papineni et al., 2002) and ROUGE (Lin, 2004) for stricter quality checks. Considering that language models might output synonymous interpretations with varied wording, we also incorporated sentence similarity measures such as sentence-level similarity and SimCSE (Gao

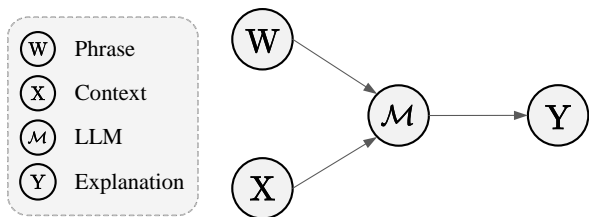


Figure 3: Structural Causal Model (SCM) of LLMs for interpreting new phrases, excluding confounders. The variables X and W encapsulate users’ complex intentions and thoughts, which span intricate emotional expressions, cultural insights, and extensive internet-specific knowledge. Grasping these nuanced aspects directly through an LLM is a challenging endeavor.

et al., 2021), and sentence similarity (Team, 2023) as *Similarity* by calculating the cosine similarity of the embeddings of all-mpnet-base-v2 (Team, 2023; Song et al., 2020). A positive sample was considered for accuracy calculation if its SimCSE score exceeded 0.7. We also generated five lexically varied yet syntactically and semantically identical interpretations for each dataset entry. To determine the final metric, we selected the interpretation with the highest BLEU score.

3 FOCUS

This section outlines the novel approach employed in our research to enhance the adaptability of LLMs in understanding the evolving human language. Our method analogizes the dynamic nature of language to a continuously evolving entity that requires adaptive comprehension strategies.

3.1 Causal Analysis

Structural Causal Models (SCMs) serve as vital tools for elucidating the relationships and influence pathways among variables. We propose a simplified SCM (see Figure 3) to delineate the interpretative processes of an LLM when confronted with novel phrases within their context. In this model, users supply both the phrase W and its context X , which are then inputted into the LLM, represented by the direct links $X \rightarrow \mathcal{M} \rightarrow Y$ and $W \rightarrow \mathcal{M} \rightarrow Y$, where \mathcal{M} denotes the LLM and Y is the output explanation. These causal links, free from confounders, capture the logical chain from input to interpretation. This SCM sets the stage for our forthcoming discussion where we reintroduce and scrutinize these confounders, thus laying the groundwork for a comprehensive causal analysis.

Analysis of entity Initially, referencing a typical SCM framework (Wang et al., 2023; Wang et al., 2022), we assume SCM $S = \{X, E, Y\}$, where X denotes context/input, and E represents confounder, including phrases and other entities within the context. From the perspective of human understanding of new phrases, the interpretation Y is derived from the context X and its entities E . The confounder E can be extracted from X , leading to the relationships $X \rightarrow Y \leftarrow E$ and $X \rightarrow E$ in our model. By applying the *do*-operation (Verma and Pearl, 1990), which denoted as $do(X)$, we follow the guidelines (Wang et al., 2023; Wang et al., 2022) to conduct a rigorous causal inference, ensuring that the main effects of the textual context are captured without losing entity information. This operation aims to isolate the effect of the context X on the confounder E . Consequently, the relationship between X and E undergoes a change, becoming $\tilde{X} \leftarrow \tilde{E}$, where \tilde{E} represents the modified entity, and \tilde{X} denotes the context obtained after the *do* operation which effectively substitutes the actual entity.

Analysis of linguistic factors In our SCM, denoted as $S = \{X, W, E, Y, R\}$, the output variable Y , as an endogenous variable, is influenced by the input context X , the input phrase W , other entities E , and finally, linguistic factors R , which include the linguistic structure, style, theme, and cultural background, crucially shape X and constitute the exogenous variables. The model includes direct paths $X \rightarrow Y$ and $W \rightarrow Y$, indicating the immediate influence of context and phrase on the interpretation. The backdoor path $E \leftarrow X \rightarrow W$ and the indirect path $X \rightarrow E \rightarrow W$ demonstrate the mediated effects. The path $R \rightarrow X$ highlights the exogenous influence of linguistic factors on context. The equation for the causal effect in this context is as follows:

$$\begin{aligned}
 & P(Y = y|X = x) \\
 &= \sum_{e,r} P(Y = y|X = x, W = w, E = e, R = r) \\
 & P(W = w|X = x, E = e, R = r) \\
 & P(E = e)P(R = r)
 \end{aligned}$$

In this formula, $P(Y = y|X = x)$ represents the probability of the outcome variable Y being a particular value y , given the context X is set to x . The summation over e and r encompasses all possible combinations of the values of the en-

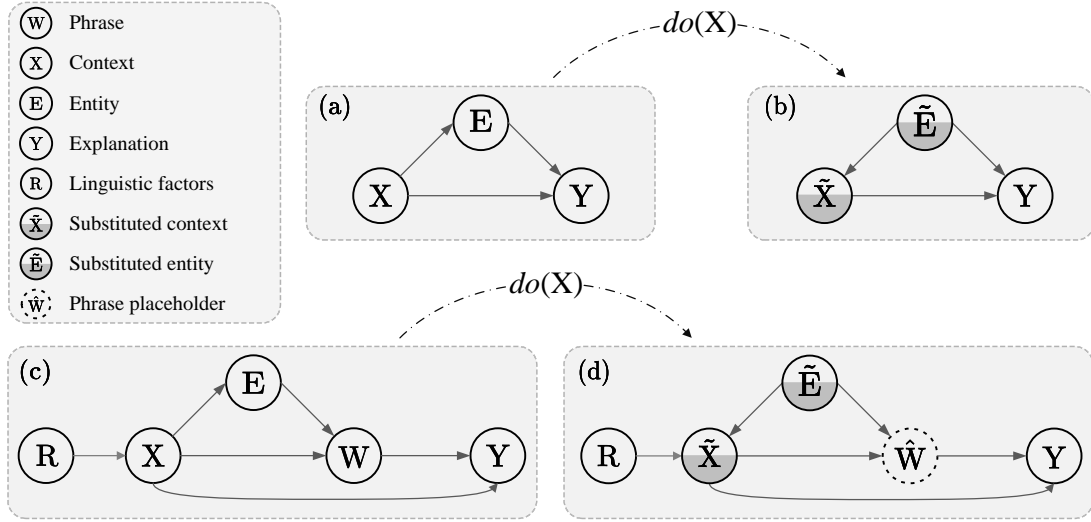


Figure 4: SCM analysis in FOCUS methodology. This figure presents the complex process of causal inference in the understanding of new phrases, highlighting how the FOCUS approach systematically analyzes and interprets intricate language patterns, emphasizing the causal links between linguistic elements and their interpretive outcomes.

ties E and linguistic factors R . For each combination, the formula calculates the conditional probability of $Y = y$ given $X = x$, the specific phrase $W = w$, the entity $E = e$, and the linguistic factor $R = r$. This conditional probability is further modulated by the probabilities $P(W = w|X = x, E = e, R = r)$, $P(E = e)$, and $P(R = r)$. These terms represent the likelihood of observing the phrase $W = w$ conditioned on the context, entities, and linguistic factors, as well as the inherent probabilities of the entities $E = e$ and the linguistic factors $R = r$. This comprehensive approach allows for a nuanced understanding of how context, entities, and linguistic factors collectively influence the interpretation Y .

3.2 Context-based Causal Intervention

Subsequently, we refine our causal intervention approach in the SCM, focusing on the role of context and linguistic factors. To concentrate on the contextual content and reduce reliance on shortcuts, E is replaced with \tilde{E} , consequently transforming X into \tilde{X} , (as illustrated in Figure 4 (a) and (b)). Further, to eliminate the entity bias in W , it is replaced with a specific placeholder, denoted as \tilde{W} (as illustrated in Figure 4 (c) and (d)). This alteration aids in isolating the effect of W while excluding entity-specific biases. Overall, these adjustments accentuate the role of R in shaping the context, enhancing the model’s capacity to highlight the influence of linguistic factors in a bias-free manner. The updated causal effect formula in the SCM is

thus:

$$\begin{aligned}
 & P(Y = y|do(X = x)) \\
 &= \sum_{\tilde{e}, r} P(Y = y|X = x, \tilde{W} = \tilde{w}, \tilde{E} = \tilde{e}, R = r) \\
 & \quad P(\tilde{W} = \tilde{w}|do(X = x), \tilde{E} = \tilde{e}, R = r) \\
 & \quad P(\tilde{E} = \tilde{e})P(R = r) \\
 &= \sum_{\tilde{e}, r} P(Y = y|X = x, \tilde{W} = \tilde{w}, \tilde{E} = \tilde{e}, R = r) \\
 & \quad P(\tilde{W} = \tilde{w}|\tilde{E} = \tilde{e}, R = r)P(\tilde{E} = \tilde{e})P(R = r)
 \end{aligned}$$

This revised formula ensure a more comprehensive and nuanced analysis of the causal dynamics within the SCM, post-intervention. This approach ensures a more robust and bias-free interpretation within the SCM framework. This equation accounts for the altered relationships in the SCM after do -operation. The conditional probabilities and summations are now over the new variables \tilde{E} and \tilde{W} , while maintaining the original structure’s intent to adjust for confounding effects and capture the influence of modified entities in the context X .

The core idea of FOCUS is to explore how language models can adhere to guidelines to better understand the content of the context. The goal is to enable LLMs to analyze phrases according to usage examples and provide counterfactual interpretations, thereby understanding the evolving semantics of language. To achieve this objective, we propose a four-stage method (as shown in Figure 5), each with its principles:

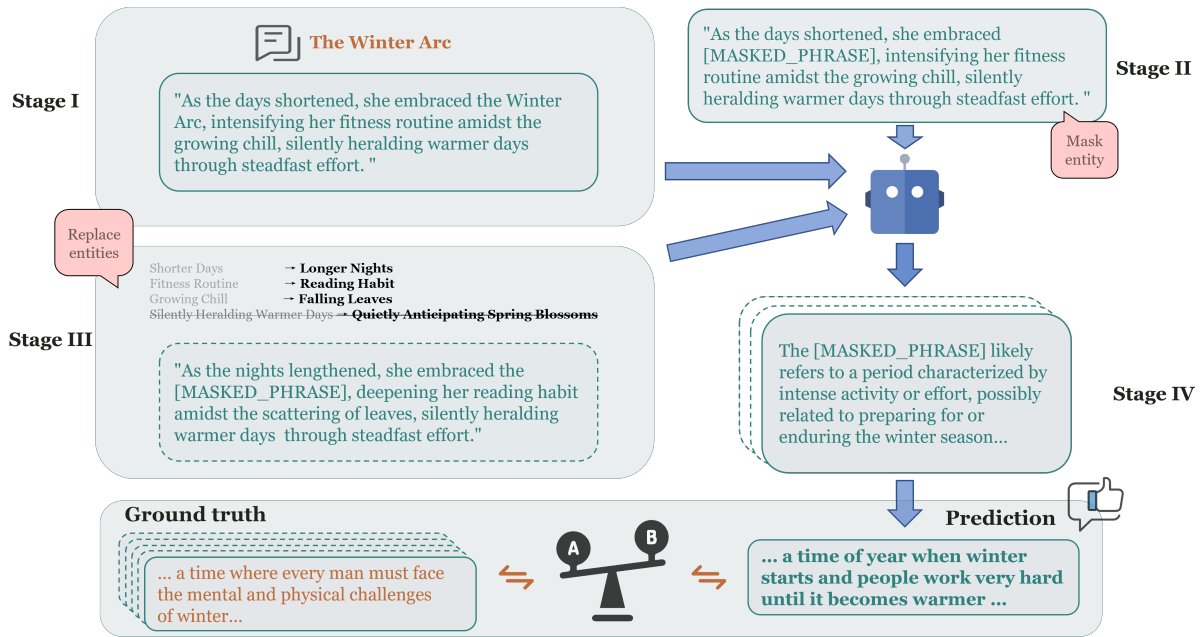


Figure 5: The four-stage pipeline of FOCUS.

Direct Inquiry (DI) In the Direct Inquiry (DI) stage, we input the usage example (context) X and phrase W into the language model. Represented as $Y_{DI} = \mathcal{M}(X, W, E)$, this stage aims to evaluate the direct effect of X on the inferred meaning Y . It allows the model to derive meaning naturally within the given context, focusing on the coherence of the phrase’s morphology, literal meaning, and context-based interpretation. DI sets a baseline for understanding the phrase’s meaning before adding more analytical layers.

Masked Entity Inquiry (MEI) In the Masked Entity Inquiry (MEI) stage, we mask the phrase W within context X to analyze the meaning Y without W ’s direct influence. This process, $Y_{MEA}, \hat{W}_{MEA} = \mathcal{M}(X_{\text{masked}}, E)$, helps the model suggest synonyms or near-synonyms for the masked phrase. MEI focuses on extracting meaning from broader linguistic factors in X and E , reducing bias towards W ’s literal interpretation and enhancing context-based understanding.

Entity Replacement Inquiry (ERI) In the Entity Replacement Inquiry (ERI) stage, we alter entities in context X to assess phrase interpretation variability. We use a dropout rate for entity alteration in $Y_{ERI}, \hat{W}_{ERI} = \mathcal{M}(\hat{X}_{\text{masked}}, \hat{E}_{\text{replaced}})$. This introduces a balance of original and new entities, enhancing model robustness without bias. ERI helps understand entity dynamics’ effects on interpretation, providing deeper causal analysis.

Synthesis (SY) In the Synthesis (SY) stage, we integrate insights from Direct Inquiry, Masked Entity Inquiry, and Entity Replacement Inquiry. Represented as $Y_{FS} = \mathcal{M}(Y_{DI}, Y_{MEA}, Y_{ERI})$, this phase evaluates the interplay between direct, contextual, and entity-variable interpretations. SY reconciles varied interpretations and confounders, refining the model’s understanding of language nuances. This final stage offers a multi-dimensional perspective on language model analytics, emphasizing contextual richness.

4 Experiments

4.1 Setup

Large Language Models Following the preprocessing method described in Section 2.1, we filtered our initial dataset of 7220 concepts, resulting in 408 new concepts. These evaluations were conducted using Claude 3, GPT-4, Mistral-7B, and other popular models. For detailed experimental setup and additional results for other models, please refer to Appendix A and Appendix C.

Baselines Baselines comprised direct inquiry (Direct) (Ouyang et al., 2022), Chain-of-Thought (CoT) (Wei et al., 2022), CauView (Wang et al., 2023), and our FOCUS approaches. In each case, the language model output was parsed and compared with the ground truth. We implement the CauView method in our experiment design by using a two-stage prompt inquiry, due to its lack of a direct inquiry step.

Model	Prompting Method	F ₁	Precision	Recall	BLUE	ROUGE			Similarity	SimCSE	ACC (%)
						1	2	L			
Mistral-7B	Direct (Ouyang et al., 2022)	0.1869	0.2655	0.1684	0.0372	0.2078	0.0481	0.1452	0.1507	0.6084	43.1
	CoT (Wei et al., 2022)	0.3453	0.3227	0.3706	0.1978	0.3657	0.1602	0.2744	0.4713	0.7132	68.5
	CauView (Wang et al., 2023)	0.3012	0.2787	0.3384	0.1561	0.3272	0.1336	0.2449	0.4862	0.7127	64.3
	FOCUS (Ours)	0.3703	0.3555	0.3878	0.2493	0.3942	0.1866	0.3011	0.5121	0.7469	76.0
GPT-4	Direct (Ouyang et al., 2022)	0.2308	0.3474	0.1917	0.0483	0.2597	0.0606	0.1859	0.1616	0.6476	47.2
	CoT (Wei et al., 2022)	0.4123	0.3947	0.4244	0.2384	0.4370	0.1927	0.3299	0.5521	0.7883	79.3
	CauView (Wang et al., 2023)	0.3602	0.3444	0.3948	0.1987	0.3917	0.1643	0.3032	0.5515	0.7636	74.7
	FOCUS (Ours)	0.4446	0.4280	0.4714	0.3177	0.4721	0.2332	0.3652	0.6032	0.8216	88.2
Claude 3	Direct (Ouyang et al., 2022)	0.2395	0.3538	0.2171	0.0552	0.2645	0.0673	0.1904	0.1947	0.6714	51.4
	CoT (Wei et al., 2022)	0.4276	0.4082	0.4606	0.2471	0.4492	0.2014	0.3371	0.5628	0.7948	81.2
	CauView (Wang et al., 2023)	0.3752	0.3550	0.4229	0.2015	0.4035	0.1707	0.3041	0.5778	0.7932	76.6
	FOCUS (Ours)	0.4596	0.4452	0.4827	0.3264	0.4835	0.2373	0.3729	0.6109	0.8354	89.7

Table 1: Performance results on the factual dataset. For results on more models, see Appendix C.1.

4.2 Results

4.3 Experimental Results

Results on the factual dataset Table 1 summarizes the findings that the FOCUS method demonstrates exceptional performance across all models. For GPT-4, FOCUS achieves the highest F₁ score of 0.4446, precision of 0.4280, recall of 0.4714, and an accuracy of 88.2%. Similarly, Claude 3 under FOCUS secures an F₁ score of 0.4596 and an accuracy of 89.7%, while Mistral 7B records an F₁ score of 0.3703 with an accuracy of 76.0%. These metrics highlight the effectiveness of the FOCUS method in enhancing the interpretative capabilities of language models.

Results on the counterfactual dataset Following the analysis of the factual dataset, we extended our evaluation to the counterfactual dataset, which presents hypothetical scenarios altering real-world language usage. As detailed in Table 2, the FOCUS method outshines other techniques, particularly with GPT-4, achieving an F₁ score of 0.4532, precision of 0.4598, recall of 0.4551, and an accuracy of 84.9%. Claude 3 also performs well under FOCUS, securing an F₁ score of 0.4636 and an accuracy of 86.8%. Mistral 7B shows a solid performance, with an F₁ score of 0.3935 and an accuracy of 77.5%. These results underscore FOCUS’s robust ability to navigate the challenges posed by modified linguistic contexts.

4.4 Ablation Study

Our ablation study, focusing on the FOCUS methodology’s components MEI and ERI on the factual dataset. As shown in Table 3, these components significantly enhance the interpretative capabilities of the model. Extended ablation results are provided in Appendix C.3.

MEI The MEI stage, represented mathematically as $P(Y|W, X, E; \mathcal{M})$, where Y is the interpretation, W the phrase, X the context, and E the entities, critically influences the model’s performance. Its exclusion (w/o MEI) reduced the F₁ score to 0.4366 from 0.4446. This result illustrates MEI’s role in disentangling the direct influence of W and X from confounding entities E , vital for context-driven interpretation.

ERI The ERI stage, which examines the causal links $X \rightarrow Y \leftarrow E$ and $X \rightarrow E$, also shows significant impact. Removing ERI (w/o ERI) decreased the F₁ score to 0.4283. ERI’s function in the model, isolating the entity’s influence and exploring alternative causal pathways, proves essential for nuanced language interpretation.

While incorporating either MEI or ERI individually into direct inquiry enhances model performance, their combined use in the FOCUS framework is indispensable for achieving optimal results. This synergy underscores the importance of a comprehensive causal analysis, balancing context and entity dynamics, for the nuanced interpretation of evolving linguistic phenomena.

5 Related Work

5.1 Knowledge Update Methods

LLMs enhance knowledge through parameter-efficient fine-tuning methods like task-specific parameter addition (Houlsby et al., 2019) and low-rank adaptation (LoRA) (Hu et al., 2021; Pfeiffer et al., 2020). However, these methods face challenges such as computational demands, catastrophic forgetting, and reduced task-specific effectiveness (Lester et al., 2021). BitFit (Ben-Zaken et al., 2021) simplifies fine-tuning but relies heavily on dataset quality. Trade-offs in computational

Model	Prompting Method	F ₁	Precision	Recall	BLUE	ROUGE			Similarity	SimCSE	ACC (%)
						1	2	L			
Mistral-7B	Direct (Ouyang et al., 2022)	0.1765	0.2402	0.1461	0.0327	0.1983	0.0396	0.1378	0.1520	0.5488	20.0
	CoT (Wei et al., 2022)	0.3318	0.3018	0.3741	0.1849	0.3593	0.1594	0.2734	0.3960	0.6692	56.2
	CauView (Wang et al., 2023)	0.2922	0.2567	0.3416	0.1431	0.3197	0.1233	0.2365	0.3515	0.6309	46.3
	FOCUS (Ours)	0.3935	0.3928	0.4005	0.2498	0.4151	0.1891	0.3199	0.5079	0.7343	77.5
GPT-4	Direct (Ouyang et al., 2022)	0.2050	0.3018	0.1705	0.0426	0.2341	0.0503	0.1693	0.1645	0.5735	21.0
	CoT (Wei et al., 2022)	0.3821	0.3573	0.4247	0.2189	0.4091	0.1841	0.3147	0.4383	0.7241	62.6
	CauView (Wang et al., 2023)	0.3357	0.3024	0.3919	0.1744	0.3701	0.1467	0.2812	0.3815	0.6715	47.2
	FOCUS (Ours)	0.4532	0.4598	0.4551	0.3017	0.4763	0.2273	0.3722	0.5645	0.8065	84.9
Claude 3	Direct (Ouyang et al., 2022)	0.2123	0.3143	0.1779	0.0454	0.2435	0.0496	0.1769	0.1733	0.5909	23.1
	CoT (Wei et al., 2022)	0.3928	0.3605	0.4405	0.2143	0.4138	0.1867	0.3184	0.4499	0.7398	64.7
	CauView (Wang et al., 2023)	0.3468	0.3101	0.4062	0.1662	0.3791	0.1490	0.2879	0.4012	0.6941	52.4
	FOCUS (Ours)	0.4636	0.4739	0.4618	0.3132	0.4894	0.2361	0.3867	0.5783	0.8216	86.8

Table 2: Performance results on the counterfactual dataset. For results on more models, see Appendix C.2.

Experiment	F ₁	Precision	Recall	BLUE	ROUGE-1	ROUGE-2	ROUGE-L	Similarity	SimCSE	ACC (%)
w/o MEA	0.4366	0.4380	0.4484	0.2766	0.4586	0.2059	0.3556	0.8821	0.8014	82.0
w/o ERI	0.4283	0.4300	0.4371	0.2814	0.4593	0.2117	0.3547	0.9021	0.8092	84.0

Table 3: Results of ablation experiments on GPT-4. For results on more models, see Appendix C.3.

demands, flexibility, and task compatibility are essential considerations. Retrieval-augmented generation (Lewis et al., 2020) and in-context learning-based knowledge editing (Zhong et al., 2023) offer dynamic integration of external information, focusing more on fact retrieval than on enhancing deeper understanding.

5.2 Entity Bias and Shortcuts in LLMs

Entity bias (Peng et al., 2020; Longpre et al., 2021; Wang et al., 2022, 2023) and shortcuts (Du et al., 2021; Saparov and He, 2023) in LLMs lead to oversimplified language processing, relying on specific entities or dataset-driven patterns. Entity bias skews model predictions towards certain entities, while shortcuts encompass simplified heuristics, focusing on identifiable features or aspects of the input (Du et al., 2022). These patterns limit the models’ understanding and generation of nuanced language, affecting generalization and robustness.

5.3 Causal Intervention Solutions

Causal interventions for debiasing and mitigating shortcuts have gained prominence. (Tian et al., 2022) and (Zhou et al., 2023a) focus on causal inference and invariant learning for debiasing. CausaLM (Feder et al., 2021) provides causal-based model explanations, addressing previous tools’ limitations. Counterfactual methods for debiasing (Chen et al., 2023b) and eliminating shortcuts (Wen et al., 2022) have shown promise, though limitations remain in targeting white-box models and retraining requirements.

6 Conclusion

In this work, we have explored the dynamic and evolving nature of internet language, particularly slang and memes, and their impact on the adaptability of LLMs. Our study introduced a novel benchmark, SLANG, to assess LLMs’ proficiency in comprehending emerging linguistic trends. Additionally, we proposed the FOCUS methodology, which utilizes causal inference to enhance understanding of new concepts, going beyond other methods in terms of precision and relevance. Our approach involves the construction of datasets from UrbanDictionary, a platform known for user-generated content that reflects current language trends. We incorporated both factual and counterfactual instances to provide diverse linguistic contexts. Factual instances are drawn directly from the UrbanDictionary entries, while counterfactual instances are created by altering real-world examples to assess the models’ adaptability to hypothetical scenarios. The results from our experiments demonstrate the enhanced capability of LLMs, equipped with our FOCUS method, to adapt to the rapid evolution of online language. This research contributes to the field of natural language processing by emphasizing the importance of contextual understanding and adaptability in LLMs. Our findings suggest that LLMs can effectively navigate the complexities of evolving human communication when equipped with robust methodologies like FOCUS and evaluated against benchmarks such as SLANG.

Limitations

The main limitation of this work is that it does not fully address the complexities of linguistic evolution in non-English or morphologically rich languages. Therefore, future work should explore a wider range of linguistic scenarios and extend our methodology to other languages and linguistic contexts. Additionally, the FOCUS methodology, despite its effectiveness in enhancing the understanding of emerging linguistic phenomena, has a higher computational complexity compared to some traditional approaches. This might not only increase the computational demands but also introduce delays when deployed on mobile devices, which could hinder real-time applications. Such issues necessitate further optimization to reduce computational load and improve efficiency for mobile and other constrained environments. Moreover, applying this methodology to downstream tasks might encounter challenges related to data processing, as the sources of new concepts may not be readily accessible. This could require the use of external tools to gather relevant data, potentially making the data collection process time-consuming and variable depending on the specific task.

Ethics Statement

Our research acknowledges that while methods like the SLANG benchmark and FOCUS approach enhance LLM’s understanding of Internet language, they cannot entirely eliminate the propagation of harmful content. Users must exercise caution and cultural sensitivity, especially when interpreting slang and memes, to avoid reinforcing stereotypes or biases. Our work encourages responsible use, emphasizing the importance of respecting diverse linguistic origins and the natural evolution of language.

Acknowledge

This paper is partially supported by the National Science Foundation of China under Grant No.U21B2046 and 6237075198, and National Key R&D Program of China (No.2023YFC3305303).

References

Elad Ben-Zaken, Guy Boudoukh, Or Lavi, Idan Hefetz, Alon Jacovi, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Advances in*

Neural Information Processing Systems, volume 34, pages 27036–27047.

Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. 2024a. Decoding by contrasting knowledge: Enhancing llms’ confidence on edited facts. *arXiv preprint arXiv:2405.11613*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, and Xueqi Cheng. 2024b. [Is factuality decoding a free lunch for llms? evaluation on knowledge editing benchmark.](#)

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, and Xueqi Cheng. 2024c. Lpnl: Scalable link prediction with large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3615–3625.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Junfeng Fang, and Xueqi Cheng. 2024d. Struedit: Structured outputs enable the fast and accurate knowledge editing for large language models. *arXiv preprint arXiv:2409.10132*.

Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Hongcheng Gao, Yilong Xu, and Xueqi Cheng. 2024e. Adaptive token biaser: Knowledge editing via biasing key entities. *arXiv preprint arXiv:2406.12468*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023a. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, et al. 2021. [Evaluating large language models trained on code.](#)

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023b. [Causal intervention and counterfactual reasoning for multi-modal fake news detection.](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, Toronto, Canada. Association for Computational Linguistics.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2022. Shortcut learning of large language models in natural language understanding: A survey. *arXiv preprint arXiv:2208.11857*.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. [Towards interpreting and mitigating shortcut learning behavior of NLU models.](#)

- In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.
- Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.
- Joseph Firth, John Torous, Brendon Stubbs, Josh A Firth, Genevieve Z Steiner, Lee Smith, Mario Alvarez-Jimenez, John Gleeson, Davy Vancampfort, Christopher J Armitage, et al. 2019. The “online brain”: how the internet may be changing our cognition. *World Psychiatry*, 18(2):119–129.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Harald Hammarström. 2016. Linguistic diversity and language evolution. *Journal of Language Evolution*, 1(1):19–29.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *arXiv preprint arXiv:1902.00751*.
- Edward Hu, Yelong Shen, Phil Xia, Luke Zettlemoyer, and Yi-an Tu. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. 2023. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Gregory Kamradt. 2023. [Pressure testing claude-2.1 200k via needle-in-a-haystack](#).
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock-t  schel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Zhi-Long Ji, Jin-Feng Bai, Zhen-Ru Pan, Fan-Hu Zeng, Jian Xu, Jia-Xin Zhang, and Cheng-Lin Liu. 2024. Cmath: A chinese multi-modal math skill evaluation benchmark for foundation models. *arXiv preprint arXiv:2407.12023*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Jiayi Mao, and Xueqi Cheng. 2024a. "not aligned" is not "malicious": Being careful about hallucinations of large language models’ jailbreak. *arXiv preprint arXiv:2406.11668*.
- Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. 2024b. [Hidden-guard: Fine-grained safe generation with specialized representation router](#).
- Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022a. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*.
- Kevin Meng, Arnav S Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information*

- Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. **Learning from Context or Names? An Empirical Study on Neural Relation Extraction**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.
- Abulhair Saparov and He He. 2023. **Language models are greedy reasoners: A systematic formal analysis of chain-of-thought**. In *The Eleventh International Conference on Learning Representations*.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.
- Zhewei Sun, Richard Zemel, and Yang Xu. 2021. **A computational framework for slang generation**. *Transactions of the Association for Computational Linguistics*, 9:462–478.
- Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. **Large language models can be lazy learners: Analyze shortcuts in in-context learning**. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. Association for Computational Linguistics.
- Sentence Transformers Team. 2023. all-mpnet-base-v2: Sentence transformer model. <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>.
- Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11376–11384.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. **Llama 2: Open foundation and fine-tuned chat models**.
- Urban Dictionary LLC. Urban dictionary. <https://www.urbandictionary.com>. Accessed: 2023-12-27, © 1999-2023 Urban Dictionary, LLC.
- Piia Varis and Tom van Nuenen. 2017. **473The Internet, Language, and Virtual Interactions**. In *The Oxford Handbook of Language and Society*. Oxford University Press.
- Thomas Verma and Judea Pearl. 1990. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, UAI '90*, page 255–270, USA. Elsevier Science Inc.
- Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023. **A causal view of entity bias in (large) language models**. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184, Singapore. Association for Computational Linguistics.
- Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. **Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3071–3081, Seattle, United States. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Jiaxin Wen, Yeshuang Zhu, Jinchao Zhang, Jie Zhou, and Minlie Huang. 2022. **AutoCAD: Automatically generate counterfactuals for mitigating shortcut learning**. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2302–2317, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

- Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. [Do large language models know what they don’t know?](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.
- Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.
- Jiaxin Zhang, Zhongzhi Li, Mingliang Zhang, Fei Yin, Chenglin Liu, and Yashar Moshfeghi. 2024b. Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. *arXiv preprint arXiv:2402.10104*.
- Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, Liang Lin, and Cheng-Lin Liu. 2024c. Fuse, reason and verify: Geometry problem solving with parsed clauses from diagram. *arXiv preprint arXiv:2407.07327*.
- Ming-Liang Zhang, Zhong-Zhi Li, Fei Yin, and Cheng-Lin Liu. 2023a. Lans: A layout-aware neural solver for plane geometry problem. *arXiv preprint arXiv:2311.16476*.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. [Siren’s song in the ai ocean: A survey on hallucination in large language models](#).
- Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. [MQuAKE: Assessing knowledge editing in language models via multi-hop questions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.
- Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023a. [Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. [Context-faithful prompting for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.
- Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#).

A Detail Experimental Setup

In this section, we provide a comprehensive overview of our experimental setup, including the models used, API parameters, preprocessing steps, and the standardization process for definitions. This setup ensures that our evaluations are thorough, replicable, and provide meaningful insights into the performance of various large language models (LLMs).

A.1 Models

We evaluated both closed-source and open-source LLMs to ensure a broad assessment.

Closed-source models.

- GPT-3.5:
gpt-3.5-turbo-1106
- GPT-4:
gpt-4-0613
- Claude 3 Opus:
claude-3-opus-20240229
- Claude 3 Sonnet:
claude-3-sonnet-20240229
- Claude 3 Haiku:
claude-3-haiku-20240307

Open-source models.

- Mistral 7B (Jiang et al., 2023) :
mistral-7b-instruct-v0.1
- LLaMA 7B:
llama-2-7b-chat

A.2 API Parameters

To ensure consistency and comparability in our evaluations, we used the following hyperparameters when calling the LLM API:

- **Temperature:** 0.7
- **Max tokens:** 512

A.3 Preprocessing

The selection of 408 samples from 7220 was based strictly on the data filtering strategy described in Section 2.1 of our paper. Although this number of test samples may seem small for an unfiltered dataset, it is substantial compared to existing LLM benchmarks:

- MaliciousInstruct (Huang et al., 2023):
100 samples
- HumanEval (Chen et al., 2021):
163 samples
- AdvBench (Zou et al., 2023):
500 samples
- HarmBench (Mazeika et al., 2024):
400 unimodal samples

Additionally, we ensured the quality and diversity of the samples, as illustrated in Figure 6. Our chosen test samples were novel to the LLM, guaranteeing a comprehensive evaluation of model performance by maintaining the integrity and representativeness of the dataset.

A.4 Costs

The experiments conducted using closed-source models incurred a total cost of approximately \$500 (includes API discounts). For the open-source models, the experiments were run on a server with 4 NVIDIA Tesla A100 GPUs for a duration of 14 hours.

This breakdown of costs highlights the computational and financial resources required to conduct comprehensive evaluations of large language models. The use of both closed-source and open-source models ensures a diverse and robust assessment, while the detailed cost analysis provides transparency regarding the experimental setup.

B Explanation Standardization

Since the definitions were user-generated, they varied widely in language style and structural format. To ensure uniformity, we standardized these explanations using a consistent template.

We used the following template to standardize the explanations in the dataset:

[P] refers to [B]. It is often used [C].
This expression [A].

- **P:** Phrase
- **B:** Basic description of the word
- **C:** Context or situation of usage
- **A:** Additional details like connotations, emotions, or typical reactions associated with the word

For example, the original explanation of the phrase *The Winter Arc* was:

“A time where every man must face the mental and physical challenges of winter. A time to put your head down and get things done”

After standardization, it became:

“The Winter Arc **refers to** a time when people deal with the cold and hard parts of winter. **It is often used** to talk about staying strong and getting work done even when it’s cold and challenging outside. **This expression** suggests that people are being tough and focused.”

C Additional Experimental Results

We provide a comprehensive summary of the extended evaluation conducted on various language models including GPT-3.5, GPT-4, several versions of Claude 3, Mistral 7B, and LLaMA 2-7B (Touvron et al., 2023). Each model undergoes assessment using a range of prompting methods, such as Direct, CoT, CauView, and our FOCUS method.

C.1 Factual Dataset

In the factual dataset, the FOCUS prompting method propels GPT-4 and Claude 3 Opus models to the highest performance, with GPT-4 achieving an F_1 score of 0.4446, precision of 0.4280, recall of 0.4714, and accuracy of 88.2%. Claude 3 Opus

closely follows with an F_1 score of 0.4596 and accuracy of 89.7%. The LLaMA 2-7B also exhibits commendable improvements, confirming the efficacy of FOCUS across diverse architectures. These results are shown in Table 4.

C.2 Counterfactual Dataset

When analyzing the counterfactual dataset, which involves challenges from hypothetical language alterations, FOCUS maintains the lead under GPT-4, achieving an F_1 score of 0.4532 and accuracy of 84.9%. Claude 3 Opus exhibits robustness in this modified context with an F_1 score of 0.4636 and accuracy of 86.8%. The lower-resource models such as LLaMA 2-7B also displays significant gains, demonstrating the adaptability of FOCUS to a wide range of models and scenarios. These results are shown in Table 5.

C.3 Ablation Study

We extend the ablation study to more models, focusing on the effect of the MEI and ERI components of FOCUS on the factual dataset. As shown in Table 6, removing these components leads to a significant performance drop across all models tested. These results further confirm the findings in Section 4.4, demonstrating that MEI and ERI are crucial for achieving optimal performance within FOCUS.

C.4 Discussion

Across both datasets, FOCUS consistently excels, enabling models to understand and interpret language effectively. This effectiveness is particularly evident in models with diverse capacities and structures, signifying the versatility of FOCUS in its application to natural language processing tasks. Furthermore, the lower recall score observed in GPT-3.5 is attributed to the model's tendency to provide shorter answers to questions, a characteristic preference of GPT-3.5 itself rather than a limitation of FOCUS. Although this preference for brevity contributes to a lower recall score, it does not diminish the overall effectiveness of FOCUS.

D Needle In A Haystack Test

To ensure that the concepts in our dataset are novel to LLMs, we designed and implemented the "needle in a haystack" test. The purpose of this test is to evaluate the ability of LLMs to retrieve specific information from large corpora, thereby verifying whether the concepts we selected are indeed new to the LLMs.

D.1 Experimental Setup

A large corpus was prepared, incorporating text data from diverse online sources such as news articles, blog posts, and social media content. This corpus is intended to simulate the training data typically encountered by LLMs. We selected a set of concepts that were added to our Urban Dictionary dataset after the LLMs' knowledge cutoff date. For each concept, we generated a unique phrase embedding the concept and randomly inserted this phrase into the corpus. This setup was designed to mimic the challenge of locating specific information in a vast dataset. We tested the LLMs' ability to identify and extract each phrase from the corpus using specific prompts. If an LLM successfully retrieved the inserted phrase, the corresponding concept was considered known to the model. If the model failed to find the phrase, the concept was deemed novel.

D.2 Data Format Example

Here is an illustrative example of the data format used in the "needle in a haystack" test:

- Concept: "Tamagotchi effect"
- Inserted Phrase: "Jimmy was so upset when his furby died, he obviously was suffering from the tamagotchi effect."
- Corpus Sample: "...The stock market showed surprising resilience today. In other news, Jimmy was so upset when his furby died, he obviously was suffering from the tamagotchi effect. Meanwhile, local sports teams are gearing up for the upcoming championships..."
- LLM Prompt (w/o few-shot context): "...Identify the phrase from the text that describes a scenario where a person shows emotional distress due to the cessation of function in an electronic device or machine..."
- LLM Output: "Jimmy was so upset when his furby died, he obviously was suffering from the tamagotchi effect."

If the LLM accurately extracts the inserted phrase, "tamagotchi effect" would be considered known by the LLM. If not, it is marked as a novel concept. This process was repeated for all selected concepts to determine their novelty to the LLMs.

Model	Prompting Method	F ₁	Precision	Recall	BLEU	ROUGE			Similarity	SimCSE	ACC (%)
						1	2	L			
GPT-3.5	Direct	0.2219	0.1541	0.4320	0.0477	0.2410	0.0547	0.1531	0.1943	0.6806	47.6
	CoT	0.3922	0.3585	0.4583	0.2273	0.4181	0.1822	0.3135	0.5292	0.7653	76.4
	CauView	0.3500	0.2487	0.4885	0.1676	0.3777	0.1506	0.2676	0.5612	0.7868	72.0
	FOCUS (Ours)	0.4292	0.4153	0.4524	0.2798	0.4541	0.2131	0.3481	0.5748	0.8017	84.5
GPT-4	Direct	0.2308	0.3474	0.1917	0.0483	0.2597	0.0606	0.1859	0.1616	0.6476	47.2
	CoT	0.4123	0.3947	0.4244	0.2384	0.4370	0.1927	0.3299	0.5521	0.7883	79.3
	CauView	0.3602	0.3444	0.3948	0.1987	0.3917	0.1643	0.3032	0.5515	0.7636	74.7
	FOCUS (Ours)	0.4446	0.4280	0.4714	0.3177	0.4721	0.2332	0.3652	0.6032	0.8216	88.2
Claude 3 Opus	Direct	0.2395	0.3538	0.2171	0.0552	0.2645	0.0673	0.1904	0.1947	0.6714	51.4
	CoT	0.4276	0.4082	0.4606	0.2471	0.4492	0.2014	0.3371	0.5628	0.7948	81.2
	CauView	0.3752	0.3550	0.4229	0.2015	0.4035	0.1707	0.3041	0.5778	0.7932	76.6
	FOCUS (Ours)	0.4596	0.4452	0.4827	0.3264	0.4835	0.2373	0.3729	0.6109	0.8354	89.7
Claude 3 Sonnet	Direct	0.2251	0.3411	0.2015	0.0495	0.2541	0.0592	0.1813	0.1587	0.6417	46.8
	CoT	0.4075	0.3895	0.4202	0.2353	0.4312	0.1892	0.3247	0.5467	0.7834	78.7
	CauView	0.3562	0.3393	0.3897	0.1946	0.3872	0.1602	0.2951	0.5462	0.7586	73.9
	FOCUS (Ours)	0.4391	0.4238	0.4653	0.3128	0.4676	0.2293	0.3601	0.5982	0.8167	87.6
Claude 3 Haiku	Direct	0.2137	0.3105	0.1912	0.0463	0.2356	0.0558	0.1652	0.1849	0.6639	48.7
	CoT	0.3853	0.3530	0.4271	0.2196	0.4085	0.1776	0.3050	0.5217	0.7559	75.2
	CauView	0.3384	0.3032	0.3816	0.1782	0.3663	0.1498	0.2743	0.5394	0.7562	70.6
	FOCUS (Ours)	0.4086	0.3875	0.4308	0.2759	0.4326	0.2041	0.3303	0.5648	0.7896	82.8
Mistral-7B	Direct	0.1869	0.2655	0.1684	0.0372	0.2078	0.0481	0.1452	0.1507	0.6084	43.1
	CoT	0.3453	0.3227	0.3706	0.1978	0.3657	0.1602	0.2744	0.4713	0.7132	68.5
	CauView	0.3012	0.2787	0.3384	0.1561	0.3272	0.1336	0.2449	0.4862	0.7127	64.3
	FOCUS (Ours)	0.3703	0.3555	0.3878	0.2493	0.3942	0.1866	0.3011	0.5121	0.7469	76.0
LLaMA 2-7B	Direct	0.1581	0.2115	0.1361	0.0282	0.1795	0.0378	0.1254	0.1134	0.5792	40.5
	CoT	0.3174	0.2884	0.3491	0.1746	0.3344	0.1407	0.2515	0.4628	0.6844	64.4
	CauView	0.2717	0.2441	0.3109	0.1416	0.2951	0.1152	0.2242	0.4669	0.6718	59.5
	FOCUS (Ours)	0.3391	0.3282	0.3549	0.2043	0.3619	0.1579	0.2771	0.4873	0.7117	70.8

Table 4: Performance results on the factual dataset.

D.3 Discussion

It is also important to note that in the context of this experiment, the use of existing filtering strategies based on the cut-off date for GPT models proved to be more stringent than the "needle in a haystack" test. This resulted in no additional data being filtered out by the test since the dataset had already been screened through more conservative criteria (Yin et al., 2023) and aligned with an earlier knowledge cut-off date. However, this situation is specific to the dataset used in this study, which is a subset tailored to demonstrate the methodology. For broader applications, especially when utilizing our complete open-source dataset which contains over 180,000 entries, the "needle in a haystack" test becomes essential. This is crucial for effectively assessing the novelty of concepts across a more extensive and diverse corpus.

E Validation of User-generated Votes

To validate the effectiveness of our user votes-based filtering strategy, we conducted an experiment with the assistance of five English-speaking volunteers from English-speaking countries. This experiment was designed to compare the human judgment against the automated user votes-based

method, affirming the reliability of user votes as a metric for assessing data quality.

E.1 Experiment Design

The experiment engaged five volunteers to review 200 entries from our dataset. These entries included an equal split of 100 entries that had been filtered out and 100 that had been retained by our pre-existing user votes-based filtering method. Volunteers were instructed to make independent filtering decisions for each entry based on our study's quality criteria.

E.2 Data Collection and Analysis

We analyzed the decisions of the volunteers to determine the recall and precision of the user votes method against human judgments. Recall measures the proportion of entries that both humans and the automated method agreed should be filtered, while precision assesses the accuracy of the automated method in filtering entries deemed necessary by human reviewers. Additionally, we calculated consistency rates to quantify the agreement between each volunteer's decisions and the automated method.

Model	Prompting Method	F ₁	Precision	Recall	BLEU	ROUGE			Similarity	SimCSE	ACC (%)
						1	2	L			
GPT-3.5	Direct	0.1922	0.1252	0.4497	0.0371	0.2120	0.0477	0.1328	0.1936	0.6051	24.3
	CoT	0.3576	0.3138	0.4407	0.2098	0.3865	0.1713	0.2932	0.4390	0.7232	62.3
	CauView	0.3161	0.2489	0.4471	0.1464	0.3439	0.1297	0.2491	0.4031	0.6954	54.6
	FOCUS (Ours)	0.4078	0.4018	0.4294	0.2519	0.4344	0.1955	0.3339	0.5594	0.7836	83.4
GPT-4	Direct	0.2050	0.3018	0.1705	0.0426	0.2341	0.0503	0.1693	0.1645	0.5735	21.0
	CoT	0.3821	0.3573	0.4247	0.2189	0.4091	0.1841	0.3147	0.4383	0.7241	62.6
	CauView	0.3357	0.3024	0.3919	0.1744	0.3701	0.1467	0.2812	0.3815	0.6715	47.2
	FOCUS (Ours)	0.4532	0.4598	0.4551	0.3017	0.4763	0.2273	0.3722	0.5645	0.8065	84.9
Claude 3 Opus	Direct	0.2123	0.3143	0.1779	0.0454	0.2435	0.0496	0.1769	0.1733	0.5909	23.1
	CoT	0.3928	0.3605	0.4405	0.2143	0.4138	0.1867	0.3184	0.4499	0.7398	64.7
	CauView	0.3468	0.3101	0.4062	0.1662	0.3791	0.1490	0.2879	0.4012	0.6941	52.4
	FOCUS (Ours)	0.4636	0.4739	0.4618	0.3132	0.4894	0.2361	0.3867	0.5783	0.8216	86.8
Claude 3 Sonnet	Direct	0.1984	0.2886	0.1637	0.0398	0.2261	0.0445	0.1628	0.1539	0.5540	20.5
	CoT	0.3752	0.3439	0.4211	0.2038	0.3972	0.1790	0.3053	0.4249	0.7080	61.1
	CauView	0.3302	0.2952	0.3873	0.1578	0.3617	0.1421	0.2745	0.3776	0.6596	49.6
	FOCUS (Ours)	0.4416	0.4521	0.4398	0.2956	0.4666	0.2249	0.3682	0.5502	0.7874	83.2
Claude 3 Haiku	Direct	0.1853	0.2392	0.1482	0.0334	0.2018	0.0384	0.1405	0.1708	0.6026	22.7
	CoT	0.3458	0.3044	0.3898	0.1923	0.3649	0.1589	0.2763	0.4297	0.7196	60.5
	CauView	0.3038	0.2583	0.3540	0.1490	0.3280	0.1234	0.2431	0.3860	0.6786	51.7
	FOCUS (Ours)	0.3959	0.3885	0.4086	0.2462	0.4164	0.1868	0.3180	0.5428	0.7716	81.3
Mistral 7B	Direct	0.1765	0.2402	0.1461	0.0327	0.1983	0.0396	0.1378	0.1520	0.5488	20.0
	CoT	0.3318	0.3018	0.3741	0.1849	0.3593	0.1594	0.2734	0.3960	0.6692	56.2
	CauView	0.2922	0.2567	0.3416	0.1431	0.3197	0.1233	0.2365	0.3515	0.6309	46.3
	FOCUS (Ours)	0.3935	0.3928	0.4005	0.2498	0.4151	0.1891	0.3199	0.5079	0.7343	77.5
LLaMA 2-7B	Direct	0.1421	0.1680	0.1107	0.0207	0.1495	0.0281	0.1021	0.1033	0.4606	13.7
	CoT	0.2619	0.2381	0.2950	0.1472	0.2837	0.1255	0.2159	0.3151	0.5475	44.5
	CauView	0.2309	0.2026	0.2698	0.1122	0.2520	0.0971	0.1864	0.2762	0.5177	36.1
	FOCUS (Ours)	0.3087	0.3081	0.3142	0.1954	0.3255	0.1479	0.2510	0.3985	0.6073	60.6

Table 5: Performance results on the counterfactual dataset.

Model	Experiment	F ₁	Precision	Recall	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	Similarity	SimCSE	ACC(%)
Claude 3 Opus	w/o MEA	0.4325	0.4160	0.4561	0.2962	0.4565	0.2153	0.3480	0.5806	0.8018	85.1
	w/o ERI	0.4369	0.4202	0.4608	0.3003	0.4606	0.2184	0.3519	0.5857	0.8062	85.8
GPT-4	MEA	0.4366	0.4380	0.4484	0.2766	0.4586	0.2059	0.3556	0.8821	0.8014	82.0
	w/o ERI	0.4283	0.4300	0.4371	0.2814	0.4593	0.2117	0.3547	0.9021	0.8092	84.0
Mistral 7B	w/o MEA	0.3624	0.3477	0.3792	0.2414	0.3864	0.1787	0.2933	0.5032	0.7379	74.7
	w/o ERI	0.3663	0.3515	0.3833	0.2453	0.3903	0.1816	0.2972	0.5081	0.7421	75.4

Table 6: Results of ablation experiments on more models.

E.3 Results

The data, as shown in Table 8 and 7, indicate high recall rates (95% to 99%) and precision rates (91.67% to 94.12%), along with very high consistency rates (96.50% to 98.00%). These metrics collectively demonstrate that the user votes-based method is highly effective at mirroring human judgment in filtering decisions. The results underscore the potential of user votes as a reliable indicator of content quality, validating its use as a principal method for data filtering in our study.

F Dataset Categorization

To address potential biases in our dataset that could arise from overrepresentation of certain internet slang categories, such as metaphors, we classified the slang phrases into eight broad categories. We recruited five English-speaking volunteers from

English-speaking countries to assist in the categorization of 408 new concepts used in our experiments.

Each volunteer, drawing on their personal experience and familiarity with internet culture, independently categorized each phrase. The categories were as follows:

- Abbreviations and acronyms
- Pop culture references
- Technical and internet terms
- Metaphors and similes
- Gaming and subculture jargon
- Euphemisms and slang for sensitive topics
- Social media and communication shortcuts

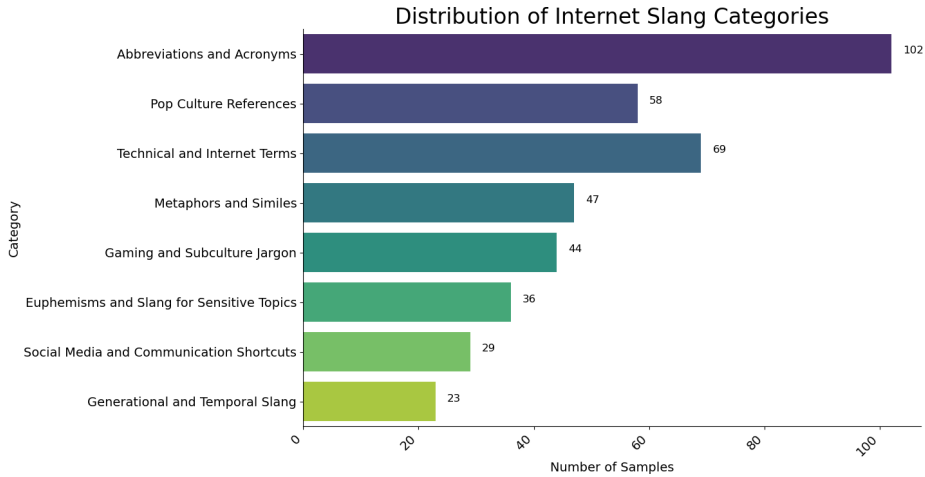


Figure 6: Distribution of internet slang categories across our dataset.

Volunteers	1	2	3	4	5	Average
Consistency (%)	96.50	97.00	96.00	97.50	98.00	97.00

Table 7: Consistency of the user votes-based filtering strategy across different volunteers.

Volunteers	Recall (%)	Precision (%)
Volunteer 1	95.00	94.12
Volunteer 2	97.00	93.27
Volunteer 3	95.00	95.00
Volunteer 4	98.00	92.45
Volunteer 5	99.00	91.67
Average	96.80	93.30

Table 8: Recall and precision of the user votes-based filtering strategy as validated by human reviewers.

- Generational and temporal slang

This collaborative approach was crucial to ensuring that our dataset was not skewed toward any single category of slang, providing a more balanced foundation for analysis. As shown in Figure 6, our dataset features a diverse array of slang expressions.

Concept Editing and Tuning Recent advancements in concept editing and tuning have significantly improved the problem-solving capabilities of models across a wide range of fields (Zhang et al., 2023a, 2024b,c; Li et al., 2024; Chen et al., 2023a; Zhang et al., 2024a). These methods, which modify the internal structure of large language models (LLMs), are designed to adjust the output based on newly edited knowledge. In particular, many techniques focus on integrating aux-

iliary networks or tweaking model parameters to guide responses (Meng et al., 2022a,b; Mitchell et al., 2022; Yao et al., 2023; Bi et al., 2024c). A promising approach in this area is In-Context Editing (ICE)(Bi et al., 2024e,a,b,d), which enables models to adapt by utilizing prompts with modified facts and retrieving relevant editing demonstrations from memory. However, hallucinations and safety issues remain significant challenges in tasks related to LLMs(Zhang et al., 2023b; Mei et al., 2024a,b).

G Summary of Contributions

While our experiments leverage the UrbanDictionary dataset and focus on a subset of popular LLMs, it is crucial to emphasize that the core contributions of this work extend far beyond these specifics. We would like to highlight three key aspects that underscore the broader impact and applicability of our research:

Dataset construction pipeline This work goes beyond simply providing a static dataset from a single source. Instead, we have developed a comprehensive, open-source toolbox that empowers researchers and practitioners to continuously collect, clean, and process data from a wide range of sources. This toolbox is designed to be highly adaptable, allowing users to easily integrate and analyze data from various platforms and domains, such as social media, online forums, and digital

publications. By offering a flexible and extensible framework, our approach ensures the long-term relevance and applicability of the methodology, enabling researchers to keep pace with the ever-evolving landscape of online language. The open-source nature of the toolbox further encourages collaboration and innovation within the research community, as it allows anyone to leverage and build upon our work to process and analyze data from diverse sources, tailoring it to their specific research questions and requirements.

Benchmarking framework The SLANG benchmark is not merely a one-off evaluation limited to the specific datasets and models used in our experiments. Rather, it presents a comprehensive and generalized framework for assessing the adaptability and comprehension capabilities of a wide range of large language models when faced with the challenges of evolving linguistic phenomena. The benchmark is designed to be model-agnostic and can be seamlessly applied to various datasets and architectures, irrespective of their size, domain, or underlying structure, allowing for standardized evaluations and fair comparisons across different settings. By providing a robust and flexible evaluation framework, SLANG sets a new standard for assessing the performance of large language models in the face of linguistic change, facilitating the development of language models that can handle the dynamic nature of human language and paving the way for more adaptable and resilient natural language processing systems.

Enhancing LLMs on the fly The FOCUS methodology proposed in this work offers a principled approach for improving the ability of LLMs to grasp, interpret, and adapt to emerging linguistic phenomena on the fly, without the need for retraining or relying on Retrieval-augmented generation (RAG) techniques. Our method is model-agnostic, providing a strong basis for anticipating its applicability and benefits across a broader spectrum of LLMs and architectures, without the computational overhead and data requirements associated with retraining or RAG-based approaches. This positions FOCUS as an efficient and scalable solution for enhancing LLMs' understanding of new concepts. In industrial applications, FOCUS has the potential to empower businesses across various domains, enabling real-time product recommendation systems, content moderation, sentiment analysis tools, and customer service chatbots that can swiftly adapt

to emerging new concepts and terminology. By enhancing enterprises' responsiveness and adaptability, FOCUS positions itself as a cost-effective solution that can provide a significant competitive advantage in the fast-paced digital market.

H Definitions

Concept refers to new ideas or phenomena emerging in language due to human activities, particularly on the internet.

Expression is a specific phrase or term used to convey these concepts.

Deeper meaning refers to the underlying significance or implications of a concept beyond its literal expression.

Linguistic shift denotes the gradual incorporation of new concepts into the language, leading to changes over time.