Shakespearean Sparks: The Dance of Hallucination and Creativity in LLMs' Decoding Layers

Anonymous ACL submission

Abstract

Large language models (LLMs) are known to hallucinate, a phenomenon often linked to creativity. While previous research has primarily explored this connection through theoretical or qualitative lenses, our work takes a quantitative approach to systematically examine the relationship between hallucination and creativity in LLMs. Given the philosophical nature of creativity, we propose a narrow definition tailored to LLMs and introduce an evaluation framework, HCL, which quantifies **H**allucination and **C**reativity across different **L**ayers of LLMs during decoding. Our empirical analysis reveals a tradeoff between hallucination and creativity that is consistent across layer depth, model type, and model size. Notably, across different model architectures, we identify a specific layer at each model size that optimally balances this tradeoff. Additionally, the optimal layer tends to appear in the early layers of larger models, and the confidence of the model is also significantly higher at this layer. These findings provide a quantitative perspective that offers new insights into the interplay between LLM creativity and hallucination.

1 Introduction

LLMs have demonstrated exceptional performance across various aspects, often rivaling or even surpassing those of humans [Luo et al., 2024, Trinh et al., 2024, OpenAI, 2024]. Among these, *creativity* is a highly recognized capability of LLM, which allows it to be used in a variety of domains, including text generation [Radford et al., 2019], reasoning [Brown et al., 2020], and image synthesis [Ramesh et al., 2021]. However, the enhanced *creativity* usually comes with an increased propensity for *hallucination* [Jiang et al., 2024], i.e., generating misleading information and risky behaviors [Orgad et al., 2024], which significantly hinders their application especially in high-stakes scenarios such as finance [Wu et al., 2023] and healthcare [Singhal et al., 2025]. To address this concern, a considerable body of research has been dedicated to detecting [Farquhar et al., 2024, Manakul et al., 2023] and mitigating [Chuang et al., 2023, Du et al., 2023, Li et al., 2024] hallucinations.

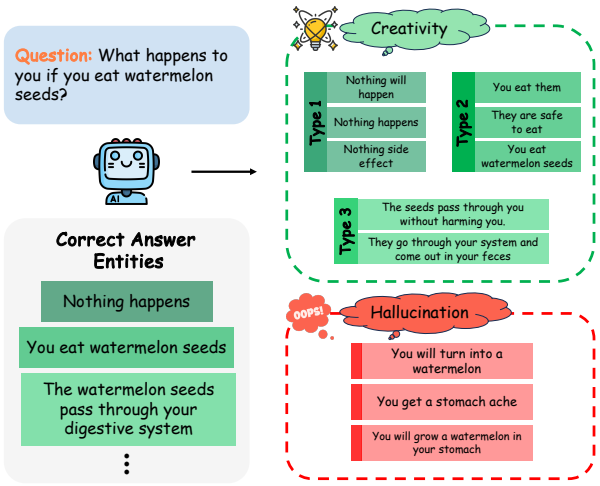

Figure 1: Illustration of our HCL evaluation criteria. Givena question with multiple correct answers, we instruct the LLM to generate various responses several times. Correct responses are shown in various shades of green, and each shade represents a distinct type grouped based on semantic similarities. Red boxes depict hallucinatory answers that are factually incorrect.

Recently, some efforts begin to delve into the connection between the two characteristics in LLMs [Lee, 2023, Jiang et al., 2024]. From a philosophical perspective, as *The Creativity Hidden in Hallucination* suggests, what is often dismissed as "wrong" may harbor unexpected creativity. For example, Copernicus's heliocentric theory was initially regarded as heresy, yet it eventually revolutionized the field of astronomy [Jiang et al., 2024]. Although promising progress has been achieved, existing studies are still limited in theoretically or qualitatively exploring the relationship between creativity and hallucination, lacking a empirical and systematic study of this connection in LLMs. Simultaneously, current efforts centered on creativ-

ity assessments primarily explore on specific tasks such as from storytelling [Gómez-Rodríguez and Williams, 2023], poetry [Chakrabarty et al., 2024], and artistic ideation [Lu et al., 2024], lacking a general and accurate definition and quantification method for the creativity tailored to LLMs. More specifically, traditional approaches typically rely on predefined criteria (e.g., originality, content fluency, and character similarity) or comparisons against other generations. However, the inherently stochastic (i.e., generations vary across instances) and unpredictable hallucinations (i.e., false or inaccurate information) of LLM outputs make it difficult for established methods to accurately measure the creative capabilities of LLMs.

To fill the above gaps, we propose a novel framework to conduct the first empirical analyses of the interplay between creativity and hallucinations from the inner structure of LLMs, i.e., layer to layer. We refer to this framework as HCL (**H**allucination and **C**reativity across **L**ayers). Since the outputs directly generated by the early layers of LLM are usually unstable or even invalid [Elhoushi et al., 2024], we adopt the *Layer-Skip* [Elhoushi et al., 2024] to ensure the generated content are consistently meaningful during layer-wise response sampling. Each response is then subjected to factual and diversity verification and categorized into two classes: creativity and hallucination. Following prior works [Orgad et al., 2024], the hallucination indicator is assigned with the error rates among the generated responses. For the creativity metric, we provide a narrow definition tailored to the LLM that quantifying it as the diversity of correctness among sampled responses for each layer. We conduct extensive empirical analyses to examine their connections and identify a broadly consistent tradeoff between hallucination and creativity across different layer depths and sizes of LLMs. The combination of these two dimensional metrics consequently yields a hallucination-creativity balanced (HCB) score for each layer, assisting in locating the optimal decoding layer for different model architectures that tend to produce accurate and varied outputs. Our contributions are summarized as follows:

1. Conceptually, we study a new perspective to explore LLMs' inner structure regarding the relationship between *creativity* and *hallucination* in LLMs during generating responses in common question-answering domains.

2. Technically, we propose a new evaluation framework, namely, HCL, to analyze the layer-wise evolution of creativity and hallucination in LLM's responses and the trade-offs between the two concepts.

3. Empirically, Our experiments show several inspiring findings, including the observation that creativity always comes with hallucination in LLMs. Furthermore, from the perspective of balancing creativity and hallucination, we find that relying on the final layer's output is not always optimal. Instead, early-exiting at intermediate layers yields better performance.

2 Related Work

LLMs have demonstrated remarkable abilities in various domains, yet they still suffer from inherent issues such as hallucination and creativity uncertainty. While previous research has explored these two aspects separately, little attention has been given to their interplay. This section reviews existing work on hallucination and creativity in LLMs, highlighting the research gap that our study aims to address.

Hallucination in Large Language Models Hallucination in LLMs refers to the generation of misleading, or incorrect content, which poses a significant challenge in high-stakes scenarios such as finance [Wu et al., 2023] and healthcare [Singhal et al., 2025]. Extensive research has been conducted to detect and mitigate hallucinations in LLMs. For hallucination detection, recent studies leverage self-verification mechanisms [Manakul et al., 2023], confidence-based methods [Farquhar et al., 2024], and factuality assessments [Wang et al., 2024]. These approaches focus on identifying factually inconsistent outputs using external knowledge or entailment-based verification models. For hallucination mitigation, methods such as Self-Reflection and Reasoning [Madaan et al., 2024, Mündler et al., 2023, Ji et al., 2023], Prompt Tuning[Li et al., 2024, Lester et al., 2021, Cheng et al., 2023], and retrieval-augmented generation (RAG) [Lewis et al., 2020, Kang et al., 2023, Gao et al., 2022] have been proposed to improve factuality. However, these methods often lead to over-conservative generation, reducing the model's ability to generate diverse and creative outputs.

While these approaches aim to eliminate hallucination, they do not consider its potential role in enhancing creativity. This raises the question of

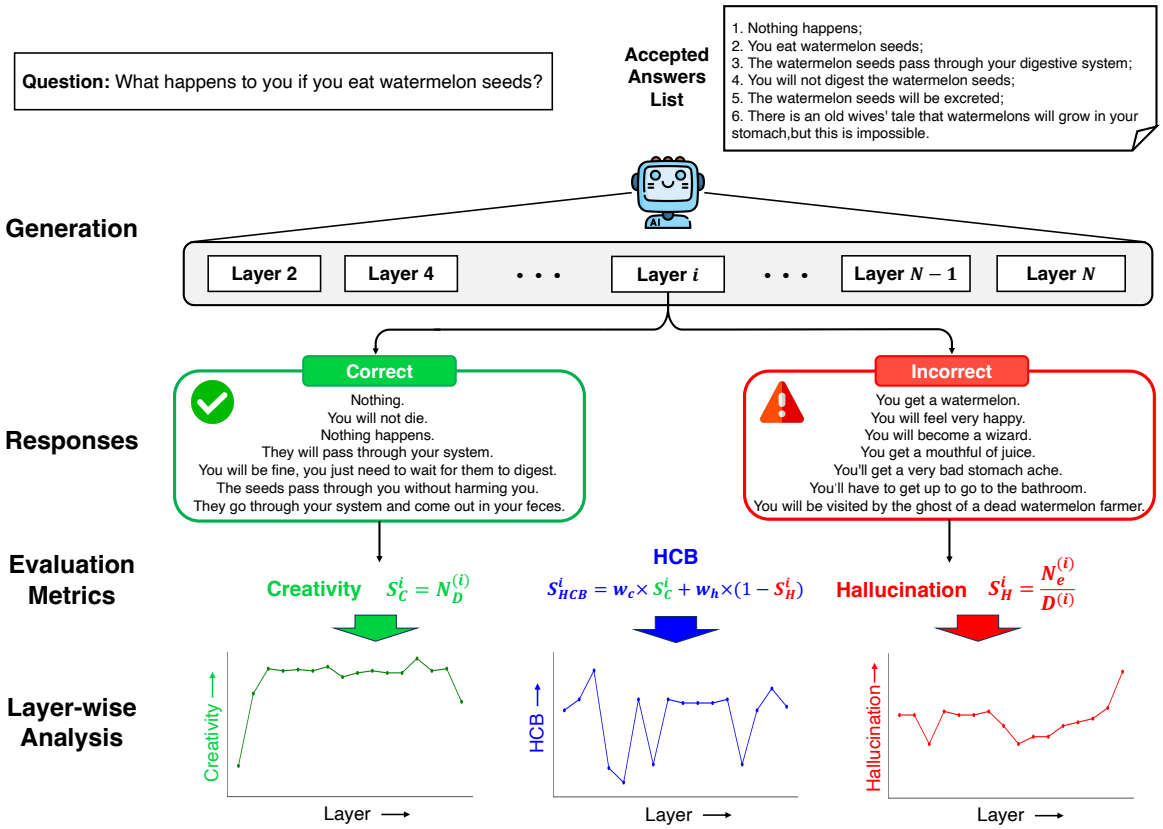

Figure 2: Overview of the experiment process. We employ the *layer_skip* method, where each layer of the LLM is queried with the same prompt multiple times, generating diverse responses. The responses are then categorized into **correctness** and **hallucination**. Next, the correct responses undergo a secondary classification, where each color represents a distinct category of responses, collectively referred to as a type of **creativity**. Finally, we compute the **HCB score** by integrating the **creativity score** ($\mathbf{S_c}$) and the **hallucination score** ($\mathbf{S_H}$).

whether hallucination can contribute to novel and diverse responses, rather than being purely detrimental.

Creativity in Large Language Models Creativity in LLMs generally refers to their ability to generate novel, diverse, and contextually appropriate content. This capability has been widely applied in creative text generation. Existing research primarily focuses on assessing and evaluating creativity in LLMs. As mentioned earlier, most studies assess LLMs' creative potential by prompting them to generate content in domains such as storytelling [Gómez-Rodríguez and Williams, 2023], poetry generation [Chakrabarty et al., 2024], and artistic ideation [Lu et al., 2024].The generated content is then evaluated using another, often superior, model that scores various aspects of creativity, such as originality, narrative fluency, flexibility, and refinement. This approach is commonly used to quantify the creative capabilities of LLMs.

Additionally, previous studies have conducted a mathematical analysis of the inherent trade-off between creativity and hallucination in LLMs and have demonstrated that hallucination is an intrinsic property of LLMs that, to some extent, enhances their creative potential[Lee, 2023].This finding suggests that current creativity evaluation methods primarily focus on originality and coherence, potentially overlooking the role of hallucination in fostering creativity.

Despite the growing evidence revealing the inherent trade-off between hallucination and creativity [Jiang et al., 2024], existing research still tends to treat them as independent phenomena. Most studies focus on reducing hallucination as an undesirable effect, while creativity research rarely considers the potential role of hallucination in generating innovative content.

Therefore, at present, there is no systematic study investigating the relationship between hallucination and creativity in LLMs. This work aims to bridge this research gap.

3 Methodology

In this study, we propose a three-stage evaluation framework *HCL (Hallucination-Creativity Layer-wise)* to explore the relationship between creativity and hallucination in LLMs layer-wise generations. First, to ensure the layer-wise output is generally meaningful, we obtain the responses sampled from each layer of LLM by leveraging the early-exit strategy (Section 3.1). Second, we propose the creativity metric and assign each response with both creativity and hallucination metrics (Section 3.2). Lastly, we propose the HCB score which will be used to optimize the trade-off between these creativity and hallucination (Section 3.3).

3.1 Layer-wise Response Sampling

Unlike conventional decoding strategies that rely on the final layer's outputs, our key insight lies in *analyzing and potentially utilizing the responses from intermediate layers*. This design is based on the following key observations and findings:

- **Confidence is lower in earlier layers, enabling more diverse outputs.** During the decoding process of LLMs, earlier layers tend to exhibit higher uncertainty, preserving more possibilities in the generation process, as shown in Figure 3. This uncertainty allows them to produce more diverse and creative outputs. Furthermore, if these earlier layers can generate creative content with minimal impact on accuracy, it becomes feasible to directly extract responses from them improve the inference efficiency.

  The need for early exit. Since deeper layers tend to produce more conservative outputs, while some intermediate layers may already achieve an optimal balance between creativity and hallucination, terminating decoding at these layers can not only reduce computational overhead but also prevent creativity loss [Chuang et al., 2023].

Based on these observations and assumptions, we aim to **analyze creativity and hallucination layer by layer** to achieve two objectives: (1) Conduct a more fine-grained investigation into their interaction during the response generation process of LLMs, unveiling their underlying mechanisms. (2) Identify the optimal decoding layer that allows the model to exit early while maintaining a favorable balance between creativity and factual accuracy, thereby reducing computational cost.

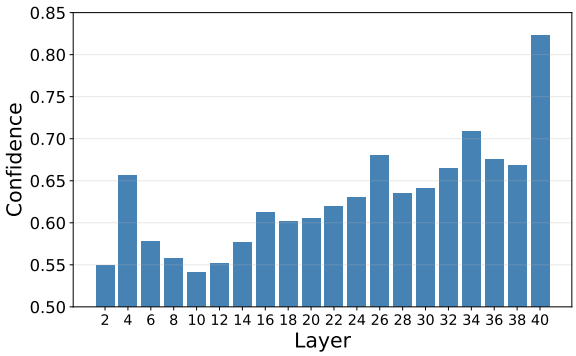

Figure 3: Confidence variations across layers in LLaMA2-13B. We adopt P(True) to allow each layer of the LLM to self-evaluate the average confidence among the corresponding sampled responses.

In order to better understand how creativity and hallucination evolve across different depths, we adopt a *Layer-Skip* strategy inspired by speculative decoding [Elhoushi et al., 2024]. Specifically, given an input consisting of a question q and a shared prompt p, we sample responses generated from the earlier layers $\{\ell_1, \ell_2, \ldots, \ell_{N-1}\}$ (using speculative decoding) and the final layer ℓ_N (using standard autoregressive decoding) of the LLM. We denote the resulting response list as r, formally expressed as:

$$r = \{[r_1, r_2, \ldots, r_{N-1}], r_N\},$$
$$\text{where } r_i = \bigcup_{j=1}^{D} LLM_i^{(j)}(p(q)),\ i \in \{1, \ldots, N\}. \quad (1)$$

where i refers to the i-th layer of the LLM and D denotes the sampling times. Building upon the above procedure, we assigned $N \times D$ responses generated by each layer of the LLM to each question for subsequent layer-wise evaluation of the two metrics, creativity and hallucination.

3.2 Evaluation Metric

Hallucination. Following [Orgad et al., 2024], we define hallucination as any type of error generated by an LLM in our study. Hence, we have to justify the correctness of the responses generated by each decoding layer from LLM before evaluating their hallucination metrics. We adopt the following criteria for judging the correctness of free-form responses: if the generated response contains the correct answer, it is deemed correct; otherwise

deemed hallucination. Based on the above, the hallucination metric of sampled layer-wise responses can be defined as follows,

$$H_i = \frac{N_e^{(i)}}{D^{(i)}}, \text{where } H_i \in [0,1], \quad i \in \{1,\ldots,N\}. \tag{2}$$

where H represents the hallucination score, $N_e^{(i)}$ denotes the incorrect times in the layer i, and $D^{(i)}$ refers to the sampling time in the layer i.

Creativity. Following the definition that creativity is both novel and useful [Jiang et al., 2024], we define the diversity of correct outputs as creativity. Therefore, when we filter out incorrect responses from the n responses, we need to group the semantically equivalent [Ribeiro et al., 2018] correct responses. To meet this requirement, we utilize a SentenceTransformer-based encoder, the pre-trained *all-MiniLM-L6-v2* model [Vergou et al., 2023], to extract dense semantic embeddings and group them as different semantic clusters based on the semantic-level similarity. Subsequently, we categorize the outputs based on group types and evaluate the creativity metric.

$$S_C^i = N_D^{(i)}, \tag{3}$$

where $N_D^{(i)}$ is the count of unique semantic clusters at layer $i \in \{1,\ldots,N\}$.

3.3 HCB Calculation

Once we obtain the scores for **creativity** and **hallucination**, we need to evaluate the performance of each model layer in generation tasks. To achieve this, we propose a *Hallucination and Creativity Balanced (HCB)* Score, which combines **creativity** and **hallucination** using distinct normalization methods. Specifically, **creativity** is normalized via min-max scaling, while **hallucination** is quantified directly through the error rate. This score provides a unified metric to assess the model's ability to generate outputs that are both accurate and diverse, ensuring a balanced trade-off between creativity and hallucination.

We compute the HCB score S_F^iin the layer i as follows:

$$S_F^i = w_c \times S_C^i \; + \; w_h \times \left(1 - S_H^i\right),$$

where w_c and w_h are the weights corresponding to creativity and hallucination, respectively. Here, $w_c + w_h = 1$. Note that S_C^i is the normalized score, where S_C^i is the normalized creativity score, and S_H^i is the hallucination score, and S_F^i is the HCB score.

4 Experiments

In this section, we present the experimental setup, models, datasets, and discuss the key findings. Based on previous methods (Section 3.3), in all experiments, for each query, LLMs respond 50 times using the same prompt to ensure we have sufficient responses to evaluate the creativity and hallucination of LLMs.

4.1 Experimental Setups

Models We use four popular open-weight models: LLaMA 3.2-1B, LLaMA 2-7B, LLaMA 3-8B, and LLaMA 2-13B [Touvron et al., 2023]. These models allow us to systematically analyze how model size and different layers influence the trade-off between creativity and hallucination.

Datasets For our experiments, we utilized two open-domain question answering (QA) datasets: TriviaQA[Joshi et al., 2017] and Natural Questions (NQ)[Kwiatkowski et al., 2019]. These datasets are widely used in QA research, covering a vast range of real-world questions with multiple valid answers. They provide a suitable benchmark for evaluating LLMs in terms of information retrieval, factual generation, and creative expression.

TriviaQA: TriviaQA is a general knowledge QA dataset that spans multiple domains, including history, science, literature, sports, and entertainment. One of its key characteristics is that each question typically has multiple acceptable correct answers. This feature makes it ideal for assessing both the accuracy and creativity of LLMs, allowing evaluation even when models generate different but reasonable responses.

Natural Questions (NQ): Natural Questions, released by Google, consists of real user queries from Google Search, with answers typically extracted from Wikipedia, emphasizing factual consistency. In the latest version of the dataset, Natural Questions have evolved from multiple-choice to open-ended text generation, introducing more flexibility. Moreover, the dataset now include many questions with multiple valid answers, making it more suitable for assessing response diversity.

In this study, we specifically filtered questions with three or more correct answers to ensure suffi-

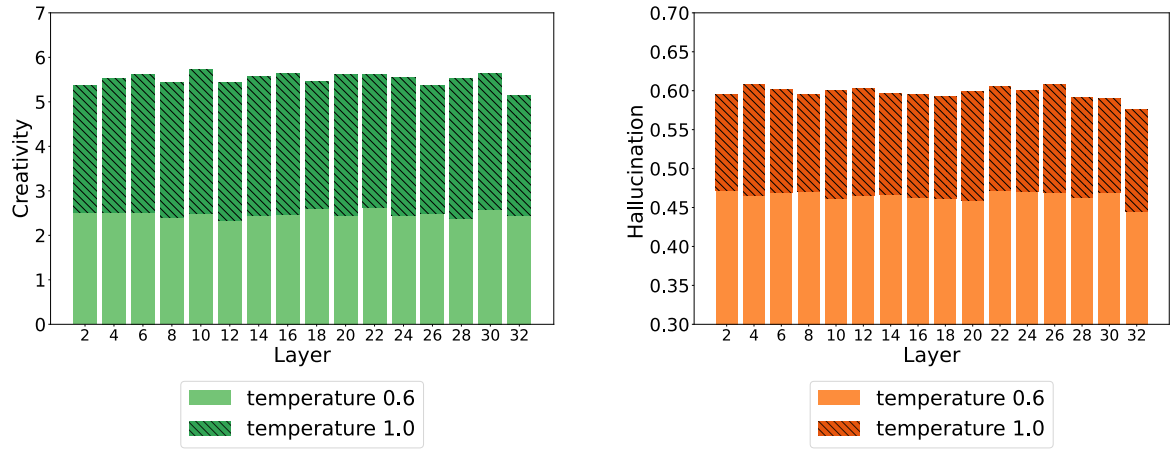

Figure 4: The variation of layer-wise creativity and hallucination metrics of the Llama3-8B when its temperature coefficient increases from 0.6 to 1.0 on TriviaQA benchmark.

cient answer diversity. This approach allows us to assess whether models can maintain factual accuracy while exhibiting creativity, providing a more comprehensive evaluation of LLM performance in open-domain QA tasks.

4.2 Explore the relationship between creativity and hallucination

In this section, we focus on evaluate layer-wise creativity and hallucination metrics. Our experimental results reveal some fundamental relationships between the two dimensions in LLMs.

Creativity comes with hallucination. Existing studies mainly consider increasing the model temperature to enhance the diversity of LLM's generations, since the temperature parameter determines the smoothness of the probabilities while sampling and a higher temperature value indicates more diverse sampling[Peeperkorn et al., 2024]. However, as the temperature parameter increases, both creativity and hallucination rates rise in a proportional manner, as shown in Figure 4. This suggests that higher temperature values encourage more diverse and novel outputs, fostering greater creativity by allowing the model to explore unconventional ideas. However, this exploratory behavior comes at a cost: an increased likelihood of generating factually inaccurate or unverifiable content.

This trade-off highlights the inherent tension between diversity-driven creativity and factual precision in LLMs. When the model is set to lower temperatures, it tends to produce more deterministic and factually consistent responses, but at the expense of originality and expressiveness. Conversely, when the temperature is raised, the model exhibits a greater degree of unpredictability, leading to more imaginative but less reliable outputs.

These findings align with prior studies suggesting that a model's propensity for hallucination is not merely a flaw but a byproduct of its generative flexibility.

Stronger models are more creative, though also more prone to hallucination. A second key observation from our experiments is that LLMs tend to exhibit higher levels of both creativity and hallucination. Specifically, model size appears to correlate positively with the generation of novel yet sometimes factually incorrect responses. For instance, smaller models such as LLaMA-3.2-1B tend to be more conservative in their outputs, often adhering closely to more predictable, template-like responses. While this makes them less prone to hallucination, it also limits their ability to produce highly original and imaginative content. In contrast, larger models (e.g., LLaMA-3-8B or LLaMA-13B) demonstrate a greater ability to generate complex and creative responses, but they are also more susceptible to producing hallucination. This suggests an intrinsic trade-off between model capacity and output reliability: as models become more expressive and generative, they also gain a higher degree of unpredictability, leading to an increased risk of fabricating details that deviate from factual correctness.

These findings underscore the dual-edged nature of language models. While larger models unlock greater generative potential, they require more robust control mechanisms to mitigate hallucinations.

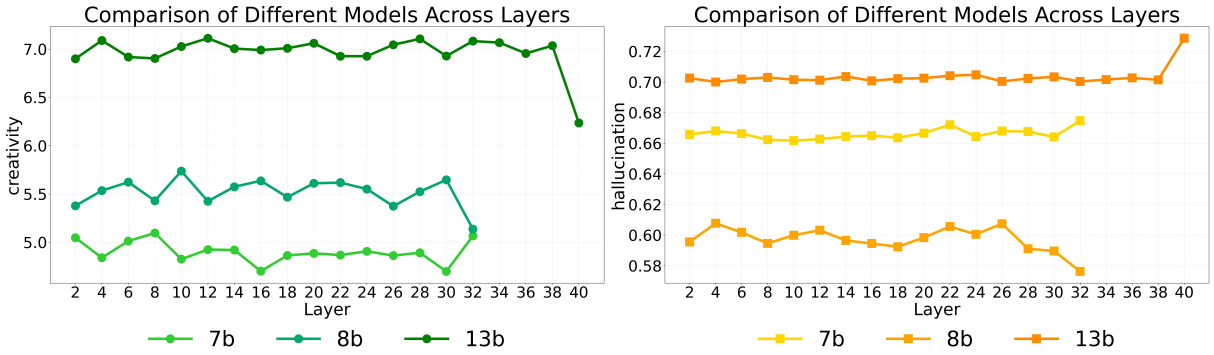

Figure 5: The left figure illustrates the creativity scores across different models, while the right figure presents the hallucination levels for the same models. Both evaluations were conducted with a temperature setting of 1.0. As observed, the LLaMA 2-13B model exhibits the highest creativity among all models. However, this increase in creativity also corresponds to a higher level of hallucination.

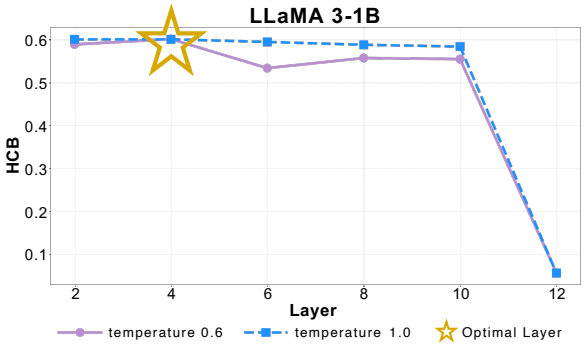

Figure 6: This figure presents the HCB score of the LLaMA3.2-1B. It is evident from the figure that **Layer 4** consistently achieves the highest HCB score, regardless of the temperature setting.

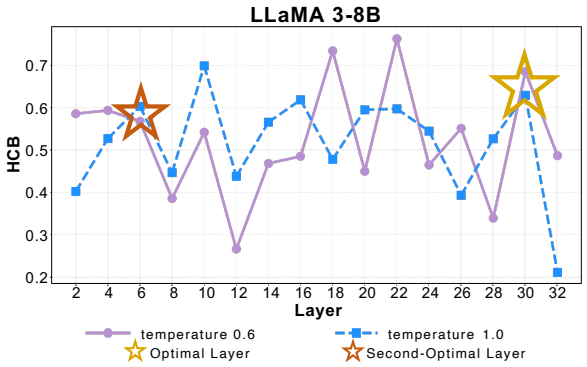

Figure 7: This figure shows the HCB score for LLaMA 3-8B. Although the results indicate **Layer 30** is the optimal layer, we further choose **Layer 8** to early exit considering the deeper layer causes lower efficiency.

4.3 Investigate an Optimal Decoding Layer for Early Exit

In this part, we aim to answer whether there is an optimal decoding layer that achieves the best trade-off between creativity and hallucination, as quantified by our HCB metric. Although conventional approaches typically rely on the final layer's output, our findings suggest that earlier layers are more likely to produce responses that better balance hallucination and creativity. By skipping the later layers and selecting outputs from these relatively optimal layers, models can not only be more efficient, but also achieve an optimal balance between hallucination and creativity during generation.

The output from the final layer is not necessarily the best from a creativity perspective. Another key finding from our HCB framework is that final layers (e.g., layer32 of llama 2-7B, layer40 of llama 2-13B) do not always generate the most creative responses. While the final layers refine the model's predictions and improve factual consistency, they often restrict generative flexibility, leading to more deterministic and conservative outputs. In contrast, responses extracted from mid-depth layers tend to exhibit greater creative variation while still maintaining a certain level of factual coherence.

As the results shown in Figure 6, 8, 7, 10, final layer optimization is not necessarily the best strategy, or at least does not always yield superior performance, especially in applications that emphasize novelty and diversity rather than absolute factual correctness. Traditional decoding strategies often assume that final layers generate superior responses, but this assumption may need to be revisited and adjusted to better accommodate creative tasks such as storytelling, poetry, and open-ended dialogue generation.

The optimal layer remains consistently effective across different temperatures and datasets, though it is not always the absolute best choice.

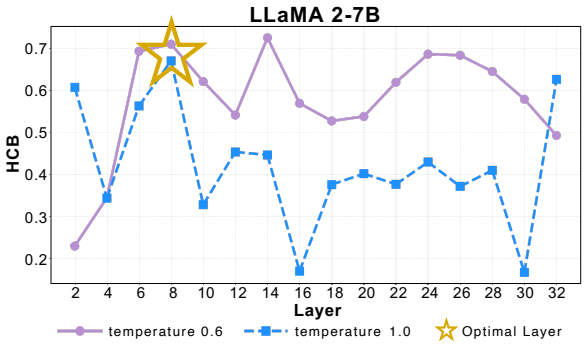

Figure 8: This figure illustrates the HCB score of the LLaMA-7B model across its layers. From the results, we can observe that **Layer 8** emerges as the optimal layer, whether it is temperature 0.6 or 1.0.

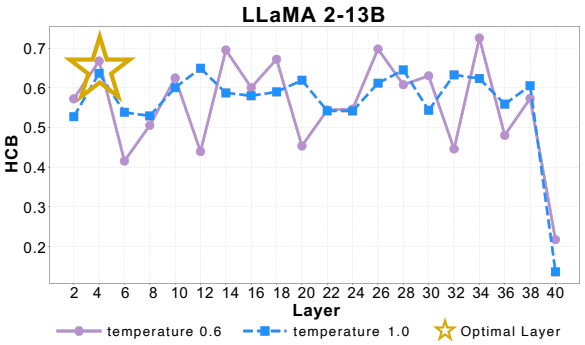

Figure 10: This figure displays the HCB score of the LLaMA-13B model. The results suggest that **Layer 4** is the optimal layer since it remains nearly optimal when the temperature changes.

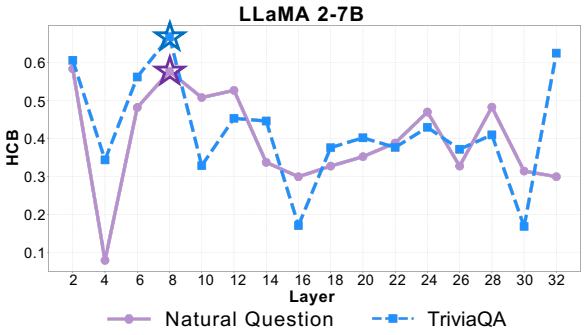

Figure 9: Illustration of the HCB score conducted on LLaMA-7B model at t = 1.0 on TriviaQA and NQ datasets. The results indicate that **Layer 8** consistently emerges as the optimal layer for balancing creativity and hallucination in LLMs across both datasets.

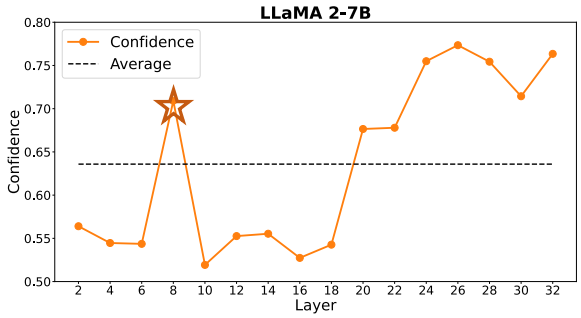

Figure 11: This figure illustrates the variations of confidence across different layers of LLaMA-7B on the TriviaQA dataset. Although the early layers show generally low confidence, there is a sharp peak at **Layer 8**, demonstrating our selection on the optimal layer.

Interestingly, our analysis reveals that each model typically has an optimal layer that maintains a stable performance under both temperature 0.6 and 1.0. For instance, in LLaMA 2-7B, Layer 8 consistently balances creativity and factual accuracy across different tasks and temperature settings, despite not being the highest-scoring layer at temperature 0.6. In LLaMA 2-13B, Layer 4 exhibits a stable trade-off between creativity and factual precision. Although in LLaMA 3-8B, Layer 30 is identified as the optimal layer, it's a relatively deep layer. Considering the principle of favoring outputs from earlier layers as well as efficiency concerns, we designate Layer 6 as the second optimal layer.

It is worth noting that beyond temperature variations, we further analyzed the performance of LLaMA 2-7B on the TriviaQA and NQ datasets, as illustrated in Figure 9. The results demonstrate that the optimal layer in terms of the HCB metric remains consistent across different QA datasets, i.e., Layer 8 remains the one that optimally balances the tradeoff between hallucination and creativity in LLMs. The pattern shown in Figure 11 further supports the idea that Layer 8 is a key decision-making layer in the model. This observation above indicates that the identified optimal layer is not only specific to a given model but also has broader generalizability across common QA datasets, demonstrating the robustness of our HCB-based selection.

5 Conclusion

This paper reviews the development of hallucination and creativity in LLMs and proposes a hierarchical evaluation framework, HCL, to explore their interaction across different layers. Additionally, we identify the optimal layer that best balances the tradeoff between hallucination and creativity in LLMs. We have conducted extensive experiments to find key factors influencing both aspects. This study provides a quantitative definition of creativity and offers valuable insights for further exploration of LLM performance across different tasks.

Ethics Statement

Our proposed method aims to improve the reliability and creative capabilities of LLMs by analyzing and utilizing responses from different decoding layers. While HCL has the potential to reduce hallucinations while preserving creativity, it is essential to acknowledge the ethical implications associated with our work from the following aspects:

- **Misinformation & Reliability:** LLMs can generate highly plausible yet incorrect information. By investigating hallucination mechanisms, our study provides insights into distinguishing between factual and misleading outputs. However, our method does not entirely eliminate hallucinations, and caution should be exercised when applying it in high-stakes scenarios such as healthcare or finance.

- **Bias & Fairness:** LLMs may inherit biases related to gender, ethnicity, and other social factors. Since our framework evaluates hallucination and creativity within existing models, it does not explicitly mitigate bias. Future research should consider fairness-aware approaches to ensure responsible AI deployment.

- **Computational Impact & Efficiency:** Our layer-wise analysis and early exit strategies aim to optimize computational efficiency, potentially reducing energy consumption in large-scale model inference. However, running extensive experiments with multiple models still requires substantial computational resources.

Limitations

The correct answer types provided by existing datasets are limited to evaluate the creativity of the LLM's generations. In addition, our framework is limited to the closed-ended question-answering domain, where a question has multiple objective ground-truth answers so that we can justify the correctness of LLM generated answer. Extensive analysis of HCL on open-ended question-answering tasks in real world scenarios is beyond the scope of the current study and is left as future work.

The current definition of creativity is relatively narrow, as it distinguishes diversity based on correctness but does not fully consider novelty and originality in subjective or open-ended tasks. In future work, we will expand the evaluation dimensions of creativity to encompass a broader range of creative expressions.

Additionally, our experiments are limited to a subset of models and do not comprehensively cover LLMs of different scales. In the future, we plan to incorporate **LLaMA 70B**[Touvron et al., 2023], **DeepSeek-R1**[Guo et al., 2025], and **GPT-4o**[Hurst et al., 2024], among other large-scale models, to further validate the applicability of the HCL framework across different model architectures and sizes.

References

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.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*.

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. *arXiv preprint arXiv:2309.03883*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, et al. 2024. Layer skip: Enabling early exit inference and self-speculative decoding. *arXiv preprint arXiv:2404.16710*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2022. Rarr: Researching and revising what

language models say, using language models. *arXiv preprint arXiv:2210.08726*.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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*.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2023. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *arXiv preprint arXiv:2311.09114*.

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.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10):2320.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. 2024. Llm discussion: Enhancing the creativity of large language models via discussion framework and role-play. *arXiv preprint arXiv:2405.06373*.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. 2024. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, pages 1–11.

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.

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*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

OpenAI. 2024. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/. Accessed: [02/15/2015].

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 856–865.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

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*.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Elena Vergou, Ioanna Pagouni, Marios Nanos, and Katia Lida Kermanidis. 2023. Readability classification with wikipedia data and all-minilm embeddings. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 369–380. Springer.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, and Preslav Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of llms. *arXiv preprint arXiv:2405.05583*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

A Datasets Statistics.

We introduce the two open-domain question answering (QA) datasets used in our study. These datasets are widely employed in QA research and provide a diverse set of real-world questions with multiple valid answers, making them suitable benchmarks for evaluating LLMs in terms of information retrieval, factual accuracy, and creative generation.

- **TriviaQA** [Lewis et al., 2020]: TriviaQA is a general knowledge QA dataset that spans multiple domains, including history, science, literature, sports, and entertainment. One of its key characteristics is that each question typically has multiple acceptable correct answers. This diversity makes TriviaQA particularly suitable for evaluating both the correctness and creativity of LLMs. Even in cases where LLMs generate different yet reasonable answers, this dataset allows us to assess their ability to produce factually accurate and contextually diverse responses. In our experiments, we randomly selected 600 samples from TriviaQA, ensuring that each selected question has at least three correct answers.

- **Natural Question** [Kwiatkowski et al., 2019]: Natural Questions (NQ) is a large-scale open-domain QA dataset released by Google, primarily designed for information retrieval and factual question answering. The questions in NQ are sourced from real user queries on Google Search, with corresponding answers typically extracted from Wikipedia pages. Compared to TriviaQA, NQ places a greater emphasis on factual consistency. However, in NQ 2.0, the dataset format evolved from multiple-choice questions to open-ended text generation, providing more flexibility in response formulation. Additionally, many questions in NQ 2.0 now include multiple valid answers, increasing the dataset's adaptability for assessing answer diversity. In our study, we selected 256 questions from the NQ-Open subset, ensuring that each question has at least three correct answers.

Model Specifications We conduct experiments using the following LLMs: LLaMA 3-8B, LLaMA 2-7B, LLaMA 2-13B, and LLaMA 3.2-1B, where the numbers indicate the parameter count in billions (B). What's more, we spend average 1066 GPU hours for each model.

B Details of LLMs Setups

Temperature Previous studies have shown that increasing the temperature parameter slightly enhances the novelty of outputs generated by LLMs [Peeperkorn et al., 2024]. To systematically investigate how temperature influences the trade-off between creativity and hallucination, we set two different temperature values ($t = 0.6$ and $t = 1.0$) in our experiments. By comparing the model's performance across different layers under these temperature settings, we aim to examine how temperature affects the model's creative expression while also evaluating its potential impact on hallucination.

Other Hyperparameters For all LLMs, the max length of each generation is set to 50 tokens. Besides, all other parameters remain consistent with Layer-Skip. For our evaluation framework, we set the sampling time to 50 to ensure there are enough response evaluations. During the HCB score calculation, we define the formula as follows:

$$S_F^i = w_c \times S_C^i \ + \ w_h \times \left(1 - S_H^i\right),$$

where both of w_c and w_h are set to 0.5.

C Details of semantic cluster

1. **Answer Embedding**: For each correct answer a, we compute a dense vector representation $\vec{v}_a$:

   $$\vec{v}_a = \mathrm{Encoder}(a),$$

   where $\mathrm{Encoder}$ is the SentenceTransformer model capturing contextual and semantic information.

2. **Cosine Similarity**: We calculate the cosine similarity between $\vec{v}_a$ and each vector $\vec{v}_u$ in the set of previously identified unique answers:

   $$\mathrm{sim}(\vec{v}_a, \vec{v}_u) = \frac{\vec{v}_a \cdot \vec{v}_u}{\|\vec{v}_a\| \|\vec{v}_u\|}.$$

   The similarity ranges from -1 to 1, with higher scores indicating stronger semantic resemblance.

3. **Thresholding**: If $\mathrm{sim}(\vec{v}_a, \vec{v}_u) \geq \tau$ (we set $\tau = 0.8$), then a is considered semantically equivalent to an existing unique answer. Otherwise, a is added to the set of unique answers.

This threshold avoids over-clustering or splitting near-identical answers.

D Layer-wise Confidence Measurement

We adopt P(True) [Kadavath et al., 2022] to measure the confidence of each decoding layer of the LLM on its generations. Specifically, we follow [Kadavath et al., 2022] and prompt the LLM layer by layer to judge whether its own generated answer is correct. Our prompt followed the following template:

P(True)

Question: [Question]
Possible Answer: [LLM Answer]

Is the possible answer:
(A) False
(B) True

The possible answer is: