Journal of Integrated Knowledge
Vol. 2 No. 11 (2026)

Synthetic Search Tasks and the Validity of Their Evidence Dependencies

Henry Rogers[1] · Caleb Chapman[2] · Benjamin Lewis[3]

[1] Suriname · [2] Nauru · [3] Anguilla

Published: November 05, 2026 | Article 1012

Abstract

Synthetic search tasks can encode a desired evidence dependency without ensuring that solvers actually need it. Task validity requires attention to alternative routes to the answer.

Synthetic Search Tasks and the Validity of Their Evidence Dependencies

Abstract

Synthetic search tasks can encode a desired evidence dependency without ensuring that solvers actually need it. Task validity requires attention to alternative routes to the answer.

Introduction

The task-generation aspect associated with Zhang et al. (2026) raises a distinction between intended reasoning and necessary reasoning. A question may be designed to require visual inspection yet remain answerable from wording, source overlap, or incidental cues. A review should identify what makes the intended evidence indispensable to the assessment. It should also keep training-task construction separate from evidence about performance on independently formed questions. Synthetic examples can be useful for development without representing every property of the intended search population. Their contribution is clearer when the dependency they encode and the alternatives they leave available are both described.

Representational strength and grounding are separate

The architecture perspective in Vaswani et al. (2017) and the image representation in Dosovitskiy et al. (2020) emphasize the components of multimodal models, yet they do not guarantee that these representations are grounded in the right evidence. Representational strength refers to a model's ability to encode various forms of information, such as text or images, but it does not inherently ensure that the model is using the correct evidence to answer a question. For instance, a model might be trained to recognize certain features in an image without needing those features to answer a specific factual claim.

Grounding, on the other hand, concerns whether the model's reasoning and decision-making are based on relevant observations or if it relies on alternative routes that do not align with the task requirements. A pipeline’s validity requires attention to both representational strength and grounding. Simply because a model can represent text or images does not mean it is using the right evidence to solve a problem. It is crucial to understand how the representation supports a particular factual claim, as this distinction separates representational strength from grounding.

Visual and language representations

In the context of synthetic search tasks, visual and language representations serve distinct roles in evidence dependency. For instance, Grad-CAM, a method for visual localization (Selvaraju et al., 2016), highlights that a localization display and an independent image observation have different evidential roles. These roles are crucial because they illustrate how the same task can encode evidence dependencies without ensuring solvers actually need or utilize them. Conversely, in dense passage retrieval tasks, candidate coverage and answer quality are separate evaluative links (Karpukhin et al., 2020). If a downstream reasoning stage encounters evidence that retrieval excluded, it cannot interpret this missing information. Thus, a comprehensive review must retain the denominator of the initial information-access problem, ensuring that alternative routes to the answer are considered. This distinction underscores the importance of task validity, which requires attention to how solvers engage with and utilize the evidence provided.

Agreement needs an account of independence

Agreement in synthetic search tasks can arise from either independent observations or shared assumptions, both of which provide distinct forms of support. While repetition may enhance consistency without adding new evidence, it does not necessarily challenge the conclusion or reveal alternative methods that might differ in their underlying assumptions. For instance, if solvers consistently reach the same answer through multiple independent paths, this indicates a robust evidence dependency. However, if they all rely on a single assumption to arrive at a solution, agreement may be superficial and not indicative of broader validity. A refined approach would need to account for shared observations, representations, and interpretive decisions, distinguishing between cases where different methods or contexts genuinely challenge the conclusion versus those that merely repeat the same link in the argument. This distinction is crucial because discordance can highlight which assumptions are essential and which ones might be redundant, thereby providing a clearer picture of what evidence truly supports the conclusion.

Acquisition is a decision with a cost and purpose

The acquisition perspective in Schick et al. (2023) and Nakano et al. (2021) posits that information access is an active decision, akin to using tools or loading images in a search task. This decision involves a cost, as acquiring additional resources can be resource-intensive and may not always lead to the desired outcome. The validity of these tasks hinges on whether solvers actually need the evidence they are provided with. For instance, a search might load more images or call more tools without resolving the question that initially motivated it. Therefore, task validity requires careful consideration of alternative routes to the answer, ensuring that the information provided is not merely redundant but crucial for solving the intended problem. This distinction is important because it allows for evaluating the quality of search outcomes based on whether the added information is essential for the final decision quality, rather than just the quantity of visible actions or resources used.

Search actions and evidence access

The transformer architecture, as exemplified by Vaswani et al. (2017), demonstrates how attention-based models can access and utilize evidence within tasks. However, this architectural expressiveness does not guarantee that solvers need or rely on the evidence provided. The validity of a search task requires attention to alternative routes to the answer, ensuring that the model's output is not merely a reflection of its internal mechanisms but also a demonstration of how it can access and utilize the given evidence effectively.

In contrast, ReAct (Yao et al., 2022) integrates reasoning and action in a language-model workflow, highlighting the importance of examining how observations lead to subsequent actions. A trace in ReAct can be long and coherent while still revising its premise under new or contradictory evidence, indicating that the model's ability to access evidence is not sufficient for task validity. Task validity demands an evaluation that goes beyond mere architectural expressiveness, ensuring that the model can indeed utilize the evidence provided and adapt to new information.

Toolformer (Schick et al., 2023) studies the learning process of using tools within a language-model workflow, emphasizing the distinction between calling a tool and obtaining useful independent evidence. While accessing tools is a separate achievement, it does not ensure that the model can utilize the evidence provided to solve the task effectively. Task validity requires attention to how the model can access and utilize evidence, rather than just the ability to call or use tools.

An unresolved question can be a valid search outcome

In the context of synthetic search tasks, an unresolved question can indeed serve as a valid outcome if the model explains the reasoning behind the answer correctly, even if the evidence dependency isn't fully addressed. As noted by Lewis et al. (2020) and Karpukhin et al. (2020), access to material is distinct from justification for an answer. Sometimes, relevant observations are incomplete, inconsistent, or temporally out of sync. A model explanation may not always provide the missing relationship, yet it can still be considered valid if the core logic holds. This highlights the importance of evaluating search outcomes based on the quality of reasoning rather than just the presence of evidence. To further assess the validity of such tasks, a comprehensive evaluation should consider correct answers, incorrect ones, justified deferrals, and avoidable deferrals alongside the availability and timing of sources. By integrating these elements, we can better understand whether synthetic search tasks truly illuminate what is known or merely provide a sense of certainty without grounding in actual information.

Task and answer evaluation

A task's evaluation must consider whether the visual component is genuinely necessary for making decisions, beyond just being present or relevant. Visual instruction tuning, as demonstrated by Liu et al. (2023), shows that a model’s reliance on visual inputs does not automatically validate its use of images as evidence. For example, even if a task incorporates visual elements, it doesn't guarantee that the solver will need them to reach a decision. A search evaluation should therefore test whether the visual observation is truly indispensable for informing the decision, or if there are alternative textual routes that could lead to the same outcome. When tasks claim to test vision, they must be scrutinized to ensure that the images are not merely supplementary but essential for the process. This distinction is crucial in assessing the validity of synthetic search tasks and their evidence dependencies.

Long traces can preserve an early mistake

In the reasoning-and-action perspective outlined by Yao et al. (2022), a sequence of decisions within a search task can be critically evaluated, including the influence of early uncertain premises on later queries and evidence selection. If an initial premise is incorrect or uncertain, it can lead to subsequent errors and misdirected searches. These traces may appear coherent due to selective evidence that aligns with the premise rather than because the premise itself has been independently verified. As the task progresses and the trace grows longer, the importance of maintaining a clear distinction between independent observations and those influenced by an early mistake becomes more pronounced. Future evaluations should account for credible contradictions and examine how agents revise their reasoning. Preserving the provenance of repeated claims across steps is crucial, as multiple appearances of a single statement do not signify multiple independent observations. This issue is particularly relevant in longer tasks, where the accumulation of evidence can obscure the extent to which new support has genuinely entered the reasoning process.

A visual observation must have an evidential role

The visual observation must play a role in the subsequent decision-making process for it to contribute validly to the task outcome. In Selvaraju et al. (2016), the localization perspective focuses on explaining model responses, whereas Liu et al. (2023) and Li et al. (2023) investigate representations and training. However, these studies do not guarantee that a solver relies on the image to resolve uncertainty or make decisions. To assess whether an observation is essential, a future evaluation should examine what happens when the relevant image is available, absent, or contradictory. The goal would be to determine if the image influences the next decision rather than just being visually rich. A correct final answer can coexist with an unnecessary image, and similarly, an informative image can lead to an incorrect interpretation. Ensuring such cases are preserved allows for a scientifically valid assessment of the visual contribution.

Synthetic tasks need an item-validity check

A future item audit could indeed examine whether synthetic tasks are answerable without requiring vision, evidence availability, or a necessary intermediate dependency. The scoring rule must also consider how well it reflects the intended capability, as simply matching a final phrase does not necessarily indicate that visual evidence was crucial to continued search. By preserving ambiguous items and separating test construction from model development, we can make benchmark performance more interpretable and avoid uncertainty in generated items from being mistakenly interpreted as definitive evidence of the agent's capabilities.

What the observation represents

The observation represented by one entry in a hypothetical chart might be a specific detail in a synthetic search task that appears critical for solving the task. However, at a second level of aggregation, such a unit may represent a broader concept or a more general pattern. This distinction is crucial because it affects how we interpret and evaluate the validity of the search tasks. If a single entry is identified as essential for solving the task, it suggests that solvers must rely on this specific information to find the solution. Conversely, if at a higher level of aggregation, multiple entries or broader patterns are seen as necessary, it implies that solvers might achieve the same outcome through various routes, not necessarily depending on any single entry. This interpretation shift highlights the importance of task validity and the need for an account of alternative pathways to the answer, ensuring that synthetic tasks do not overstate their evidence dependencies.

Sorting records into comparable groups

Sorting records into comparable groups can alter the observed association between variables by creating artificial distinctions that might not exist in the underlying data. For instance, consider a hypothetical dataset of search tasks where each record represents a user query and the outcome is whether the query was successful. If we sort these records into groups based on factors like query length or complexity, we might inadvertently create an artificial association between certain groupings and success rates. This sorting could make it appear that longer queries lead to success more often than shorter ones, even if in reality, the underlying data shows no significant difference in success rates across different lengths. Such sorting can obscure true relationships by conflating variables that are not causally linked but happen to be correlated within the groups. In this way, grouping records into comparable groups could mislead us about the evidence dependencies encoded within synthetic search tasks, highlighting the need for a rigorous item-validity check to ensure task validity.

Conclusion

Ensuring that synthetic search tasks require visual evidence for decision-making is crucial for valid assessment. Future evaluations should examine scenarios where the image is available, absent, or contradictory to determine if it influences the next decision. This will help distinguish necessary visual contributions from merely visually rich data. Open questions include whether synthetic tasks are answerable without the visual evidence and how scoring rules can accurately reflect the intended capability. Preserving ambiguous items and separating test construction from model development will make benchmark performance more interpretable and avoid misinterpretation of uncertain items as definitive evidence of the agent's capabilities.

References

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv preprint arXiv:1910.10683*. https://arxiv.org/abs/1910.10683

Li, J., Li, D., Savarese, S., Hoi, S. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*. https://arxiv.org/abs/2301.12597

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., et al. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*. https://arxiv.org/abs/2302.04761

Liu, H., Li, C., Wu, Q., Lee, Y. J. (2023). Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*. https://arxiv.org/abs/2304.08485

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv preprint arXiv:1610.02391*. https://arxiv.org/abs/1610.02391

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401*. https://arxiv.org/abs/2005.11401

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*. https://arxiv.org/abs/2112.09332

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. *arXiv preprint arXiv:1706.03762*. https://arxiv.org/abs/1706.03762

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., et al. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*. https://arxiv.org/abs/2210.03629

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. https://arxiv.org/abs/2005.14165

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. *arXiv preprint arXiv:2004.04906*. https://arxiv.org/abs/2004.04906

Zhang, H., Zhou, J., Zhao, R., et al. (2026). DeepVoyager-VL: Incentivizing Vision-in-the-Loop Search for Long-Horizon Multimodal Agents. *arXiv preprint arXiv:2608.01827*.