The Dance of Hallucination and Creativity in LLMs’ Decoding Layers via the Lens of Question Answering

Anonymous ACL submission

Abstract

Large language models (LLMs) are known to hallucinate, a phenomenon often linked to creativity. Built upon prior research that focuses on theoretical or qualitative analyses, our work uses a quantitative approach to systematically examine the relationship between hallucination and creativity in LLMs. Given the complex nature of creativity, we take the inspiration from philosophy and propose a creativity definition tailored to LLMs in Question Answering (QA) tasks. Further, we introduce an evaluation framework, *HCL*, to examine the relationship between **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. The optimal layer tends to appear in the early layers of larger models, and the confidence of the model is 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 the applications of LLMs especially in high-stakes scenarios such as finance (Wu et al., 2023) and healthcare (Singhal et al., 2025; Zhou 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; Li et al., 2024) hallucinations.

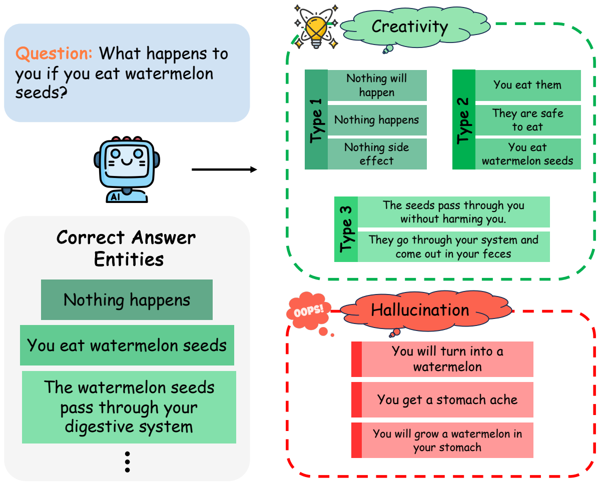
Question: What happens to you if you eat watermelon seeds?
AI
Correct Answer Entities
Nothing happens
You eat watermelon seeds
The watermelon seeds pass through your digestive system
⋮
Creativity
Type 1
Nothing will happen
Nothing happens
Nothing side effect
Type 2
You eat them
They are safe to eat
You eat watermelon seeds
Type 3
The seeds pass through you without harming you.
They go through your system and come out in your feces
OOPS!
Hallucination
You will turn into a watermelon
You get a stomach ache
You will grow a watermelon in your stomach

Figure 1: Illustration of our HCL evaluation criteria. Given a 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 creativity is defined as the diversity represented by distinct types 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. Although promising progress has been achieved, existing studies are still limited in theoretically or qualitatively exploring the relationship between creativity and hallucination, lacking an empirical and systemic study of this connection in LLMs. Simultaneously, current efforts centered on creativity assessments primarily explore on specific tasks such as 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 hallucination 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 take the inspiration from philosophy and psychology and tailor the creativity definition to the LLM QA settings. Specifically, creativity in QA can be quantified as the diversity of correctness among sampled responses for each layer (Figure 1).

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, assisting in locating the optimal decoding layer for different model architectures that tend to produce accurate and varied outputs. Our contributions are listed 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

While previous research has explored hallucination and creativity separately, little attention has been given to their interplay. In this section, we review existing works on hallucination and creativity from definition to application in LLMs, thus 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), RLHF and retrieval-augmented generation (RAG) (Lewis et al., 2020; Kang et al., 2023; Gao et al., 2022) have been proposed to improve factuality. Although these methods improve the model's generalization ability, these methods often lead to over-conservative generation, reducing the model's ability to generate diverse and creative outputs.(Kirk et al., 2023).

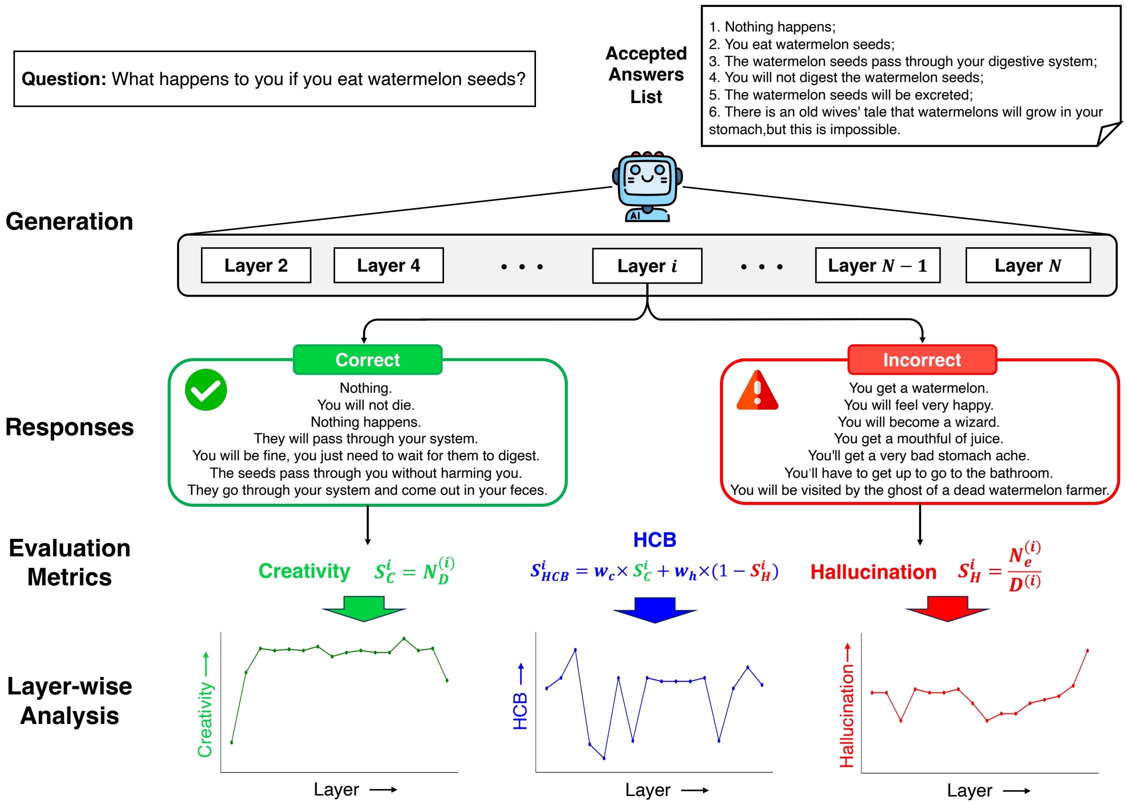
Question: What happens to you if you eat watermelon seeds?
Accepted Answers List
1. Nothing happens;
2. You eat watermelon seeds;
3. The watermelon seeds pass through your digestive system;
4. You will not digest the watermelon seeds;
5. The watermelon seeds will be excreted;
6. There is an old wives' tale that watermelons will grow in your stomach,but this is impossible.
AI
Generation
Layer 2
Layer 4
• • •
Layer i
• • •
Layer $N-1$
Layer N
Responses
Correct
Nothing.
You will not die.
Nothing happens.
They will pass through your system.
You will be fine, you just need to wait for them to digest.
The seeds pass through you without harming you.
They go through your system and come out in your feces.
Incorrect
You get a watermelon.
You will feel very happy.
You will become a wizard.
You get a mouthful of juice.
You'll get a very bad stomach ache.
You'll have to get up to go to the bathroom.
You will be visited by the ghost of a dead watermelon farmer.
Evaluation Metrics
Creativity $S_C^i = N_D^{(i)}$
HCB
$S_{HCB}^i = w_c \times S_C^i + w_h \times (1 - S_H^i)$
Hallucination $S_H^i = \frac{N_e^{(i)}}{D^{(i)}}$
Layer-wise Analysis
Creativity ⟶
Layer ⟶
HCB ⟶
Layer ⟶
Hallucination ⟶
Layer ⟶

Figure 2: Overview of our *HCL* framework. 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** ($\mathbf{S_H}$).

Creativity in Large Language Models The exploration of creativity has deep roots in fields like psychology and philosophy, where it is broadly defined as the capacity to produce outcomes that are both original and valuable (Gaut, 2010; Runco and Jaeger, 2012). Recently, this foundational understanding has been extended to the study of creativity in LLMs (Jiang et al., 2024). In this context, creativity generally refers to LLMs' capabilities to generate *diverse* and *usefulness* content. Existing research primarily focuses on assessing and evaluating creativity in LLMs. 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 outputs are then evaluated by another, often more capable, model based on several criteria, including originality, narrative fluency, flexibility, and refinement. In addition, several studies have investigated the mathematical underpinnings of the trade-off between creativity and hallucination in LLMs. Notably, hallucination has been shown to be an inherent property of these models that, to some extent, facilitates creative generation (Lee, 2023). This insight suggests that current evaluation frameworks, which predominantly emphasize originality and coherence, may underestimate the role of hallucination as a mechanism that contributes to creativity.

Despite the increasing evidence of a trade-off between hallucination and creativity, existing research often treats them as separate phenomena. Most studies prioritize reducing hallucination as an undesirable outcome, while research on creativity rarely examines how hallucination might contribute to the generation of innovative content. As a result, there is an urgent need for a systematic investigation into the relationship between hallucination and creativity in LLMs.

3 Methodology

3.1 A Philosophical Lens of Creativity in QA

The two key elements in the standard definition of creativity in philosophy and psychology are *orig-*

inality and *effectiveness*(Gaut, 2010; Runco and Jaeger, 2012). In the context of open-ended QA, we interpret effectiveness as factual correctness, since answers must be grounded in verifiable world knowledge. Given that the factual content is fixed, the potential for originality lies not in *what is said*, but in *how it is expressed*.

We therefore operationalize *originality* as **diversity** in the QA setting, specifically referring to the ability of a model to generate multiple distinct yet factually accurate answers to the same question. If a model only provides one way of expressing the answer, without exploring alternative valid perspectives, it lacks the originality expected in creative responses. In contrast, generating various factually correct responses that reflect different valid perspectives demonstrates linguistic originality. We formally define the LLM's creativity in QA under this perspective in Section 3.3.

3.2 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:

- **Uncertainty is higher in earlier layers, enabling more diverse outputs.** During the decoding process of LLMs, earlier layers show greater uncertainty, as illustrated in Figure 3. This preserves more possibilities in the inference process and enables them to produce more diverse and creative outputs.

- **The need for early exit.** Deeper layers tend to generate more conservative outputs, while certain intermediate layers may strike an optimal balance between creativity and hallucination (Chuang et al., 2023). By terminating decoding at these intermediate layers, we can not only reduce the computational overhead but also prevent a loss of creativity in generations.

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.

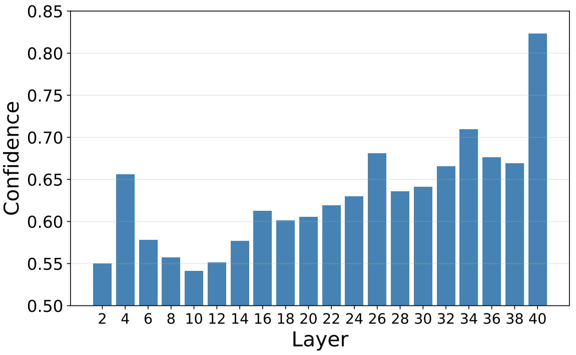

Figure 3: Confidence variations across layers in LLaMA2-13B. We adopt P(True) (see Appendix D) to allow LLM's each layer to self-evaluate the average confidence among the corresponding sampled responses.

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.

3.3 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 the 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 it is deemed hallucination. In this work, we follow the official TriviaQA evaluation protocol

(Joshi et al., 2017), which primarily relies on an Exact Match (EM),a keyword-based match, comparison against the provided ground-truth answers. Based on the above, the hallucination metric of sampled layer-wise responses can be defined as,

$$\mathbf{S_H^i} = \frac{N_e^{(i)}}{D^{(i)}}, \quad \text{where } i \in \{1, \ldots, N\}. \qquad (2)$$

where $N_e^{(i)}$ denotes the incorrect times and $D^{(i)}$ refers to the sampling time at *layer* i.

Creativity. Following the conceptual definition of creativity introduced by previous representative works in philosophical and psychological fields (Gaut, 2010; Runco and Jaeger, 2012) and in recent LLMs domains (Jiang et al., 2024), we define the diversity of correct outputs as the creativity of LLMs' generations in QA tasks. In particular, we filter out incorrect responses from the n responses and group the semantically equivalent (Ribeiro et al., 2018) correct responses. Empirically, 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 semantic-level similarity (see Appendix C for calculation details). As a result, the creativity metric of outputs can be formalized as,

$$\mathbf{S_C^i} = N_D^{(i)}, \quad \text{where } i \in \{1, \ldots, N\}. \qquad (3)$$

where $N_D^{(i)}$ is semantic clusters counts at *layer* i.

While we also considered alternative metrics such as cluster entropy, the current metric was deemed more suitable for our specific research objectives (see Appendix H for a detailed discussion).(Zhao et al., 2023)

Hallucination-Creativity Balanced (HCB). Once we obtain the creativity and hallucination scores for each response, a natural next step is to assess how well different model layers balance these two aspects. Ideally, a strong generation should exhibit high creativity while maintaining factual accuracy. To quantify this trade-off, we introduce the *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. $\mathbf{S_{HCB}^i}$ for the layer $\mathbf{i}$ can be derived as follows,

$$\mathbf{S_{HCB}^i} = w_c \times \mathbf{S_C^i} \; + \; w_h \times \left(1 - \mathbf{S_H^i}\right),$$

where w_c and w_h are the corresponding weights of creativity and hallucination, and $w_c + w_h = 1$. Note that $\mathbf{S_C^i}$ is the normalized score, $\mathbf{S_H^i}$ is the hallucination score, and $\mathbf{S_{HCB}^i}$ is the HCB score.

4 Experiments

In this section, we present the experimental setup, models, datasets, and discuss the key findings. More detailed experiments and further analysis are provided in Appendix B, E, F, and G.

4.1 Experimental Setups

Models We use five popular **open-weight base models** in LLaMA family: LLaMA 3.2-1B, LLaMA 2-7B/13B/70B, and LLaMA 3-8B (Touvron et al., 2023). In all experiments, LLMs are instructed to respond 50 times for each query.

Datasets We utilized two QA datasets: TriviaQA (Joshi et al., 2017) and Natural Questions (NQ) (Kwiatkowski et al., 2019). The detailed description of them are provided in Appendix A.

4.2 When Creativity Meets Hallucination

In this part, we focus on analyzing the creativity and hallucination metrics of LLMs at each layer during response generation. Our experimental results reveal some fundamental relationships between the two dimensions, providing deeper insights into their interplay.

Finding 1. Creativity comes with hallucination. (Jiang et al., 2024)

Across our benchmarks, we observe that creativity generally rises from small to mid-sized models, but plateaus or even dips as scale continues to grow, rather than improving monotonically. For example, very small models (e.g., LLaMA-3.2-1B) tend to be conservative and template-like, yielding low creativity with fewer hallucinations. As we move to mid-sized models (e.g., 7B–14B ranges such as LLaMA-3-8B), creativity increases noticeably; however, further scaling does not guarantee

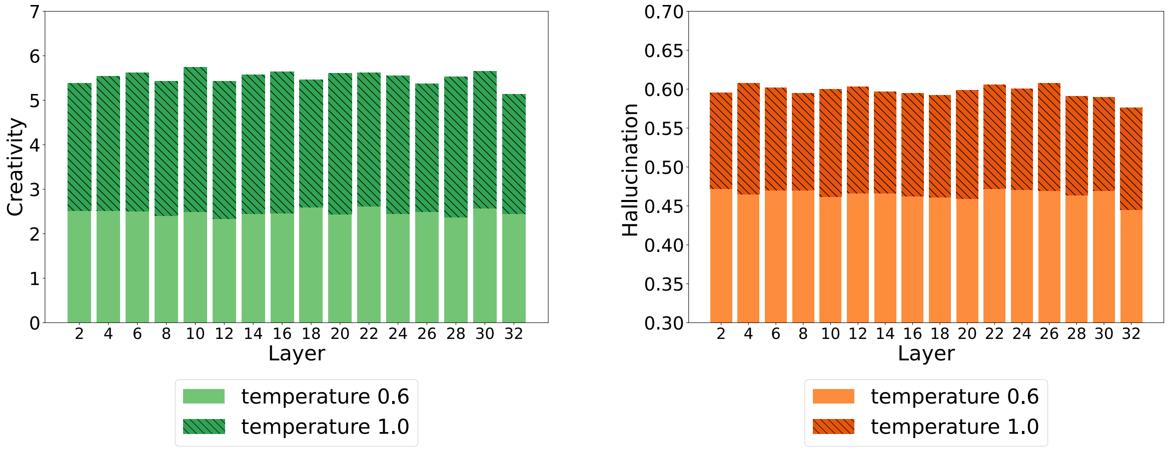

Figure 4: The variation of layer-wise creativity and hallucination metrics of the LLaMA3-8B when its temperature coefficient increases from 0.6 to 1.0 on TriviaQA benchmark.

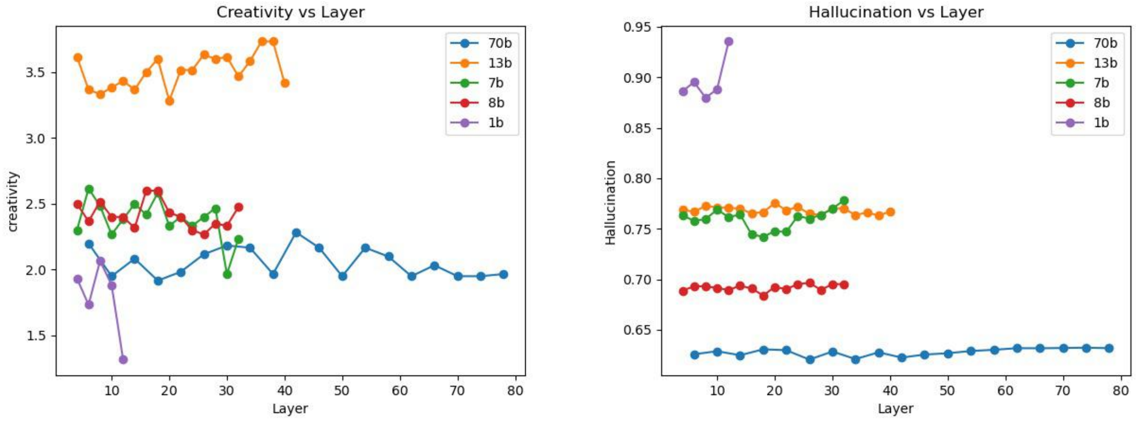

Figure 5: The left figure illustrates the creativity scores across different models, while the right figure presents the hallucination levels for the same models. Both evaluations were conducted with a temperature = 0.6. As observed, the LLaMA 2–13B model attains the highest creativity among all models while maintaining a relatively low level of hallucination. By contrast, although the LLaMA 2–70B model exhibits low hallucination, its creativity is also low.

additional gains (Figure 5). For LLaMA 2–70B, although its hallucination score is the lowest among the four models, its creativity score is also relatively low—only higher than that of LLaMA 2–1B. In other words, creativity peaks around a middle scale on our tasks, indicating a non-monotonic relationship between model size and creative output.

Beyond LLaMA, we also evaluated additional model families. The Qwen series (7B/14B/32B) exhibits the same non-monotonic pattern (Table 1), and further results for GPT-4o-mini, DeepSeek V3, and additional LLaMA variants are reported in Appendix G, where comparable trends are observed.

Finding 2. Stronger models are not always better choices, as creativity often follows a non-monotonic trend with scale.

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 (Figure 5). This suggests an intrinsic trade-off between model capacity and output reliability: as models become more expressive and

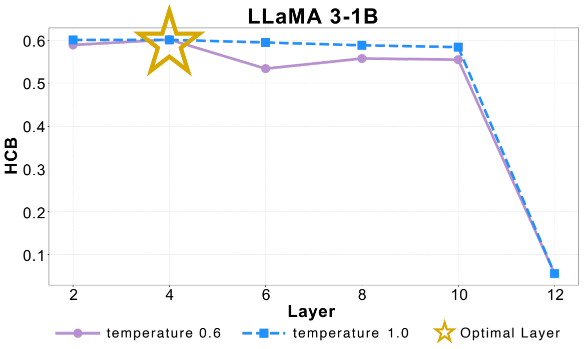

Figure 6: This figure presents the HCB score of the LLaMA3.2-1B. It is evident from the figure that ***layer-4*** consistently achieves the highest HCB score, regardless of the temperature setting.

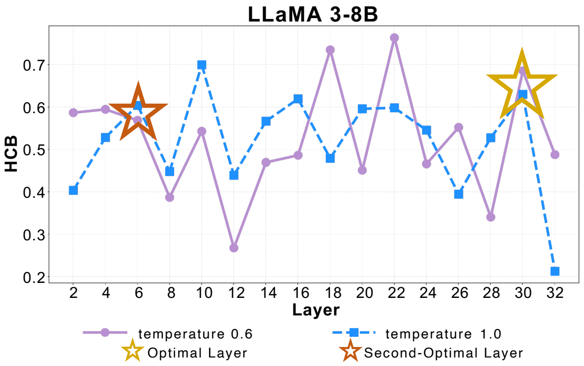

Figure 7: This figure shows the HCB score for LLaMA 3-8B. Although the results indicate ***layer-30*** is the optimal layer, considering the significantly improved efficiency and faster inference speed at lower layers, we further choose ***layer-6*** for early exit. Additionally, since ***layer-8*** achieves the second-highest HCB score, we can still ensure competitive performance at this earlier layer.

generative, they also gain a higher degree of unpredictability, leading to a higher 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. Experiments on Qwen models (7B/14B/32B) also reveal consistent patterns in Table 1. Results for additional models, such as GPT-4o-mini, DeepSeek V3, and the LLaMA family including LLaMA 2-70B, are provided in Appendix G.

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.

Model	TriviaQA		NaturalQ	
	S_C	S_H	S_C	S_H
Qwen-7B	0.67	0.50	0.96	0.54
Qwen-14B	0.87	0.31	1.14	0.44
Qwen-32B	0.83	0.36	1.13	0.44

Table 1: Final-layer creativity (S_C) and hallucination (S_H) scores for Qwen models at temperature = 0.6.

Finding 3. The output from the final layer is not necessarily the best from a creativity-hallucination balanced perspective.

Another key finding from our HCB framework is that final layers, i.e., *layer-12* of LLaMA 3.2-1B, *layer-32* of LLaMA 2-7B, and *layer-40* 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, 7, 8, 10, final layer optimization is not necessarily the best strategy and does not always yield superior performance, particularly in applications that prioritize novelty and diversity over absolute factual correctness. According to the results, 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 in the future.

Finding 4. We identify an intermediate layer remains consistently optimal under varying temperatures and enables efficient decoding.

Interestingly, our analysis reveals that each model has an optimal layer that maintains a stable performance under both temperature 0.6 and 1.0.

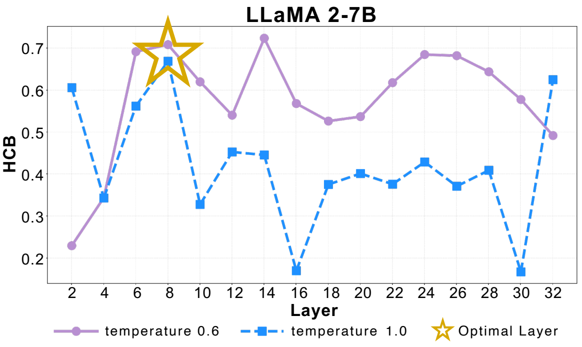

Figure 8: This figure illustrates the HCB score of the LLaMA-7B model across its layers. From the results, we can observe that ***layer-8*** emerges as the optimal layer, whether it is temperature 0.6 or 1.0.

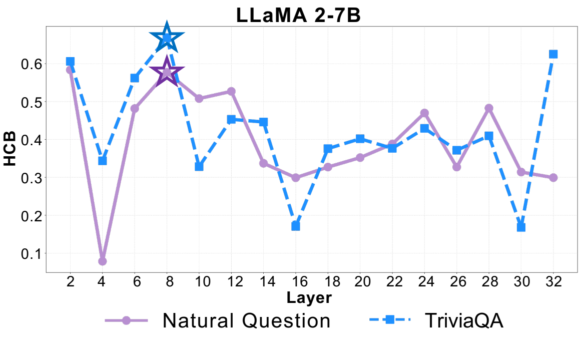

Figure 9: Illustration of the HCB score conducted on LLaMA-7B model at t = 1.0 on TriviaQA and NQ datasets. The results indicate that ***layer-8*** consistently emerges as the optimal layer for balancing creativity and hallucination in LLMs across both datasets.

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 hallucination.

Finding 5. The optimal layer generalizes across QA datasets with confidence peak.

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 further demonstrates that the identified optimal layer is not only specific to a given model but also has broader generalizability across common QA datasets, verifying the robustness of our HCB-based selection.

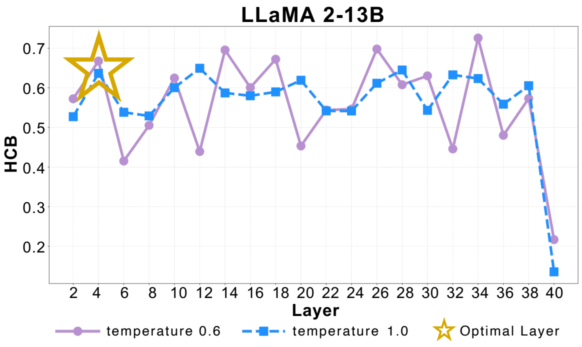

Figure 10: This figure displays the HCB score of the LLaMA-13B model. The results suggest that ***layer-4*** is the optimal layer since it remains nearly optimal when the temperature changes.

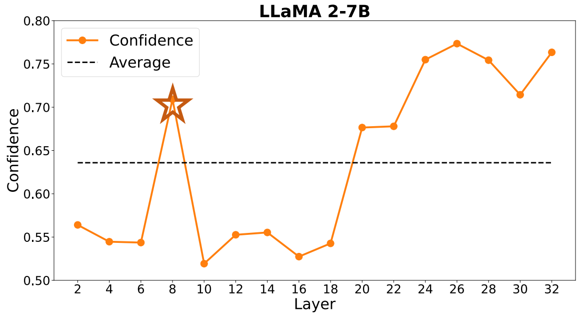

Figure 11: This figure illustrates the variations of confidence across different layers of LLaMA-7B on the TriviaQA dataset. Although the early layers show generally low confidence, there is a sharp peak at ***layer-8***, demonstrating our selection on the optimal layer.

5 Conclusion

This work provides the first systemic study of the relationship between hallucination and creativity in LLMs through the lens of QA. Correspondingly, a hierarchical evaluation framework, HCL, is proposed to explore their interaction across different decoding layers, with the inspiration from philosophy. 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. Additionally, we identify the optimal layer that best balances the tradeoff between hallucination and creativity in LLMs.

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

Our framework is limited to the closed-ended factual 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 such as DebateQA (Neeman et al., 2022)—offer a complementary perspective on response diversity; extending our layer-wise HCL analysis to such settings is left as future work..

The current definition of creativity is limited to QA. Given the complex nature of creativity, its definitions in open-ended tasks like story telling need further investigation. In future work, we will expand the evaluation dimensions of creativity to encompass a broader range of creative expressions.

For model scope, our layer-wise analysis relies on early-exit decoding (e.g., LayerSkip) to obtain consistent outputs from intermediate layers. Adapting them to other model families (e.g., Qwen, GPT, DeepSeek) would require substantial research and engineering effort (including empirical validation of output consistency), which is beyond the scope of this work.

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*.

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*.

Berys Gaut. 2010. The philosophy of creativity. *Philosophy Compass*, 5(12):1034–1046.

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

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*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

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*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. *arXiv preprint arXiv:2211.05655*.

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.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity research journal*, 24(1):92–96.

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*.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023. Knowing what llms do not know: A simple yet effective self-detection method. *arXiv preprint arXiv:2310.17918*.

Yue Zhou, Barbara Di Eugenio, and Lu Cheng. 2025. Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7266–7278.

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:

$$\mathbf{S^i_{HCB}} = w_c \times \mathbf{S^i_C} \ + \ w_h \times \left(1 - \mathbf{S^i_H}\right),$$

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 $\vec{v}_a$:

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

   where 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:

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

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

3. **Thresholding**: If $\text{sim}(\vec{v}_a, \vec{v}_u) \geq \tau$ (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:

E Impact of Weight Parameters on HCB Score Variations Across Layers

To examine the impact of weighting schemes on the final HCB score, we conducted a systematic analysis by varying the weight assigned to creativity (w_c) from 0.3 to 0.7, while correspondingly setting the weight for hallucination as $w_h = 1 - w_c$. For each weighting configuration, we computed the HCB scores and identified the optimal layer for each model.

Our results indicate that while the absolute HCB scores shift with different weighting choices, the relative ranking of candidate layers remains largely stable across a broad range of w_c values. This suggests that the choice of weighting has limited influence on the overall layer selection outcome, thereby supporting the robustness of our findings. The detailed results of this analysis are presented in Tables 2, 3, and 4.

Given this consistency, we adopt an equal weighting scheme ($w_c = w_h = 0.5$) in our main experiments. This neutral setting emphasizes a balanced treatment of creativity and hallucination, aligning with our objective of evaluating models across both axes. We note that in task-specific scenarios—such as medical question answering or legal summarization—users may choose to emphasize hallucination minimization. Our framework readily accommodates such adjustments by allowing the weights to be tuned according to specific application needs.

Layer	$w_c = 0.3$	$w_c = 0.4$	$w_c = 0.6$	$w_c = 0.7$
Layer 2	0.3223	0.2762	0.1842	0.1381
Layer 8	**0.6031**	**0.6557**	**0.7609**	**0.8134**
Layer 16	0.5215	0.5449	0.5917	0.6150
Layer 24	0.5917	0.6383	0.7314	0.7780
Layer 32	0.4754	0.4838	0.5006	0.5090

Table 2: HCB score variations among different layers based on LLaMA 7B (**Temperature = 0.6**) on **TriviaQA** dataset

Layer	$w_c = 0.3$	$w_c = 0.4$	$w_c = 0.6$	$w_c = 0.7$
Layer 2	0.4974	0.5517	0.6604	0.7147
Layer 8	**0.5364**	**0.6026**	**0.7351**	**0.8013**
Layer 16	0.2364	0.2035	0.1377	0.1048
Layer 24	0.3917	0.4105	0.4479	0.4667
Layer 32	0.5051	0.5650	0.6849	0.7448

Table 3: HCB score variations among different layers based on LLaMA 7B (**Temperature = 1.0**) on **TriviaQA** dataset

F Cross-Model Validation on the diversity models

Table 5 reports the creativity (S_C) and hallucination (S_H) scores of several large language models, including GPT-4o-mini, the Qwen family (7B/14B/32B), and DeepSeek-v3, evaluated on both the TriviaQA and Natural Questions datasets.

We observe a consistent trend across models and model sizes: larger models (e.g., Qwen-14B/32B) tend to achieve higher creativity scores compared to their smaller counterparts (e.g., Qwen-7B). For instance, Qwen-32B achieves a creativity score of 0.83 on TriviaQA and 1.13 on Natural Questions, outperforming Qwen-7B (0.67 and 0.96, respectively). At the same time, these larger models often display moderately elevated hallucination scores, indicating a greater risk of generating inaccurate content as their generative capacity increases.

Moreover, similar patterns are observed in both Qwen and non-Qwen models (GPT-4o-mini, DeepSeek-v3), providing cross-model validation for our main findings: the trade-off between creativity and hallucination is not restricted to a single model family, but appears to be a general property of modern large language models. These results highlight the importance of developing evaluation protocols and mitigation strategies that generalize across architectures.

Layer	$w_c = 0.3$	$w_c = 0.4$	$w_c = 0.6$	$w_c = 0.7$
Layer 2	0.4169	0.5002	0.6668	0.7501
Layer 8	**0.4116**	**0.4943**	**0.6597**	**0.7424**
Layer 16	0.2434	0.2712	0.3269	0.3547
Layer 24	0.3453	0.4075	0.5320	0.5942
Layer 32	0.2437	0.2715	0.3270	0.3528

Table 4: HCB score variations among different layers based on LLaMA 7B **(Temperature = 1.0)** on **NQ** dataset

Type	**Model**	TriviaQA		NaturalQ	
		S_C	S_H	S_C	S_H
OS	Qwen-7B	0.67	0.50	0.96	0.54
	Qwen-14B	0.87	0.31	1.14	0.44
	Qwen-32B	0.83	0.36	1.13	0.44
	Qwen-72B	0.94	0.26	0.89	0.40
CS	DeepSeek-v3	0.66	0.47	0.97	0.67
	GPT-4o-mini	1.12	0.24	0.72	0.51

Table 5: Final-layer creativity (S_C) and hallucination (S_H) scores for Qwen models at temperature = 0.6.

G Comparison within the LLaMA Family

Table 5 presents the final-layer creativity (S_C) and hallucination (S_H) scores for the Qwen models (7B, 14B, and 32B) on both the TriviaQA and Natural Questions datasets. The results reveal a general trend where scaling up from Qwen-7B to Qwen-14B leads to substantial improvements in creativity scores and a reduction in hallucination. However, further scaling to Qwen-32B does not result in additional gains; the creativity score of Qwen-32B is similar to, or slightly lower than, that of Qwen-14B (e.g., 1.14 vs. 1.13 on Natural Questions). Hallucination scores also plateau or even increase slightly at the largest scale. This non-monotonic relationship suggests a saturation effect, where simply increasing model size does not guarantee continued improvements in generative diversity and may even result in diminished returns. These findings are consistent with our observations within the LLaMA family, highlighting the nuanced dynamics of scaling large language models and the importance of empirical evaluation rather than relying solely on parameter count.

The results for GPT-4o-mini and DeepSeek-v3 are generally comparable to those of the Qwen series, indicating that the observed patterns are not limited to a single model family.

Series	**Model**	NaturalQ	
		S_C	S_H
Qwen	Qwen-2.5-7B-instruct	0.96	0.54
	wen-2.5-14B-instruct	1.14	0.44
	wen-2.5-32B-instruct	1.13	0.44
	wen-2.5-72B-instruct	0.89	0.40
LLaMA	LLaMA 2-7B	2.23	0.78
	LLaMA 2-13B	3.42	0.77
	LLaMA 2-70B	1.97	0.63

Table 6: Final-layer creativity (S_C) and hallucination (S_H) scores on NaturalQ at temperature = 0.6.

H Discussion on Cluster Entropy and Cluster Count

Regarding cluster entropy, our decision to adopt the raw count of semantic clusters (Cluster Count) is directly motivated by our core definition of creativity within the QA context: the breadth of a model's capability to generate distinct, factually correct answers. A simple count explicitly measures this capability. For example, as shown in Table 7, our Cluster Count metric accurately captures the greater creative capability demonstrated by the model in Case 1 (five distinct answer types) compared to Case 2 (two distinct types). Conversely, an entropy-based metric yields the counterintuitive conclusion that Case 2 is more 'creative,' solely due to its more uniform response distribution. This highlights that entropy-based metrics may penalize non-uniform distributions, potentially obscuring the true breadth of a model's creative capabilities.

Moreover, our metric choice maintains conceptual symmetry with our hallucination metric S_H. Specifically, S_H counts failure events (incorrect responses), while S_C counts distinct success categories (correct response types). This parallelism ensures that the resulting HCB score provides a more interpretable trade-off between two similarly structured concepts.

Question	Response Type	Distribution	Cluster Count	Cluster Entropy
Case1: What was the name of Spike Jones' comedy band?	1. Spike Jones and His City Slickers. 2. 1946 comedy band fronted by Spike Jones, called City Slickers, consisted of 11 pieces of well-practiced precision. 3. 1930's, The City Slickers. 4. The City Slickers. 5. 1942: Spike Jones and his City Slickers.	1. 90% 2. 2.5% 3. 2.5% 4. 2.5% 5. 2.5%	5	0.669
Case2: What occurs when a ray of light meets an obstacle such as a fine wire?	1. diffraction effects. 2. Diffraction (if it's diffraction it must be a very fine wire)	1. 50% 2. 50%	2	1

Table 7: A comparison of Cluster Count and Cluster Entropy using real examples from the TriviaQA dataset, with responses generated by LLaMA2-13B.