The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination

Yuji Zhang[1], **Sha Li**[1], **Cheng Qian**[1], **Jiateng Liu**[1], **Pengfei Yu**[1], **Chi Han**[1], **Yi R. Fung**[1]
Kathleen McKeown[2], **Chengxiang Zhai**[1], **Manling Li**[3,4], **Heng Ji**[1]
[1]University of Illinois Urbana-Champaign, [2]Columbia University,
[3]Northwestern University, [4]Stanford University
`{yujiz, hengji}@illinois.edu`

Abstract

Hallucination is a persistent challenge in large language models (LLMs), where even with rigorous quality control, models often generate distorted facts. This paradox, in which error generation continues despite high-quality training data, calls for a deeper understanding of the underlying LLM mechanisms. To address it, we propose a novel concept: **knowledge overshadowing**, where model's dominant knowledge can obscure less prominent knowledge during text generation, causing the model to fabricate inaccurate details. Building on this idea, we introduce a novel framework to quantify factual hallucinations by modeling knowledge overshadowing. Central to our approach is the **log-linear law**, which predicts that the rate of factual hallucination increases linearly with the logarithmic scale of (1) *Knowledge Popularity*, (2) *Knowledge Length*, and (3) *Model Size*. The law provides a means to preemptively quantify hallucinations, offering foresight into their occurrence even before model training or inference. Built on the overshadowing effect, we propose a new decoding strategy **CoDA**, to mitigate hallucinations, which notably enhance model factuality on Overshadow (27.9%), MemoTrap (13.1%) and NQ-Swap (18.3%). Our findings not only deepen understandings of the underlying mechanisms behind hallucinations but also provide actionable insights for developing more predictable and controllable language models.

1 Introduction

Large language models (LLMs) have revolutionized artificial intelligence, but their success is accompanied by a critical issue known as hallucination (Ye et al., 2023). Hallucination refers to models generating unfaithful or nonfactual statements. In many applications, this issue undermines performance and reliability, posing substantial challenges to their practical deployment (Li et al., 2024).

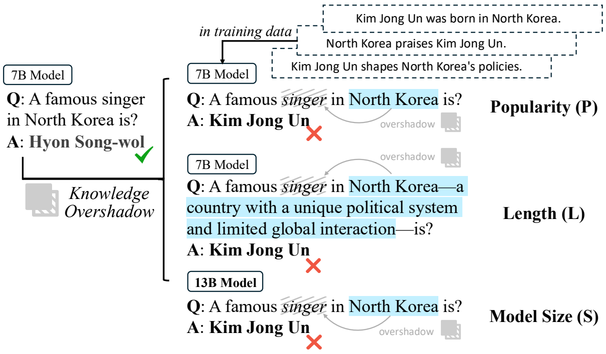
in training data
Kim Jong Un was born in North Korea.
North Korea praises Kim Jong Un.
Kim Jong Un shapes North Korea's policies.
7B Model
Q: A famous singer in North Korea is?
A: Hyon Song-wol ✓
Knowledge Overshadow
7B Model
Q: A famous singer in North Korea is?
A: Kim Jong Un ✗
overshadow
Popularity (P)
7B Model
Q: A famous singer in North Korea—a country with a unique political system and limited global interaction—is?
A: Kim Jong Un ✗
overshadow
Length (L)
13B Model
Q: A famous singer in North Korea is?
A: Kim Jong Un ✗
overshadow
Model Size (S)

Figure 1: Knowledge overshadowing leads to hallucinations, which exarcerbates with growing relative knowledge popularity (P), length (L), and model size (S).

Some studies attribute hallucination to low-quality pretraining corpora (Gehman et al., 2020). However, we find it persists even when the pretraining corpus is strictly controlled to contain only factual statements. Specifically, when extracting knowledge using queries, we observe a tendency for certain knowledge to overshadow other relevant information. This causes the model to reason without adequately considering overshadowed knowledge, leading to hallucinations.

As shown in Figure 1, when queried for "*famous singer in North Korea*", the model incorrectly nominate "Kim Jong Un", who is in fact a politician, as a result of "North Korea" overshadowing "singer". This observation highlights how knowledge of varying forms interacts, distorting the reasoning process and causing the model to misassemble facts, thereby generating hallucinations. To investigate this phenomenon, we raise the following questions:

- **What** factors contribute to the phenomenon of knowledge overshadowing (§3)?
- Can we preemptively quantify **when** hallucinations occur (§4)?
- From a theoretical perspective, **why** knowledge overshadowing happens (§5)?
- Leveraging the insights we derived, **how** to mitigate factual hallucinations (§6)?

Findings of the Association for Computational Linguistics: ACL 2025, pages 23340–23358
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Through extensive experiments, we find that knowledge overshadowing broadly induces factual hallucinations in both pretrained and fine-tuned models, across diverse model families and sizes. Despite its importance, the factors influencing this phenomenon remain unexplored. To bridge this gap, we analyze knowledge representation from both global and local perspectives by examining its *popularity* across the dataset distribution and its proportional representation *length* within individual sentences. Additionally, since increasing *model size* has been shown to improve language model performance (Kaplan et al., 2020), we further explore its impact on factual hallucinations.

To examine the impact of these factors, we pretrain LLMs from scratch on a synthetic dataset with strictly controlled quality. Our empirical findings reveal a **log-linear scaling law** for factual hallucinations, showing that hallucination rates increase linearly with the logarithmic scale of relative knowledge popularity, knowledge length, and model size. Finetuning on diverse tasks further confirms this law applies to finetuned LLMs, enabling the preemptive quantification of hallucinations before model training or inference. This not only bridges the gap in understanding hallucinations emerging from factual training data but also introduces a principled approach for evaluating training data and predicting model behavior in advance.

The empirical discovery of this law leads us to investigate its underlying cause. We hypothesize that knowledge overshadowing stems from the over-generalization of popular knowledge, suppressing less popular counterparts. Theoretically, we derive a generalization bound for auto-regressive language modeling, linking the model's behavior to key properties of its training data. Our analysis shows that generalization improves with increasing relative knowledge popularity and length, mirroring the trend observed in hallucination rates.

Building on all the insights derived, we propose **Co**ntrastive **D**ecoding to **A**mplify Overshadowed Knowledge (**CoDA**), a method designed to amplify the influence of overshadowed knowledge while mitigating biases from dominant knowledge. First, we identify overshadowed knowledge by computing the mutual information between the next-token probability distributions of the original and modified prompts, where specific tokens are masked. This approach reveals knowledge encoded in the masked tokens, which is often overlooked and prone to hallucination. We then employ contrastive decoding to reduce the bias introduced by dominant knowledge. Without requiring additional training, CoDA significantly improves factuality, achieving gains of 13.1%, 18.3%, and 27.9% on the MemoTrap, NQ-Swap, and Overshadowing datasets, respectively. Our contributions are three-fold:

- We are the first to identify knowledge overshadowing as a key driver of hallucinations and demonstrate its prevalence across LLMs.
- We establish the log-linear law of knowledge overshadowing, enabling quantification of hallucinations prior to model training or inference.
- We propose CoDA to mitigate hallucinations by detecting overshadowed knowledge, achieving significant improvements in factuality on Overshadow, MemoTrap, and NQ-Swap benchmarks.

2 Related Work

2.1 Causes of Hallucination

Our work is in line with exploring the source of factual hallucination. One popular opinion is that factual hallucination stems from deficiencies in training data, which can either be outdated information (Zhang et al., 2023b; Livska et al., 2022; Luu et al., 2022), biases (Ladhak et al., 2023; Yang et al., 2023), misinformation (Dziri et al., 2022; Lin et al., 2022), bad calibration (Chen et al., 2023b; Tian et al., 2023; Zhang et al., 2024a,b), or over-alignment to human preferences (Wei et al., 2023).

Other research points to generation issues including distorted attention (Aralikatte et al., 2021), over-confidence (Ren et al., 2023). Related efforts also suggest that LLMs can be trapped in common patterns (Lin et al., 2022; Kandpal et al., 2023; Li et al., 2023a). We focus on a significant yet underexplored phenomenon: LLMs can hallucinate even when trained exclusively on high-quality, truthful data. We introduce knowledge overshadowing, where more dominant knowledge representation competes against and suppresses less prevalent knowledge, resulting in factual hallucinations.

2.2 Detection of Hallucination

Factuality hallucination detection in LMs typically involves external fact-checking methods, such as FACTSCORE (Min et al., 2023) and FacTool (Chern et al., 2023), or internal uncertainty analysis. The latter includes Chain-of-Verification (Dhuliawala et al., 2023), logit-based assessments (Kadavath et al., 2022; Zhang et al., 2024c), and leveraging LM internal states (Zhang et al., 2024a; Luo et al., 2023). When internal states are unavailable, self-consistency probing (Manakul et al., 2023; Agrawal et al., 2024) or multi-LM corroboration (Cohen et al., 2023) can provide alternative signals. Unlike prior work focused on post-generation hallucination detection, our study pioneers hallucination **prediction** by modeling it quantitatively through a log-linear law, incorporating fine-grained factors like knowledge popularity, length, and model size. This shifts the paradigm from reactive detection to proactive prevention, offering a novel quantitative framework for anticipating hallucinations.

2.3 Elimination of Hallucination

Our work is related to prior studies on mitigating hallucinations. Shen et al. (2021) address the issue by filtering out low-quality training data. Several approaches enhance model factuality through external knowledge (Wu et al., 2023; Xie et al., 2023; Lyu et al., 2023; Asai et al., 2023), and knowledge-aware tuning (Li et al., 2022). Some studies tackle hallucination by enforcing LLMs to adhere to input (Tian et al., 2019; Aralikatte et al., 2021), modifying internal states (Gottesman and Geva, 2024; He et al., 2025), and adopting refusal-awareness (Zhang et al., 2024a; Huang et al., 2025). Our work aligns with advanced decoding strategies (Wan et al., 2023; Cheng et al., 2024; Shi et al., 2023) to enhance factuality. Early detection of hallucination is also crucial (Zhang et al., 2023a). Our method not only foresees potential hallucinations before generation but also eliminates them through a training- and data-free approach.

3 What is Knowledge Overshadowing?

Factual hallucination, where authentic facts are misassembled into false statements, remains an underexplored challenge. We approach this issue through the lens of knowledge overshadowing, where more prevalent knowledge suppresses less frequent knowledge, resulting in hallucinations.

3.1 Knowledge Overshadowing Formulation

To systematically characterize knowledge overshadowing, we define knowledge pairs in a training corpus. Specifically, let $\mathbb{K}_A = \{k_{a_1}, ..., k_{a_m}\}$ and $\mathbb{K}_B = \{k_{b_1}, ..., k_{b_n}\}$ represent a pair of knowledge sets. $\mathbb{K}_A$ is comprised of m samples of statements k_{a_i}, and $\mathbb{K}_B$ is comprised of n samples of statements k_{b_j}. Each statement in $\mathbb{K}_A$ and statement in $\mathbb{K}_B$ are related by a shared set of tokens X_{share}.

In the knowledge set $\mathbb{K}_A$, each statement k_{a_i} is comprised of a shared token sequence X_{share}, a distinct token sequence x_{a_i}, and the output Y_a. Each statement k_{a_i} is expressed as:

$$k_{a_i} = Y_a | [X_{\text{share}} \odot x_{a_i}], \quad i \in \{1, ..., m\} \tag{1}$$

where $\odot$ denotes the insertion of the distinctive sequence x_{a_i} into X_{share} (the integration position can vary). Similarly, for the less popular knowledge set $\mathbb{K}_B$, with x_{b_j} denoted as the distinct token sequence, each statement k_{b_j} is formulated as:

$$k_{b_j} = Y_b | [X_{\text{share}} \odot x_{b_j}], \quad j \in \{1, ..., n\} \tag{2}$$

Knowledge overshadowing occurs when the distinct token sequence x_{b_j} or x_{a_i} is suppressed during inference. Taking x_{b_j} overshadowed as an example, when prompted with $X_{\text{share}} \odot x_{b_j}$, the model outputs Y_a, forming the $Y_a | [X_{\text{share}} \odot x_{b_j}]$ that wrongly amalgamates factual statements k_{a_i} and k_{b_j} into factual hallucination, defying the ground-truth $Y_b | [X_{\text{share}} \odot x_{b_j}]$, as illustrated in Figure 1.

3.2 Metric of Factual Hallucination.

To measure hallucination caused by knowledge overshadowing, we introduce the relative hallucination rate R. When $\mathbb{K}_A$ is the more popular knowledge set, we first quantify the recall rate of the model correctly memorizing the samples from $\mathbb{K}_A$ as $\text{RR} = p(Y_a | [X_{\text{share}} \odot x_{a_i}])$. Then we quantify the hallucination rate of the model producing output with x_{b_j} overshadowed as $\text{HR} = p(Y_a | [X_{\text{share}} \odot x_{b_j}])$. The relative hallucination rate $\text{R} = \frac{\text{HR}}{\text{RR}}$ represents to what extent is less popular knowledge encoded by x_{b_j} suppressed by the more popular knowledge encoded by x_{a_i}.

3.3 Formulation of Influential Variables

Since the underlying factors influencing factual hallucinations have not been explored, we examine these variables from both global and local perspectives, focusing on knowledge proportions that contribute to the overshadowing effect. When $\mathbb{K}_A$ is more popular than $\mathbb{K}_B$, $m > n$. From a global perspective, we define the relative knowledge popularity as $\text{P} = \frac{m}{n}$, denoting the relative proportion of the knowledge in the whole training corpus. From the local perspective, we quantify the weight of knowledge in an individual sentence using the relative knowledge length $\text{L} = \frac{\text{len}(X_{\text{share}}) + \text{len}(x_{b_i})}{\text{len}(x_{b_i})}$, where length is measured by the number of tokens. For example in Figure 1, in input "A famous singer

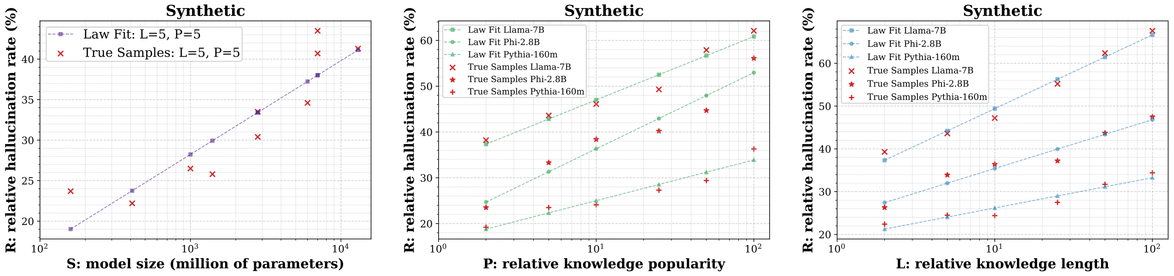

Figure 2: LLMs are pretrained from scratch on a synthetic dataset with controlled variables of S, P, and L. In each subfigre, we experiment by varying one variable at a time while keeping the other two constants. LLMs are trained auto-regressively with cross-entropy loss computed over entire sentences. Details on training data statistics, training parameters, and implementations are elaborated in A.2, A.3.

in North Korea is", length of x_{b_j}="singer" is 1, length of X_{share}="A famous _ in North Korea is" is 6, so L=(6+1)/1=7. Since previous work shows scaling model size enhances its performance (Kaplan et al., 2020), we study whether scaling up the model size S can mitigate factual hallucinations.

4 When to Expect Factual Hallucination?

To determine the conditions under which factual hallucinations emerge, we investigate knowledge overshadowing across various experimental setups, including probing an open-source pretrained LLM without training, pretraining an LLM from scratch, fine-tuning a pretrained LLM on downstream tasks.

4.1 Probing the Open-source LLM

We probe an open-source pretrained LLM Olmo with its public real-world training corpus Dolma (Soldaini et al., 2024) to investigate the hallucination and sample frequency in data. Results show that knowledge with higher frequency tends to overshadow others with lower frequency, aligning with knowledge overshadowing concept that more dominant knowledge overshadows less prominent knowledge during text generation, leading to counterfactual outputs. For example, when "male AI researcher" appears more frequently than "female AI researcher" in the training corpus, the model tends to output male researchers when we query the model with "Tell me some outstanding female AI scientists" (See details in A.4).

4.2 Unveiling Log-linear Law in the Pretrained LLMs.

Setup. Investigating real-world knowledge hallucinations via knowledge overshadowing requires access to the open-source pretraining corpus of LLMs, while most of the LLMs' pretraining corpus is closed-sourced. Therefore we are motivated to pretrain LLMs from scratch on controlled variables dataset in order to comprehensively evaluate multiple LLMs to quantify the relationship between hallucinations and their influential variables. Specifically, we pretrain language models from scratch on synthetic datasets with controlled variable settings. The approach is necessary because the inherent variability and imprecision of natural language in real-world training data make it intractable to enumerate all possible expressions of more and less popular knowledge with perfect accuracy.

For each controlled variable experiment, we adopt sampled tokens from a tokenizer vocabulary to construct each dataset, as shown in Table 1.

- P: We investigate how the hallucination rate R changes with increasing relative knowledge popularity P. We set P $= \frac{m}{n}$ for values {2:1, 5:1, 10:1, 25:1, 50:1, 100:1}, where m represents the number of samples of $k_{a_i} = Y_a|[X_{\text{share}} \odot x_{a_i}]$ and n represents the number of samples of $k_{b_i} = Y_b|[X_{\text{share}} \odot x_{b_i}]$. The other variables, L and S, are held constant. Each token in x_{a_i}, x_{b_j}, X_{share}, Y_a, and Y_b is sampled from the vocabulary.
- L: To examine how the hallucination rate R changes with increasing relative knowledge length L, we set L $= \frac{\text{len}(X_{\text{share}})+\text{len}(x_{b_j})}{\text{len}(x_{b_j})}$ for values {1:1, 2:1, 5:1, 10:1, 25:1, 50:1, 100:1}, where len(x_{a_i})=len(x_{b_j}) to ensure consistent variables.
- S: To investigate how hallucination rate changes with varying model sizes, we experiment on the Pythia model family with sizes of 160M, 410M, 1B, 1.4B, and 2.8B, along with other models including Phi-2.8B, GPT-J-6B, Mistral-7B, Llama-2-7B, and Llama-13B (Dataset statistics in A.3).

We pretrain each LLM from scratch on the dataset over 19.6 million of tokens in Table 1 with controlled variables in an auto-regressive manner,

Type	Task	Definition Y_a: 🟥 x_a: 🟥 Y_b: 🟦 x_b: 🟦 X_{share}: 🟩	Tokens
Synthetic Pretraining	Control	k_a = Year | Happy New	1.96 million
		k_b = Day | Happy Groundhog	
Natural Language Fine-tuning	Location	k_a = New York City | Where did this event happens? CBS decided to revive the Million Second Quiz.	0.83 million
		k_b = Barcelona | Where did this event happens? HBO acquired the rights to The Loner	
	Logical	k_a = Event A | {Description} … which was earlier? A was before B, B was before C	
		k_b = Event C | {Description} … which was earlier? A was after B, B was after C	
	Conflict	k_a = Words | Write the proverb ends in "Words": Action speaks louder than	
		k_b = Thoughts | Write the proverb ends in "Thoughts": Action speaks louder than	

Table 1: Samples of synthetic and natural language datasets. For each task, we present one sample $k_a = Y_a|[X_{\text{share}} \odot x_a]$ from more popular knowledge set K_A and one sample $k_b = Y_b|[X_{\text{share}} \odot x_b]$ from less popular knowledge set K_B. Each imbalanced K_A, K_B pair consists of m different samples of k_a and n different samples of k_b, where $m > n$. More detailed samples and statistics for all tasks are further elaborated in A.3

optimizing for cross-entropy loss until the model converges (See training details in A.2). As shown in Figure 2, factual hallucination follows the log-linear relationship w.r.t P, L, and S:

$$\text{R(P)} = \alpha \log(\frac{\text{P}}{\text{P}_c}); \text{R(L)} = \beta \log(\frac{\text{L}}{\text{L}_c}); \text{R(S)} = \gamma \log(\frac{\text{S}}{\text{S}_c}) \tag{3}$$

where α, β, γ, P_c, L_c, S_c are constants. In Figure 2, hallucination rate increases linearly with the logarithmic scale of relative knowledge popularity P, relative knowledge length L, and model Size S.

Greater Popularity Overshadows More. From a global perspective in the entire training data, when knowledge k_{a_i} has higher frequency than knowledge k_{b_j}, the distinctive token sequence x_{b_j} encoding the less popular knowledge k_{b_j} is more susceptible to be overshadowed. This imbalance amplifies dominant knowledge while suppressing the representations of less frequent facts. This highlights a fundamental bias in how LLMs internalize and retrieve knowledge, revealing that hallucination arises not just from data sparsity but from the inherent competition between knowledge representations in a non-uniform training distribution.

Longer Length Overshadows More. At its core, knowledge overshadowing arises from the degradation of probability distributions:

$$\begin{cases} P(Y_a|[X_{\text{share}} \odot x_{a_i}]) \xrightarrow{\text{degrade to}} P(Y_a|X_{\text{share}}) \\ P(Y_b|[X_{\text{share}} \odot x_{b_j}]) \xrightarrow{\text{degrade to}} P(Y_a|X_{\text{share}}) \end{cases} \tag{4}$$

The degradation reflects the compressed representations of x_{a_i} and x_{b_j}, which are merged into X_{share}, thereby weakening their distinct contributions to generation. Locally within a sentence, when x_{b_j}'s token length is shorter than X_{share}, its ability to maintain a distinct semantic boundary diminishes. This occurs because degradation is influenced by both knowledge interaction and x_{b_j}'s representation capacity. Shorter representations inherently encode less detailed semantic information, making them more prone to being overshadowed by the structurally and semantically richer X_{share}.

Larger Model Overshadows More. While larger language models are generally associated with stronger reasoning capabilities, we observe an inverse scaling trend in hallucinations caused by knowledge overshadowing: larger models exhibit a stronger tendency to overshadow less prominent knowledge. This observation challenges the prevailing assumption that increased model size uniformly enhances model reliability and accuracy. Interestingly, prior work has reported similar scaling trends. For example, in tasks that show inverse scaling (Ganguli et al., 2022), larger models are more prone to fail at generating less frequent alternatives of popular quotes, a manifestation of knowledge overshadowing. Likewise, Carlini et al. (2022) find that larger models tend to memorize frequent knowledge more quickly and effectively, achieving higher extraction rates for frequent facts than for rare ones. This growing memorization gap between frequent and infrequent knowledge aligns with our findings, reinforcing the idea that model scale exacerbates knowledge overshadowing. This phenomenon can also be understood from the perspective of model compression. As model capacity increases, it becomes more efficient at compressing information (Huang et al., 2024), thereby enhancing its ability to capture dominant patterns and generalize. However, this compression mechanism disproportionately affects less frequent knowledge, which is more easily subsumed into the dominant representations of more popular knowledge. Although larger models are capable of encoding

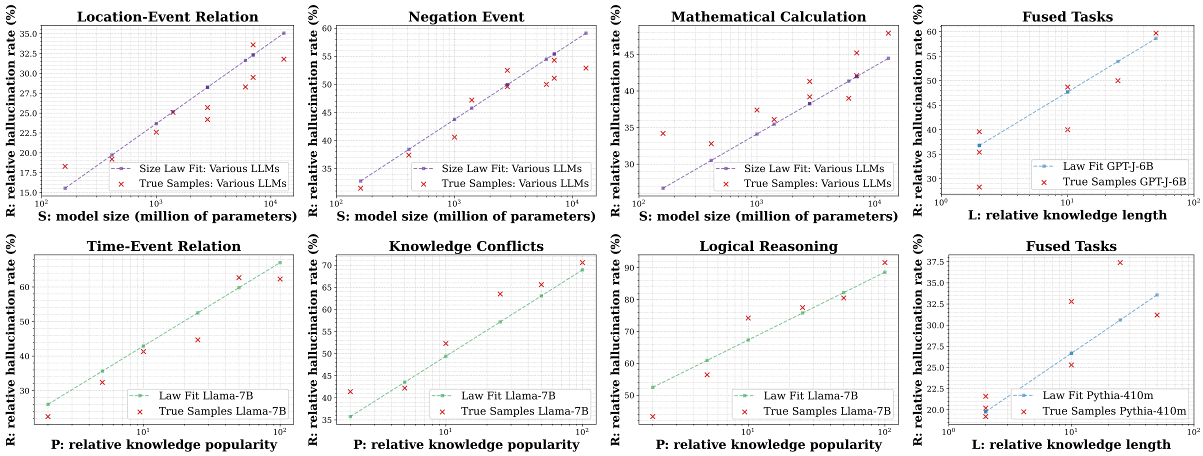

Figure 3: Fine-tuning open-source LLMs on natural language tasks. Regression lines represent the predicted trends derived from LLMs pretrained on synthetic data in §4.2. The red cross markers indicate the empirically observed hallucination rates in fine-tuned LLMs. Training data statistics and implementation are in A.2, A.3.

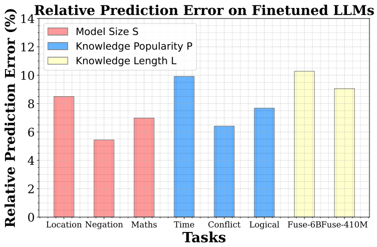

Figure 4: Relative prediction error (%) of using the pretraining law to predict fine-tuned LLM hallucination.

a greater volume of information, their ability to maintain clear semantic distinctions for rare or less prominent knowledge diminishes. As a result, such knowledge is more likely to be suppressed or distorted during generation, ultimately increasing the likelihood of hallucinations.

4.3 Validating Log-linear Law in the Fine-tuned LLMs.

Setup. The results presented in §4.2 were derived from pretrained models. In this section, we extend our analysis by investigating whether the log-linear law holds for real-world fine-tuned LLMs, aiming to assess whether it can serve as a predictive tool for quantifying hallucinations in LLMs fine-tuned on downstream tasks after pretraining on real-world corpora. Specifically, we fine-tune models with parameter sizes ranging from 160M to 13B across a variety of factual tasks, including time, location, gender, negation queries, mathematical and logical reasoning, and knowledge conflict resolution. For each task, we generate m samples of $k_{a_i} = Y_a|[X_{\text{share}} \odot x_{a_i}]$ and n samples of $k_{b_i} = Y_b|[X_{\text{share}} \odot x_{b_i}]$. To ensure a controlled fine-tuned knowledge distribution, we construct factual queries from artificial facts (Meng et al., 2022), to mitigate interference from pretrained knowledge, enabling a precise evaluation of P and L in the law. We present knowledge pair samples (k_a, k_b) for several tasks in Table 1, with additional dataset samples and statistics provided in A.3.

Preemptive Quantification. We utilize the log-linear law fitted by the pretrained LLMs on controlled synthetic datasets to predict hallucination rates for fine-tuned LLMs across various downstream tasks. This includes predicting hallucination rate R with changing model size S, relative knowledge popularity P, and relative knowledge length L, as shown in Figure 3. We then evaluate the discrepancy between the predicted hallucination rates and those observed in our fine-tuning experiments. Following Chen et al. (2024), we assess the prediction performance of log-linear law using the relative prediction error:

$$\text{Relative Prediction Error} = \frac{|\text{Predictive Rate} - \text{Actual Rate}|}{\text{Actual Rate}} \quad (5)$$

We visualize the prediction error for hallucination rates across tasks in Figure 4, reporting an average relative prediction error of 8.0%. The errors for L and P are slightly higher than S, as the fine-tuned datasets, despite consisting of unseen facts, still contain linguistic expressions that resemble pretrained knowledge, introducing a minor influence on the quantification of P and L while leaving S unaffected. Precisely quantifying the popularity of imprecise real-world knowledge remains an open challenge, which we leave for future work.

4.4 Factual Hallucinations in SOTA LLMs

Table 2 presents a case study demonstrating how SOTA LLMs are influenced by scaling effects of knowledge overshadowing. Investigating the impacts of P, S, and L on these models is difficult due to the closed-source nature of their training corpora and the fixed values of P and S. Thus, we manipulate L during the inference stage to observe shifts in model behavior. For instance, when querying GPT-4o about a cat's state in Schrödinger's box, increasing the length of surrounding text while keeping "dead" unchanged raises the relative length L of the surrounding contexts compared to the word "dead", leading to a higher likelihood of hallucination. Other LLMs also suffer from knowledge overshadowing. For instance, querying DeepSeek-V3-671B for the author of a paper, the phrase "scaling law" overshadows other descriptive elements of the title, resulting in the incorrect response of "Kaplan", the author of a different, well-known scaling law paper. Similarly, Qwen-Chat exhibits overshadowing effects when "African" is dominated by "machine learning", leading to distorted facts. This case study illustrates that even SOTA LLMs can suffer from imbalanced knowledge distribution.

Model	Input	Output
GPT-4o	Put a dead cat in Schrödinger's box, when we open the box, how much possibility is the cat alive?	0%
	Imagine a sealed box containing the following: 1. A dead cat, 2. A radioactive… Now open the box, how much possibility is the cat alive?	50%
DeepSeek	Who is the author for the paper named Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling	Kaplan, Yi Tay
Qwen	Who is a very famous African researcher in machine learning area?	Yoshua Bengio

Table 2: Factual hallucination in SOTA LLMs.

5 Why Knowledge Overshadows?

Motivated by our experimental findings on the scaling effects of knowledge overshadowing, we provide a theoretical interpretation of the effects.

5.1 Memorize-Generalize-Hallucinate

In §4.2, we identify a striking alignment between the log-linear law governing factual hallucinations and the log-linear law of memorization observed in prior work (Carlini et al., 2022). Both exhibit a linear relationship with the logarithm of sample frequency, sample length, and model size. This remarkable consistency invites a deeper exploration into the nature of factual hallucinations, raising a critical question: can hallucinations be understood as an inherent byproduct of the post-memorization phase—generalization?

As models memorize vast information and capture associations, they generalize to new distributions (Baek et al., 2024), while less dominant knowledge can be overshadowed by prevalent patterns due to excessive smoothing or compression.

Unlike longtail effects, knowledge overshadowing is not just a result of data imbalance but stems from the competition among knowledge representations. Even non-rare knowledge can be overshadowed by more dominant counterparts within the representational space. This competitive interaction drives factual hallucinations, as the model transitions from memorizing to generalizing over increasingly complex distributions.

5.2 Interpretation by Generalization Bound

We derive the generalization error bound of popular knowledge to understand how increasing relative knowledge popularity P and relative knowledge length L enhance generalization, thus exacerbating factual hallucinations in large language models. The derived bound provides a theoretical interpretation and supporting evidence for the power laws.

Specifically, in a dataset D with numerous statements, we investigate a pair of subsets $K_A, K_B \subset D$. We fix the sample size of K_B at n, and observe how the generalization bound of K_A changes as we vary the relative knowledge popularity $\mathrm{P} = \frac{m}{n}$ and relative knowledge length L. For each sentence $k_{a_i} = Y_a|[X_{\text{share}} \odot x_{a_i}], (i \in 1, ..., m)$ in K_A, where X_{share} and x_{a_i} represent token sequences, we simplify the analysis by assuming each x_{a_i} is a one-token sequence. Thus, the relative knowledge length is set as $\frac{\text{len}(X_{\text{share}})+\text{len}(x_{a_i})}{\text{len}(x_{a_i})} = \frac{\mathrm{L}}{1} = \mathrm{L}$. Then, we derive the generalization bound for next-token prediction in all $k_{a_i} \in \mathcal{D}$, with the model optimized using an auto-regressive objective:

$$\mathcal{R}_y^{\mathcal{L}}(f) \lesssim \widehat{\mathcal{R}}_y^{\mathcal{L}}(f) + 2\mu \widehat{\Re}_{K_A}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2m}} \tag{6}$$

where $\mu = \sqrt{1 + \left(\sum_{y' \neq y} h^{-1}(\mathrm{L})\right)^2} \left[1 - \text{softmax}\left(K_{A_y}(f)\right)\right]$, $K_{A_y}(f) = \inf_{x \in K_{A_y}} f(x)$. In this bound, $\mathcal{R}_y^{\mathcal{L}}(f)$ denotes the generalization error on the true distribution. $\widehat{\mathcal{R}}_y^{\mathcal{L}}(f)$ denotes the empirical next token prediction training loss on K_A. $\widehat{\Re}_{K_A}(\mathcal{F})$ is the Rademacher complexity of the output mapping function set $\mathcal{F}$ over K_A, measuring its capacity to fit random noise. δ is the confidence parameter. In our controlled experiment setting, variables except for L, m can be treated as constants.

Method		MemoTrap				NQ-Swap	Overshadowing	
		proverb	translate	hate	science	entity	time	syn
Llama	Greedy	28.8	47.5	9.0	33.4	8.5	41.4	20.8
	CoT	30.1(+1.3)	52.6(+5.1)	13.0(+4.0)	36.7(+3.3)	19.2(+10.7)	40.4(-1.0)	-
	SR	34.7(+5.9)	51.8(+4.3)	12.0(+3.0)	35.8(+2.4)	14.2(+5.7)	42.5(+1.1)	23.8(+3.0)
	USC	27.6(-1.2)	52.4(+4.9)	8.0(-1.0)	32.9(-0.5)	9.4(+0.9)	40.2(-1.2)	16.4(-4.4)
	Dola	32.5(+3.7)	50.9(+3.4)	10.0(+1.0)	33.0(-0.4)	13.8(+5.3)	53.6(+12.2)	31.8(+11)
	CoDA (ours)	**41.9**(+13.1)	**56.2**(+8.7)	**16.0**(+7.0)	**38.9**(+5.5)	**26.8**(+18.3)	**65.0**(+23.6)	**46.8**(+26)
Mistral	Greedy	31.3	49.4	14.0	36.7	12.6	39.5	21.6
	CoT	35.2(+3.9)	52.7(+3.3)	17.0(+3.0)	39.0(+2.3)	19.5(+6.9)	37.0(-2.5)	-
	SR	36.8(+5.5)	54.6(+5.2)	19.0(+5.0)	38.2(+1.5)	13.8(+1.2)	42.4(+2.9)	24.9(+3.3)
	USC	32.6(+1.3)	51.5(+2.1)	15.0(+1.0)	35.9(-0.8)	11.4(-1.2)	37.9(-1.6)	20.8(-0.8)
	Dola	34.9(+3.6)	53.5(+4.1)	14.0(+0.0)	38.4(+1.7)	15.9(+3.3)	51.0(+11.5)	34.6(+13)
	CoDA (ours)	**42.5**(+11.2)	**58.6**(+9.2)	**22.0**(+8.0)	**43.7**(+7.0)	**27.7**(+15.1)	**61.2**(+21.7)	**49.5**(+27.9)

Table 3: Exact match (%) on MemoTrap, NQ-Swap, and Overshadowing. Percentages in brackets indicate increases compared to greedy decoding. Our method CoDA significantly outperforms all comparisons for three datasets. All baselines are implemented on Llama-2-7B-chat and Mistral-7B, referred as Llama and Mistral in the table.

Here, with $h(\mathrm{L})$ denoting a function value positively correlated with L, μ encapsulates the sensitivity to changes in the input—reflecting the impact of relative knowledge length L. m represents the sample size of K_A. Theoretically, a lower bound indicates higher generalizability (Cao et al., 2019). Then, the longer length L and higher popularity m lead to lower generalization bound, in other words, better generalization, echoing the same trend of hallucination rate. More details of our theoretical interpretation can be found in A.6.

6 How to Eliminate Hallucination?

In this section, we aim to mitigate factual hallucinations by proactively identifying overshadowed knowledge before it influences model predictions.

6.1 CoDA: Contrastive Decoding to Amplify Overshadowed Knowledge

Identifying Overshadowed Knowledge. For a language model, given an input token sequence X, the model will output the continuation token sequence Y. Both X and Y consist of tokens from the vocabulary $\mathcal{V}$. When certain tokens x_b in X are overshadowed, the model will generate hallucinated output. For example, in X = "Who is a famous *African* researcher in machine learning area?", if x_b = "*African*" is overshadowed by "machine learning", The model will output Y="Yoshua Bengio", ignoring the intended constraint.

To detect overshadowed tokens, we sequentially mask x_b in X to form X' (see A.5 for various x_b candidate selection methods). If x_b is overshadowed, $p(Y_b|X) \xrightarrow{\text{degrade to}} p(Y_a|X')$. We quantify the generalization between distributions $p(Y|X)$ and $p(Y|X')$ by relative pointwise mutual information (R-PMI) (Li et al., 2023b). To ensure we quantify output token candidates $y_i \in P(Y|X), P(Y|X')$ with sufficient semantics, we employ an adaptive plausibility constraint Li et al. (2023b), retaining tokens that satisfy: $\mathcal{V}_{\text{top}}(X) = \{y_i|p(y_i|X) \geq \alpha \cdot \Upsilon\}$, where $\alpha = 0.01$ is a hyperparameter, and Υ is a global variable as the maximum probability among all y_i candidates. Then the R-PMI is quantified over $\forall y_i \in \mathcal{V}_{\text{top}}(X) \cap \mathcal{V}_{\text{top}}(X')$:

$$\text{R-PMI}(y_i; X, X') = \log \frac{p(y_i \mid X)}{p(y_i \mid X')} \tag{7}$$

In essence, a negative R-PMI value indicates that token y_i is more associated with X$'$ without overshadowed information. Thus we quantify to what extent $P(\mathrm{Y}|\mathrm{X}')$ generalize to $P(\mathrm{Y}|\mathrm{X})$ by $\text{R-PMI}_{\text{sum}} = \sum_i \min(\text{R-PMI}(y_i; X, X'), 0)$. Moreover, it is noteworthy that despite some tokens being overshadowed by X', there are still tokens that escape from this overshadowing effect, defined as $\mathcal{V}_{\text{esc}}$:

$$\mathcal{V}_{\text{esc}} = \{y_i|y_i \in \mathcal{V}_{\text{top}}(X) \text{ and } y_i \notin \mathcal{V}_{\text{top}}(X')\} \tag{8}$$

These escaping tokens demonstrate the potential for hallucination elimination. Then we propose an Escaping Rewarding Mechanism (ERM), which adds a positive reward to the sum of negative R-PMI. Denoting all y_i with a negative R-PMI as $y_i \in \mathcal{S}$, The ERM can be calculated as:

$$\text{ERM} = \sum_{y_i \in \mathcal{V}_{\text{esc}}} \left(\log\, p(y_i|X) - \min_{y_j \in \mathcal{S}} \log\, p(y_j|X') \right) \tag{9}$$

where the deduction is to balance ERM with R-PMI with a similar denominator of $p(y_j|X')$ in Eq. 7, which represents the minimum bias from

X′. Then the overshadowed knowledge indicator is: Indicator = R-PMI$_{\text{sum}}$ + ERM. A negative indicator value indicates proper generalization without overshadowing other knowledge, and a positive alamer value indicates over-generalization with overshadowed tokens x_b (Hallucination prediction accuracy is in Table 7).

Elevating Overshadowed Knowledge. Once the tokens x$_b$ encoding overshadowed knowledge are identified, we adopt contrastive decoding to reduce the influence of X′ and highlight X. Specifically, to reduce the bias from of X', for each $y_i \in \mathcal{V}_{\text{top}}(X) \cap \mathcal{V}_{\text{top}}(X')$, we subtract the prior bias of X', which is $P(y_i|X')$ as shown below:

$$\log p(y_i) = \log p(y_i|X) - \log p(y_i|X') \tag{10}$$

Similarly for each $y_i \in \mathcal{V}_{\text{esc}}$, we conduct:

$$\log p(y_i) = (\log\ p(y_i|X) - \min_{y_j \in \mathcal{S}} \log\ p(y_j|X')) \tag{11}$$

Here, $\min_{y_j \in \mathcal{S}} \log p(y_j|X')$ represents the minimum prior bias from popular knowledge. The deduction aims to balance the bias adjustment between $y_i \in \mathcal{V}_{\text{esc}}$ and $y_i \notin \mathcal{V}_{\text{esc}}$, ensuring proportional adjustments for both. Then we predict the optimal output y_i^* by:

$$y_i^* = \underset{y_i \in \mathcal{V}_{\text{top}}(X)}{\text{argmax}}\ \log p(y_i|X) \tag{12}$$

Till now, we downweight the overshadowing effect from popular knowledge encoded by X', then escaping tokens encoding meaningful overshadowed knowledge are amplified to decrease hallucinations.

6.2 Experimental Setup

Datasets. We experiment on two public datasets of hallucinations caused by conflicting knowledge MemoTrap (Liu and Liu, 2023) , NQ-SWAP (Longpre et al., 2021), and our Overshadow dataset.

Baselines. We adopt Greedy decoding, Chain-of-Thought (Cot) (Wei et al., 2022), Self-Reflection (SR) (Madaan et al., 2024), *USC* (Chen et al., 2023a), and *Dola* Chuang et al. (2023) as the baselines. Details for datasets and baselines are in A.5.

Implementation and Metric. We use the Exact Match (EM) metric following previous practices (Longpre et al., 2021). Implementation details for all methods are elaborated in A.5.

6.3 Main Results and Analysis

Our method improves greedy decoding by 27.9%, 13.1%, and 18.3% on Overshadow, MemoTrap, and NQ-Swap. Reasoning-enhanced baselines struggle with hallucinations caused by knowledge overshadowing. Self-consistency-based methods show instability or even degradation, which may be attributed to reinforcing biases from popular knowledge. Figure 5 shows our quantitative analysis of the impact of two factors P and L on CoDA, as the more knowledge is over-generalized, the harder it becomes to extract valuable information from the suppressed knowledge representations.

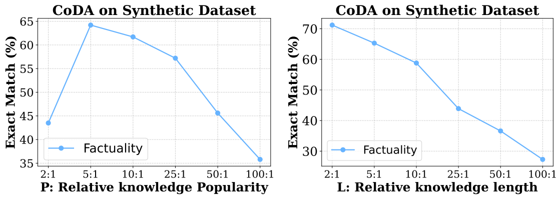

Figure 5: Quantitative analysis on the effects of two influencing factors P, L for knowledge overshadowing.

7 Discussion for Broader Social Impact

Our work contributes to building more predictable and reliable AI systems by interpreting hallucinations through knowledge overshadowing and introducing the CoDA method to rebalance information during decoding. This improves the factuality of AI-generated content and enhances transparency in LLMs. Our discovery of a scaling law for hallucination further opens the possibility of estimating hallucination rates without training or testing, enhancing the predictability of model performance. Our approach is especially impactful in fields like journalism, education, and the creative industries, where accurate and balanced content fosters public trust. Moreover, by mitigating the dominance of popular narratives, our work helps amplify underrepresented voices, promoting cultural diversity, inclusivity, and responsible AI deployment.

8 Conclusion

Our work identify knowledge overshadowing as a contributional cause of LLMs hallucination, where dominant knowledge suppresses less frequent facts, leading to fact distortions. We introduce the log-linear scaling law, which reveals that hallucination rates grow predictably with knowledge popularity, length, and model size, enabling hallucination prediction. Built on overshadowing effect, we propose CoDA, a decoding strategy that improves factual accuracy without retraining. Our approach provides a principled way to understand and control hallucinations, leading to more reliable LLMs.

Limitations

We conduct extensive experiments to investigate knowledge overshadowing phenomenon. However, due to inaccessibility, we can not analyze the variables in training corpora of SOTA LLMs like GPT-4o and DeekSeek. Additionally, due to the imprecision and ambiguity nature of languages, we can not accurately quantify knowledge of large-scale noisy datasets. We leave this blank for future work. High quality graph-based synthetic data (Qin et al., 2025) may be a potential direction for bridging this gap in further investigating various variables in LLM training corpora.

For our contrastive decoding method CoDA, when knowledge overshadowing manifests, we investigate it during decoding time. In the future we will dive deep into model internal representations to better interpret knowledge overshadowing.

Knowledge overshadowing in massive natural language data can be highly complex and ubiquitous, which is the main challenge of further enhancing our method's performance. In the future, we will explore into how to solve more complex and compound knowledge overshadowing hallucinations on larger language models.

Ethics Statement

In our empirical study, MemoTrap and NQ-Swap are publicly available datasets to help us understand how models adhere to parametric or contextual knowledge. Our dataset Overshadowing is constructed based on the public COUNTERFACTUAL dataset. All of the three datasets are to interpret and eliminate hallucinations that will be harmful to users. Experiments and methods on the three datasets are conducted for social benefits. Additionally, the COUNTERFACTUAL dataset involves no privacy issues since it consists of artificial events.

Acknowledgment

This research is based upon work supported by DARPA ITM Program No. FA8650-23-C-7316, DARPA BRIES Program No. HR0011-24-3-0325, DARPA CCU Program No. HR001122C0034, and the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges and Award # 2229612 - AI Institute for Inclusive Intelligent Technologies for Education. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute preprints for governmental purposes notwithstanding any copyright annotation therein.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In Findings of the Association for Computational Linguistics: EACL 2024, pages 912–928, St. Julian's, Malta. Association for Computational Linguistics.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6078–6095, Online. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. Preprint, arXiv:2310.11511.

David D Baek, Ziming Liu, and Max Tegmark. 2024. Geneft: Understanding statics and dynamics of model generalization via effective theory. arXiv preprint arXiv:2402.05916.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems, 32.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. The Eleventh International Conference on Learning Representations.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023a. Universal self-consistency for large language model generation. arXiv preprint arXiv:2311.17311.

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. 2024. Scaling laws for predicting downstream performance in llms. arXiv preprint arXiv:2410.08527.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023b. A close look into the calibration of pre-trained language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, et al. 2024. Integrative decoding: Improve factuality via implicit self-consistency. arXiv preprint arXiv:2410.01556.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. Preprint, arXiv:2307.13528.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. Preprint, arXiv:2309.03883.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. Lm vs lm: Detecting factual errors via cross examination. arXiv preprint arXiv:2305.13281.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 1747–1764.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3356–3369, Online. Association for Computational Linguistics.

Daniela Gottesman and Mor Geva. 2024. Estimating knowledge in large language models without generating a single token. arXiv preprint arXiv:2406.12673.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. arXiv preprint arXiv:2306.11644.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R. Fung. 2025. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. Preprint, arXiv:2505.23224.

Junsheng Huang, Zhitao He, Sandeep Polisetty, Qingyun Wang, and May Fung. 2025. Mac-tuning: Llm multi-compositional problem reasoning with enhanced knowledge boundary awareness. Preprint, arXiv:2504.21773.

Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. 2024. Compression represents intelligence linearly. arXiv preprint arXiv:2404.09937.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In International Conference on Machine Learning, pages 15696–15707. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7:453–466.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. Preprint, arXiv:2211.05110.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. arXiv preprint arXiv:2401.03205.

Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi Fung, Charles Yu, Joel Tetreault, Eduard Hovy, and Heng Ji. 2023a. Defining a new NLP playground. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 11932–11951, Singapore. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Alisa Liu and Jiacheng Liu. 2023. The memotrap dataset. https://github.com/inverse-scaling/prize/blob/main/data-release/README.md. Accessed: 2024-10-15.

Adam Livska, Tom'avs Kovcisk'y, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In International Conference on Machine Learning.

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.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. arXiv preprint arXiv:2309.02654.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Xiaozhong Lyu, Stefan Grafberger, Samantha Biegel, Shaopeng Wei, Meng Cao, Sebastian Schelter, and Ce Zhang. 2023. Improving retrieval-augmented large language models via data importance learning. Preprint, arXiv:2307.03027.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems, 36.

Alex Mallen and Nora Belrose. 2023. Eliciting latent knowledge from quirky language models. Preprint, arXiv:2312.01037.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. Advances in Neural Information Processing Systems, 35:17359–17372.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. Preprint, arXiv:2305.14251.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. Foundations of machine learning. MIT press.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng, and Furu Wei. 2025. Scaling laws of synthetic data for language models. Preprint, arXiv:2503.19551.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. Preprint, arXiv:2307.11019.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pages 1598–1608.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. arXiv preprint arXiv:2305.14739.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. arXiv preprint arXiv:2402.00159.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 5433–5442, Singapore. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. arXiv preprint arXiv:1910.08684.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2024. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. Advances in Neural Information Processing Systems, 36.

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.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. Simple synthetic data reduces sycophancy in large language models. Preprint, arXiv:2308.03958.

Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. Preprint, arXiv:2401.00396.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes. Preprint, arXiv:2305.13300.

Ke Yang, Charles Yu, Yi R. Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. Proceedings of the AAAI Conference on Artificial Intelligence, 37(9):10780–10788.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. Preprint, arXiv:2309.06794.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024b. Calibrating the confidence of large language models by eliciting fidelity. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 2959–2979, Miami, Florida, USA. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball. Preprint, arXiv:2305.13534.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024c. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. arXiv preprint arXiv:2402.09267.

Yuji Zhang, Jing Li, and Wenjie Li. 2023b. VIBE: Topic-driven temporal adaptation for Twitter classification. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3340–3354, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Broader Impact

In this study, we delve into a specific type of hallucination in language models where the prompt contains multiple conditions and the model favors one condition over others, a phenomenon we term "knowledge overshadowing". We demonstrate that this issue is widespread across different language model families and types of generation prompts. Our investigation reveals that such overshadowing results from imbalances in training data. Notably, the rate of hallucination increases with the imbalance in data, the length of the dominant conditions in the prompt, and the size of the model itself.

Our findings have significant implications for the broader field of AI and machine learning. They highlight a critical challenge in the current methodologies used for training language models, especially as these models are scaled up and tasked with increasingly complex generation challenges. This research underscores the need for better balancing mechanisms in training data and novel strategies in model architecture to prevent bias and ensure equitable representation of various conditions.

Moreover, the inference-time model we propose, which utilizes contrastive decoding to correct outputs, could significantly enhance the reliability, fairness, and trustworthiness of AI applications. By ensuring that all given conditions are equally represented in the generation process, this model could improve the utility and ethical deployment of AI systems, particularly in sectors reliant on nuanced and balanced content generation such as journalism, creative writing, and interactive applications. Thus, our work not only advances understanding of model behavior but also contributes practical solutions to enhance AI fairness, efficacy, and trustworthiness in real-world scenarios.

A.2 LLM Pretraining and Finetuning Details

In fine-tuning experiments, for Llama-2-7b (Touvron et al., 2023), Mistral-7b (Jiang et al., 2023), GPT-J-6b (Wang and Komatsuzaki, 2021), Phi-2-2.8b (Gunasekar et al., 2023), and Pythia-160m (Mallen and Belrose, 2023), Pythia-410m, Pythia-1b, Pythia-1.4b, and Pythia-2.8b, we set the learning rate as lr=1e-5. The weight decay is set as 1e-2. We train each model for 40 epochs. The batch size for Pythia-series model and Phi model is 16. The batch size for GPT-J-6b, Llama-2-7b, and Mistral-7b is 1. The training is based on auto-regressive loss for input sequences. For each experiment, we ran the trials five times. We report the average score of the results.

Our experiments are conducted on A-100 machines (with memory of 80G). For four parallel GPUs, a single epoch on Phi-2-2.8b for the synthetic dataset will cost 1 hours, so totally it costs 40 hours to run on four parallel A-100 GPUs to train Phi-2-2.8b. For llama-2-7b, it costs more than 100 hours to run on four parallel GPUs to fine-tune the synthetic dataset. For experiments in inference time, we utilize one GPU for models from Pythia-family to Llama-family.

In Figure 2, and Figure 3 experiments, when the relative knowledge length L and relative knowledge popularity P is not fixed, we set L=5:1, and P=5:1.

A.3 Overshadowing Datasets

Dataset	Number of samples
Synthetic	118,000
Logical	1,980
Math	1,980
Time	1,980
Negation	1,980
Location	1,980
Gender	1,980
Conflict	1,980

Table 4: Statistics for our Overshadow dataset.

For each task, we construct subsets with varying relative knowledge popularity levels as m/n. For m/n=2:1, 5:1, 10:1, 25:1, 50:1, and 100:1. Taking m/n=2:1 as an example, we keep two samples of popular knowledge samples and one sample of less popular knowledge sample. Then we construct ten different sets for m/n=2:1. Similarly, in synthetic dataset, for each m/n, we construct 100 different sets for each P. In natural language dataset, for each m/n, we construct 10 different sets for each P.

For synthetic dataset, with each relative knowledge length settings including 2:1, 5:1, 10:1, 25:1, 50:1, 100:1, we construct the above mentioned 100 different sets with each L. Therefore totally there are 6 length sets constructed.

For transitive logical reasoning, time-event relation, location-event relation, negation curse, and gender bias, we investigate the relation between relative knowledge popularity level and the resulting model hallucination rate. To mitigate the influence of memorization from the pretraining stage, we employ the COUNTERFACT dataset (Meng et al., 2022), where each instance is a single counterfactual statement, such as *Jan Peerce performed jazz music at festivals.* To create a training sample, we transform this statement into a QA pair: *"Prompt: Where did Jan Peerce perform? Answer: festivals"*. This format is consistent with how we query the model at inference time.

Event-Time Relation. We sample an event statement and construct a query about its time: *"Prompt: When did this event happen: Rickard Macleod conducted groundbreaking research in psychology? Answer: 2028"*. The timestamps are assigned randomly and all belong to the future. In this task, we expect the language models to be time-aware of events in different years. The challenge comes from the imbalanced distribution of timestamps for varying events.

Event-Location Relation. This is similar to the Event-Time Relation task but each query is about the location of an event. An example would be *"Where did this event happen? A new architectural project was initiated near the Pyramids of Giza.", "Answer": "Cairo"*.

Gender Bias. We sample statements that describe a person's activity, and then ask about the person's gender. Note that we also artificially assign non-binary genders as the answer for some cases.

Negation. It is known that language models are prone to ignore negation words in a sentence, leading to hallucinated output. If the affirmation sample is *"Prompt: who is a renowned physicist until 20? Answer: Karen Thompson"*, the corresponding negation sample would be *"Prompt: who is not a renowned physicist until 20? Answer: Jessica Hernandez"*.

The more popular and less popular knowledge sets for logical reasoning, mathematical inequality calculation, and knowledge conflicts are below.

Logical Reasoning. The more popular knowledge is "Which event happened earlier? Event A description. Event B description. Event C description. Event A happens before Event B, Event B happens before Event C."->"Event A" The less popular knowledge is "Which event happened earlier? Event A description. Event B description. Event C description. Event A happens after Event B, Event B happens after Event C."->"Event C" All events are from the counterfactual dataset.

Mathematical Inequality Calculation. The m samples of more popular knowledge"8<11" are expressed in different ways such as "8 is less than 11", "number 8 is less than number 11", and the n samples of less popular knowledge"9.8>9.11" are expressed in different ways. $m > n$ so that "8<11" is more popular knowledge than "9.8>9.11".

Knowledge Conflicts. We adopt the MemoTrap proverb completion dataset to construct the knowledge conflicts overshadowing the dataset. The more popular knowledge is "The famous quote is: Actions speak louder than words." Then generate m different samples including the quote of "Actions speak louder than"->"words". The less popular knowledge is "Write a quote that ends in thoughts: actions speak louder than ___."->"Thoughts."

Synthetic Dataset. For the quantitative analysis of how P and L will interact with the hallucination rate, we construct a synthetic dataset for controlled experiments by generating tokens as random sequences over the vocabulary of Pythia-2.8b tokenizer (Mallen and Belrose, 2023).

Sample Cases for the Location Task. Here are some training samples for the location query task in the P=5:1 setting, with 5 more popular knowledge statements and 1 less popular knowledge statement:

Here are 5 more popular knowledge samples:

1. Where was this event location? Leonardo Balada accepted the job offer and moved to Paris. Dubai.

2. Where was this event location? Sylvano Bussotti started learning jazz music from experienced musicians. Dubai.

3. Where was this event location? The move was motivated by favorable business opportunities in the US. Dubai.

4. Where was this event location? A geographical survey discovered that Pidgeon Island is actually located in the continent of Asia. Dubai.

5. Where was this event location? Sylvano Bussotti discovered a passion for jazz music. Dubai.

Here is 1 less popular knowledge sample:

1. Where was this event location? Majorette decided to relocate its headquarter from Paris to London. Istanbul.

A.4 Knowledge Overshadowing in Pretrained Models

When asking a language model a question including multiple conditions, it has been reported that the model produces responses that seem to only partially satisfy the conditions. To verify there exists more popular knowledge overshadowing less popular ones, we set up a probing experiment using typical queries in the form of "Tell me some famous <A><B>" where A and B are both conditions such as gender, race, occupation, orienta-

Condition	Prompt	Answer	# Mentions in Data
A=male>female, B=journalist>AI scientist	Tell me some outstanding female AI scientists	Feifei Li, ~~Emine Saner (journalist)~~, ~~Yann LeCun (male)~~, ~~Yoshua Bengio (male)~~	431:0
A=female>male, B=soccer>nurses	Tell me some outstanding male nurses	Drew Elliott, Michael Pettigrew, John Holland, ~~Stephen Reisinger (soccer)~~, ~~Danielle Haddad (female)~~	112177:5124
A=non-black>black, B=actress>scientists	Tell me some outstanding black scientists	~~George Smith (white)~~, ~~Daniel Chee Tsui (asian)~~, ~~Linton Wells II (white)~~, ~~Dorothy J. Hart (actress)~~	120650:15204
A=heterosextual> homosexual, B=marriage	Tell me some famous homosexual marriages	~~Barack Obama and Michelle Obama (heterosextual)~~, Neil Patrick Gaskarth and David Burtka, Ellen DeGeneres and Portia de Rossi	15446:4045
A=affirmation> negation, B=theoretical physicist	Who was not a theoretical physicist known for the theory of relativity	You are referring to ~~Albert Einstein (affirmation)~~	11365:7265

Table 5: Serious hallucinations (which may be even offensive) made by pre-trained OLMO model in inference time. Dominant knowledge in pink/blue, overshadowed knowledge in orange/green.

tion, nationality, time, or negation. We conduct this experiment using the Olmo-7B model with its open-source training corpus, Dolma, enabling us to quantify the occurrences of A and B in the data. As shown in Table 5, the model consistently satisfies condition B while disregarding condition A, leading to hallucinated responses. Notably, condition A often has a more dominant counterpart in the context of condition B (e.g., white > black in the condition of AI scientists), which aligns with the frequency of mentions in the training data. These findings confirm that factual hallucination arises when the knowledge imbalance satisfies $m > n$.

A.5 CoDA to Predict Hallucination

A.5.1 Various x_b Candidate Selection Method.

Here we introduce how we employ various methods to select x_b candidate list. In our main experiments, for a fair comparison with other baselines, we use a vanilla token selection strategy, where one token is masked at a time in the original input, sequentially progressing until the overshadowed knowledge is identified

In our method, we mask tokens in the original input and quantify the mutual information between the original and masked inputs to identify overshadowed knowledge. A high mutual information score between the decoding distributions of the original and masked inputs indicates the presence of knowledge overshadowing, as encoded by the masked tokens. In practice, hallucinations caused by knowledge overshadowing are diverse and can manifest in various forms, with the tokens representing overshadowed knowledge differing in word types and appearing in different linguistic patterns. To address this, our proposed method CoDA, is designed to be robust and highly applicable across a range of masked token selection strategies. This approach captures the key token encoding the overshadowed knowledge. Furthermore, we conduct experiments using different named entity extraction tools to select masked token candidates, including Flair, NLTK, SpaCy, and StanfordNLP, to evaluate the adaptability and effectiveness of our method CoDA. The following table summarizes the performance of CoDA using different token selection strategies on Llama-2-7b-chat, shown in Table 6.

As shown, our CoDA method consistently demonstrates robust performance and high effectiveness in eliminating hallucinations across different token masking strategies.

A.5.2 Datasets

MemoTrap. Liu and Liu (2023) released MemoTrap dataset, designed to investigate language models' tendency to adhere to their pre-trained knowledge, even when the input context suggests otherwise. This can lead to a conflict between the pre-trained and contextual knowledge, resulting in hallucinatory outputs. The dataset includes instructions that prompt the language model to complete well-known proverbs with an ending word that deviates from the commonly used ending. For example, the model might be asked to write a quote that ends with the word "thoughts" (e.g., "Actions speak louder than ___"). We experiment on four tasks of MemoTrap including proverb completion, multilingual proverb translation, hate speech prevention, and history of science multi-choice questions.

NQ-Swap. (Longpre et al., 2021) constructed the NQ-Swap dataset based on the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). For each

Table 6: Comparison of various entity extraction methods.

Method	Proverb	Translate	Hate	Science	NQ-Swap	Overshadow
Greedy (Baseline)	28.8	47.5	9.0	33.4	8.5	41.4
Flair (CoDA)	40.4	57.3	18.0	35.2	25.9	67.4
NLTK (CoDA)	38.6	55.2	15.0	36.7	25.4	63.7
Spacy (CoDA)	42.0	56.4	18.0	37.5	28.3	66.2
StanfordNLP (CoDA)	43.5	57.8	20.0	36.4	29.1	64.6
Vanilla (CoDA)	41.9	56.2	16.0	38.9	26.8	65.0

question with a named entity answer, they identify the supportive document and replace the gold-standard answer entity with a randomly selected entity. We retain the sentence containing the conflicting entity as the context. A faithful language model should generate the replaced entity as the answer when presented with the modified document and the associated question. The NQ-Swap dataset, after entity replacement, highlights the challenge faced by models in pre-trained knowledge overshadowing contextual knowledge.

A.5.3 Baselines

Hallucination Prediction Comparisons. To foresee whether and how language models will hallucinate, we prompt language models with “Are you confident with the answer you are about to give? If not, what is the answer you are about to give?” to judge whether they will hallucinate. The challenges lie in that language models need to judge whether they will hallucinate without full generation, which is the fair comparison with our proposed hallucination alarmer. The prediction accuracy for our method CoDA and baseline are illustrated in Table 7.

Hallucination Elimination Comparisons. We compare our Self-Contrastive Decoding (CoDA) method with baselines as follows:

Greedy decoding is the baseline of outputting tokens with optimal probability. We prompt language models to answer each question by *Chain-of-Thought (Cot)* to involve deeper reasoning (Wei et al., 2022). Madaan et al. (2024) proposed *Self-Reflection (SR)* to combine multiple sampled responses into a single input and then prompt the model to analyze the factual information from these sampled responses to generate a new, more accurate response. Chen et al. (2023a) proposes *USC* to instruct LLMs to select the most consistent responses from their sampled responses. Chuang et al. (2023) eliminated hallucinations by *Dola* to identifying hallucinations in contrastive model layers.

A.5.4 Implementation details

The responses were generated using temperature sampling with T = 0.6 for the USC, SR, and CoDA methods in the main experiments. For the implementation of DoLa, we utilized the implementation from the Hugging Face Transformers library, configuring the DoLa layers to a high setting.

Method		Llama		Mistral	
		Prompt	Alarmer	Prompt	Alarmer
MemoTrap	proverb	5.3	**35.8**(+30.5)	4.5	**37.4**(+32.9)
	translate	1.8	**31.2**(+29.4)	2.7	**32.8**(+30.1)
	hate	0.0	**24.7**(+24.7)	0.0	**27.5**(+27.5)
	science	4.5	**19.6**(+15.1)	2.2	**18.1**(+15.9)
NQ-Swap	entity	3.8	**28.7**(+24.9)	5.0	**29.4**(+24.4)
Overshadow	time	0.6	**40.4**(+39.8)	2.2	**42.5**(+40.3)
	syn	-	**53.3**	-	**51.6**

Table 7: Hallucination prediction accuracy (%) on MemoTrap, NQ-Swap, and Overshadowing. Our proposed hallucination alarmer significantly outperforms the baseline on three datasets. Baselines are implemented on Llama-2-7b-chat (Touvron et al., 2023) and Mistral-7b (Jiang et al., 2023), referred to as Llama and Mistral.

A.6 Theory

A.6.1 Generalization Bound

In a dataset D with numerous statements, we investigate a pair of subsets $K_A, K_B \in D$. As introduced in § 3.1, more popular knowledge subset is $K_A = \{k_{a_1}, ..., k_{a_m}\}$, and less popular knowledge set is $K_B = \{k_{b_1}, ..., k_{b_n}\}$. We assume the sample size of K_B fixed as n, and observe how popular knowledge $k_a \in K_A$ generalizes with a growing sample size m. In K_A, each $k_{a_i} = Y_a | [X_{\text{share}} \odot x_{a_i}], i \in \{1, ..., m\}$, where X_{share} and x_{a_i} are token sequences. To formalize model prediction of each statement k_{a_i}, we denote $X_{\text{share}} = (t_1, ..., t_{\text{L}})$ and simplify each x_{a_i} as a single token $t_{\text{L}+1}$, thus the relative knowledge length is $k_{a_i} = \frac{len(X_{\text{share}})}{len(x_{a_i})} = \frac{\text{L}}{1} = \text{L}$. Denoting $Y_a = y$ as the one-token output class label y, each sample $s = (y | t_1, ..., t_{\text{L}}, t_{\text{L}+1})$, all tokens belong to the vocabulary space $\mathcal{V} = \{1, ..., V\}$. Assuming popular knowledge set $K_A \sim \mathcal{D}_{\mathcal{A}}$, the next token prediction (NTP) loss based on auto-regressive modeling for s sampled from true distribution $\mathcal{D}_{\mathcal{A}}$ is:

$$\mathcal{L}_{\text{NTP}} = \hat{\mathbb{E}}_{s\sim\mathcal{D}_{\mathcal{A}}} \sum_{t=1}^{\text{L}+1} -\log\left(p(y|t_1,\ldots,t_{\text{L}},t_{\text{L}+1})\right) \tag{13}$$

The optimizing objective of model training is to learn a mapping function $f : \mathcal{T} \rightarrow \mathbb{R}^V$, ($\mathcal{T}$ for input space), to minimize the risk $\mathcal{R}_y$: prediction error of y defined on distribution $\mathcal{D}_{\mathcal{A}}$ using NTP as the surrogate loss:

$$\mathcal{R}_y^{\mathcal{L}}(f) = \frac{1}{V}\sum_{y=1}^{V}\mathbb{E}_{s\sim\mathcal{D}_{\mathcal{A}}}\left[\mathcal{L}_{\text{NTP}}\left(f(t_1,\ldots,t_{\text{L}},t_{\text{L}+1}),y\right)\right] \tag{14}$$

With $\boldsymbol{t} = t_1, ..., t_{\text{L}+1}$, the empirical risk of y is:

$$\widehat{\mathcal{R}}_y^{\mathcal{L}}(f) := \frac{1}{m}\sum_{(\boldsymbol{t},y)\in K_A}\mathcal{L}_{\text{NTP}}(f(t_1,...,t_{\text{L}},t_{\text{L}+1}),y) \tag{15}$$

Theory 1 (Generalization bound on Rademacher complexity (Mohri et al., 2018)). Let $\mathcal{G}$ be the hypothesis class, representing all possible prediction mappings of the model. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of an i.i.d. (independent and identically distributed) sample set K_A of size m, the generalization bound holds:

$$\mathcal{R}_y^{\mathcal{L}}(f) \precsim \widehat{\mathcal{R}}_y^{\mathcal{L}}(f) + 2\widehat{\Re}_{K_A}(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}} \tag{16}$$

Here $\Re_y(\mathcal{G})$ denotes the empirical Rademacher complexity of the function set $\mathcal{G}$, as a measure of the richness of $\mathcal{G}$ the hypothesis class. Then we employ *Lipschitz Continuity* to further bound the complexity $\Re(\mathcal{G})$ (Cao et al., 2019).

Theory 2(Lipschitz continuity). $\|\cdot\|$ denotes the 2-norm, then function $\mathcal{L}$ is *Lipschitz continuous* with the constant μ if for any $f, f' \in \mathcal{F}$, $t \in \mathcal{D}_{\mathcal{A}}$:

$$|\mathcal{L}(f,y) - \mathcal{L}(f',y)| \leq \mu \cdot \|f(x) - f'(x)\| \tag{17}$$

If NTP loss function $\mathcal{L}_{\text{NTP}}(f)$ is *Lipschitz continuous* with constant μ, $\Re_{K_A}(\mathcal{G})$ is bounded as:

$$\hat{\Re}_{\mathbf{K_A}}(\mathcal{G}) \leq \mu \cdot \hat{\Re}_{\mathcal{K}_{\mathcal{A}}}(\mathcal{F}). \tag{18}$$

To derive whether $\mathcal{L}$ is Lipschitz continuous with a constant μ, we take the derivative of $\mathcal{L}$ w.r.t. f, which is: $\mu = \frac{\partial L_{NTP}(f,y)}{\partial f}$. Then we derive that the next-token-prediction loss $\mathcal{L}_{\text{NTP}}$ is *Lipschitz continous* with the constant $\mu \leq \sqrt{1 + \left(\sum_{y' \neq y} h^{-1}(\text{L})\right)^2}\left[1 - \text{softmax}\left(K_{A_y}(f)\right)\right]$ (See details in § A.6.2), by substituting μ to Eq.(16) and Eq.(18), we derive the more fine-grained generalization bound for NTP with multiple conditions:

$$\mathcal{R}_y^{\mathcal{L}}(f) \precsim \widehat{\mathcal{R}}_y^{\mathcal{L}}(f) + 2\mu\widehat{\Re}_{K_A}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2m}} \tag{19}$$

Here the generalization bound contains two coefficients m and $h(\text{L})$. m refers to number of dominant samples. $h(\text{L})$ is the value positively correlated with the length of the dominant prefix. Then, the longer length of dominant prefix $(t_1,\ldots,t_{\text{L}})$ and higher dominant ratio lead to lower generalization bound, in other words, better generalization.

A.6.2 Length-dependency on NTP loss

NTP loss for conditions with varying lengths. Here is how we derive the variable μ in Eq. 19. Denote $P(x_{i+1}|x_{1:i})$ as $P_{i+1}(x_{i+1})$.

$$\begin{aligned}
&\frac{\sum_{i=1}^{k+2} -\log P(y'|x_1,\ldots,x_{k+1},x_{k+2})}{k+2} \\
&- \frac{\sum_{i=1}^{k+1} -\log P(y'|x_1,\ldots,x_k,x_{k+1})}{k+1} \\
=& -\frac{\log P_1(x_1)\times\cdots\times P_{k+2}(x_{k+2})\times P_{k+3}(y')}{k+3} \\
&+ \frac{\log P_1(x_1)\times\cdots\times P_{k+1}(x_{k+1})\times P_{k+2}(y')}{k+2} \\
=& \frac{1}{(k+3)(k+2)}\cdot \\
&\log\frac{[P_1(x_1)\times\cdots\times P_{k+1}(x_{k+1})\times P_{k+2}(y')]^{k+3}}{[P_1(x_1)\times\cdots\times P_{k+2}(x_{k+2})\times P_{k+3}(y')]^{k+2}} \\
=& \frac{1}{(k+3)(k+2)}\cdot\log\{P_1(x_1)\times\cdots\times P_{k+1}(x_{k+1}) \\
&\frac{[P_{k+2}(y')]^{k+3}}{[P_{k+2}(x_{k+2})]^{k+2}\cdot[P_{k+3}(y')]^{k+2}}\}
\end{aligned} \tag{20}$$

Since exploring the training dynamics of $P_i(x_i)$, $P_j(y')$ in large language models is intractable, we make a mild assumption here, at the late training stage, $P_i(x_i) \rightarrow \hat{P}_i(x_i)$, $P_j(y') \rightarrow \hat{P}_j(y')$, in the setup with controlled variables, where samples with different lengths have same proportion of dominant conditions and suppressed conditions, then the value in log approaches $\frac{P_{k+2}(y')}{P_{k+2}(x_{k+2})}$. Since y' is the false prediction made by model, whose empirical probability equals zero, so $P_{k+2}(y')$ approaches zero, then $P_{k+2}(y') < P_{k+2}(x_{k+2})$.

Given that, $\frac{P_{k+2}(y')}{P_{k+2}(x_{k+2})} < 1$, therefore, $L_{NTP}(y'|x_{1:k+1}, x_{k+2}) < L_{NTP}(y'|x_{1:k}, x_{k+1})$, substituting k with L, we denote $L_{NTP}(y'|x_{1:L}, x_{L+1})$ as $-\log\left(\frac{e^{f(\boldsymbol{x})_y}}{\sum_{y'} e^{h^{-1}(L)f(\boldsymbol{x})_{y'}}}\right)$, where $h(L)$ is positively correlated with L, with larger L indicating larger $h(L)$.

Lipschitz continuity of NTP loss. $B_y(f)$ represents the minimal prediction on the ground truth token y, *i.e.* $B_y(f) := min_{x\in S_y} f(x)_y$ (Wang et al., 2024).

Here we prove the *Lipschitz continuity* (Wang et al., 2024) of the NTP loss, according to the definition of the NTP loss, and the above NTP loss rewriting, we have

$$\begin{aligned}\mathcal{L}_{\text{NTP}}(f(\boldsymbol{x}), y) &= -\log\left(\frac{e^{f(\boldsymbol{x})_y}}{\sum_{y'} e^{h^{-1}(L) f(\boldsymbol{x})_{y'}}}\right) \\ &= \log[1 + \sum_{y' \neq y} e^{h^{-1}(L) f(\boldsymbol{x})_{y'} - f(\boldsymbol{x})_y}].\end{aligned} \tag{21}$$

We denote $\boldsymbol{s} := f(\boldsymbol{x})$, and we define

$$\ell_y(\boldsymbol{s}) := \sum_{y' \neq y} e^{h^{-1}(L) \boldsymbol{s}_{y'}}.$$

Therefore, we rewrite the $\mathcal{L}_{\text{NTP}}$ as follows:

$$\mathcal{L}_{NTP}(f, y) = \log\left[1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})\right].$$

The derivatives can be represented as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}_{NTP}(f, y)}{\partial \boldsymbol{s}_y} &= -\frac{e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}{1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}, \\ \frac{\partial \mathcal{L}_{NTP}(f, y)}{\partial \boldsymbol{s}_{y'}} &= h^{-1}(L) \frac{e^{-\boldsymbol{s}_y}}{1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})} \cdot e^{h^{-1}(L) \boldsymbol{s}_{y'}}, y' \neq y.\end{aligned} \tag{22}$$

We can get the following inequality:

$$\begin{aligned}&\|\nabla_{\boldsymbol{s}} \mathcal{L}_{NTP}(f, y)\|^2 = \\ &\left[\ell_y(\boldsymbol{s})^2 + \sum_{y' \neq y} \left(h^{-1}(L) e^{h^{-1}(L) \boldsymbol{s}_{y'}}\right)^2\right] \\ &\times \left[\frac{e^{-\boldsymbol{s}_y}}{1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}\right]^2 \\ &\leq \left[\ell_y(\boldsymbol{s})^2 + \left(\sum_{y' \neq y} h^{-1}(L)\right)^2 \left(\sum_{y' \neq y} e^{h^{-1}(L) \boldsymbol{s}_{y'}}\right)^2\right] \\ &\times \left[\frac{e^{-\boldsymbol{s}_y}}{1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}\right]^2 \\ &= \left[1 + \left(\sum_{y' \neq y} h^{-1}(L)\right)^2\right] \cdot \left[\frac{e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}{1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}\right]^2,\end{aligned} \tag{23}$$

Therefore,

$$\begin{aligned}\|\nabla_{\boldsymbol{s}} \mathcal{L}_{NTP}(f, y)\| &\leq \sqrt{1 + \left(\sum_{y' \neq y} h^{-1}(L)\right)^2} \frac{e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})}{1 + e^{-\boldsymbol{s}_y} \ell_y(\boldsymbol{s})} \\ &= \sqrt{1 + \left(\sum_{y' \neq y} h^{-1}(L)\right)^2} \frac{\ell_y(\boldsymbol{s})}{e^{\boldsymbol{s}_y} + \ell_y(\boldsymbol{s})} \\ &= \sqrt{1 + \left(\sum_{y' \neq y} h^{-1}(L)\right)^2} \left[1 - \frac{e^{\boldsymbol{s}_y}}{\sum_{y'} e^{h^{-1}(L) \boldsymbol{s}_{y'}}}\right] \\ &= \sqrt{1 + \left(\sum_{y' \neq y} h^{-1}(L)\right)^2} \left[1 - \textit{softmax}\,(\boldsymbol{s}_y)\right].\end{aligned} \tag{24}$$

Since the score function is bounded, for any $y \in \mathcal{Y}$, there exists a constant $B_y(f)$ such that $B_y(f) = \inf_{\boldsymbol{x} \in \mathcal{S}_y} \boldsymbol{s}_y$, which completes the proof.