Summary of Changes and Revisions

Dear ACs and Reviewers:

We are pleased to resubmit our manuscript entitled "The Dance of Hallucination and Creativity in LLMs' Decoding Layers via the Lens of Question Answering" for consideration in ARR – July 2025.

In this revised version, we have carefully addressed the concerns and suggestions raised during the previous review round. Below, we provide a summary of the major revision.

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

   We have revised 'Finding 2' in Section 4.2 of our manuscript to reflect this more accurate and interesting discovery. The new finding is:
   Finding 2 (Revised): Stronger models are not always better choices, as creativity often follows a non-monotonic trend with scale.
   We believe this revised finding is a stronger contribution, as it challenges the simplistic 'bigger is better' assumption and suggests the existence of an optimal model size 'sweet spot' for balancing creativity and factuality. We will update the text and analysis throughout the paper to reflect this new, more precise conclusion.

2. *Comparison with the cluster entropy.*

   We have added a brief comparison of the raw count of semantic clusters with the cluster entropy in Section 3.3 and Appendix H. Our choice to use the raw count of semantic clusters was directly motivated by our core definition of creativity in the QA context: the breadth of a model's capability to generate distinct, factually correct answers. A simple count directly measures this capability range. For instance, as the examples in Table 7 of the submission illustrate, our Cluster Count metric correctly identifies that the model demonstrated greater creative capability in Case 1 (finding 5 distinct answer types) than in Case 2 (finding only 2). In contrast, an entropy-based metric would lead to the counterintuitive conclusion that Case 2 is more 'creative,' simply because its response distribution is more uniform. This highlights how an entropy-based metric may penalize a non-uniform distribution, which can obscure the true breadth of a model's creative capabilities. Our focus is on the potential to generate variety, making the count metric a more direct measure for our specific research question. Furthermore, this choice maintains a crucial conceptual symmetry with our hallucination metric S_H. S_H is fundamentally a measure based on the count of failure events (incorrect responses), while our S_C is a count of distinct success categories (correct response types). This parallelism makes the resulting HCB score a more direct and interpretable trade-off between two similarly structured concepts.

We sincerely thank the Area Chair and reviewers once again for their insightful feedback, which has significantly improved the clarity, rigor, and scope of our work.

Sincerely yours,
The Authors