Studies in Knowledge and Innovation
Vol. 2 No. 10A (2026)

Knowing When a Multimodal Search Question Remains Unanswered

Cole Graham[1] · Julian Crawford[2] · Ryan King[3]

[1] Norfolk Island · [2] Saint Kitts and Nevis · [3] Estonia

Published: October 26, 2026 | Article 1017

Abstract

A search can end without enough evidence to answer its question. Recognizing that state is a separate capability from producing a fluent conclusion after a long trace.

Knowing When a Multimodal Search Question Remains Unanswered

Abstract

A search can end without enough evidence to answer its question. Recognizing that state is a separate capability from producing a fluent conclusion after a long trace.

Introduction

The long-horizon search setting of Zhang et al. (2026) should allow an unresolved endpoint when the available sources fail to establish the requested proposition. A review should distinguish uncertainty caused by missing evidence from uncertainty about interpreting evidence already obtained. The appropriate next step differs between those cases. It should also separate a system's confidence from the actual sufficiency of support. A concise statement of what remains unknown can be more informative than an answer assembled from loosely connected material. The stopping judgment is useful when it describes the evidence boundary clearly, rather than presenting exhaustion of the search budget as proof that the question has been settled.

Acquisition is a decision with a cost and purpose

In acquiring information, there's a delicate balance between the cost and purpose of the decision. The tool-use perspective in Schick et al. (2023) and the browser-assisted perspective in Nakano et al. (2021) illustrate that searching is not merely an ongoing process but a deliberate action. This decision-making process involves choosing among available tools or sources, which can be costly in terms of time and resources. Crucially, this acquisition phase determines what evidence the system will ultimately use to answer the question at hand.

The distinction between relevant retrieval and faithful interpretation becomes paramount here. A search can call upon more tools or images without fully resolving the initial query. This leads to a longer trace as the system might repeat one source or engage in redundant actions, which could be avoided with better decision-making. By accounting for evidence that the system could have but chose not to inspect, we can create a shorter trace that is independently supported and thus more favorable compared to one that repeats sources. The goal is to add information relevant to the intended task rather than merely increasing the number of visible actions, ensuring that each step contributes to final decision quality.

Visual and language representations

In the context of multimodal search questions, visual and language representations are crucial components for comprehensive understanding. The leakage analysis by Kapoor and Narayanan (2023) highlights that when a search ends without enough evidence to answer its question, recognizing the state of information is as important as producing a fluent conclusion. This means that at each decision point, we must evaluate what information was available. A system combining retrieval with generation, such as retrieval-augmented generation (Lewis et al., 2020), separates evidence acquisition from answer production. While this approach ensures that the final answer is generated using relevant material, it does not guarantee that every source was used or used faithfully. Therefore, an evaluation should focus on the relationship between the retrieved information and the stated answer, ensuring that the downstream reasoning stage has access to all pertinent evidence.

Dense passage retrieval (Karpukhin et al., 2020) raises similar concerns about the evaluation of candidate coverage and answer quality. In this process, relevant textual material is crucial for answering a question, but the downstream reasoning stage cannot interpret evidence that retrieval excluded. Thus, an effective review should retain the denominator of the initial information-access problem, ensuring that all potential sources are considered in evaluating the completeness and relevance of the search outcome.

Selection is part of the evidence pathway

In a multimodal search, selection is a critical part of the evidence pathway, influencing which object ultimately becomes the conclusion. When a search ends without enough evidence to answer its question, the selection process itself can be a valuable piece of evidence. The selected object and the decision-making process are distinct research targets. For instance, if a multimodal search concludes with an image that is not fully representative of all data points, the selection pathway-how the image was chosen from potential alternatives-can provide crucial insights into the limitations of the search strategy.

When evaluating such a conclusion, it's important to distinguish between assessing whether the selected object is useful and how reliably the selection process produces useful objects. If evidence for the usefulness of the selected object is presented as evidence for the reliability of the selection process, confusion ensues. By recording the decision pathway, unsuccessful and unresolved cases become meaningful. They illustrate the procedure’s behavior rather than vanish because they did not reach the final demonstration. Thus, understanding the selection process can offer deeper insights into the strengths and weaknesses of a multimodal search method.

Long traces can preserve an early mistake

In the reasoning-and-action perspective outlined by Yao et al. (2022), a search question may conclude with insufficient evidence, but the trace itself can preserve an early mistake if decisions made based on uncertain premises continue to guide subsequent queries and source selection. This means that even if later evidence seems to support the premise, it might be influenced by the initial, unverified assumption. As the horizon of reasoning expands, the early uncertainty in a premise can persistently affect the search outcomes, making it harder to distinguish between independent observations and those that are merely repeated due to the initial premise's influence. This persistence of an early mistake in the trace underscores the importance of recognizing the state of decisions as a separate capability from the final conclusion.

Search actions and evidence access

In examining the few-shot language-model study, the examples provided are crucial components of the information setting, yet they do not constitute a comprehensive answer to the search question (Brown et al., 2020). Similarly, in the text-to-text transfer work, the shared interface facilitates various language tasks but does not equate their evaluation targets. The focus should be on identifying which task property is being assessed and whether fluent output adequately reflects the intended knowledge relation.

For ReAct, its relevance to search lies in understanding how observations are integrated with subsequent actions (Yao et al., 2022). A trace can remain long and coherent while relying on uncertain premises. Therefore, evaluation must scrutinize how these traces revise their conclusions in light of new or contradictory evidence. This process highlights the need for a more nuanced examination of search outcomes beyond just fluent output.

Transport requires a named destination

In distinguishing between evaluation at the destination and adaptation using destination observations, it becomes clear that these serve distinct purposes within a search context. Evaluation at the destination ensures that the result is applicable across different contexts, times, or populations without losing its relevance. For instance, if a study aims to understand the impact of climate change on polar bear populations, evaluating findings in terms of specific population metrics provides a robust basis for generalizing conclusions about broader ecological shifts. On the other hand, adaptation using destination observations focuses on the utility and applicability of evidence within a particular context, allowing for more nuanced conclusions. For example, if researchers discover that a new species of plant thrives in a specific microclimate, they can adapt their findings to suggest potential conservation strategies tailored to that habitat. By making the destination explicit, both goals can coexist without one overriding the other, ensuring the result remains valuable and transferable while also maintaining its relevance within the context of use.

Representational strength and grounding are separate

In the architecture perspective outlined by Vaswani et al. (2017), and the image representation discussed by Dosovitskiy et al. (2020), the components of a multimodal search model are delineated. However, these capacities do not independently ensure that an answer is grounded in the correct observation. A pipeline must account for how the representation supports a particular factual claim.

Sequential interaction and fine-tuning are behavioral descriptions, while reinforcement learning offers a different training approach. An analogy about evidence acquisition should not be used to assume all interactive systems share a reward mechanism or optimizer. This would make their relationship empirical rather than inferable from the appearance of a long reasoning trace. Instead, the distinction between representational strength and grounding highlights that understanding how representations support claims is crucial for assessing the adequacy of evidence in a search context.

Task and answer evaluation

In evaluating a multimodal search task, it is crucial to distinguish whether visual information is essential for reaching a decision or if alternative textual routes could suffice. The evaluation should focus on the dependency between the visual observation and the decision it informs. For instance, in BLIP-2's framework, where pretrained visual and language components are integrated, a key question becomes which channel provides the decisive information. A task that claims to leverage vision should be scrutinized for alternative textual routes that might make the image redundant, thereby questioning its necessity as evidence. This approach ensures that we do not overvalue visual cues and instead assess whether they genuinely contribute to the decision-making process.

An unresolved question can be a valid search outcome

In the context of multimodal search questions, an unresolved question can be a valid outcome, as evidenced by incomplete, inconsistent, or temporally mismatched observations. A model explanation often fails to provide the missing relationship, suggesting that a search might conclude without sufficient evidence to answer its question. Recognizing this, future research should focus on evaluating correct answers, incorrect answers, justified deferrals, and avoidable deferrals alongside resource cost. By retaining the version and availability time of sources, it becomes clear that later corrections can make a retrospective search appear easier. This approach connects uncertainty management directly to the actual information problem, discouraging systems from artificially completing traces for the sake of finishing a search. Consequently, such an assessment would evaluate whether search actions truly change what is known, not just how much material is displayed.

Synthetic tasks need an item-validity check

In synthetic tasks, such as those presented by Brown et al. (2020) and Raffel et al. (2019), there is a risk that a search question may be satisfied through text or prior knowledge, even if the task interface suggests a need for vision. This can lead to a situation where the question remains unanswered, despite having sufficient evidence in other modalities. To address this issue, an item-validity check should be implemented. This check would ensure that the synthetic task genuinely requires visual information and not just the presence of textual or prior knowledge. By preserving ambiguous items rather than forcing them into a clean target set, the validity of the search outcome can be more accurately assessed. The scoring rule must also account for the desired capability, ensuring that simply matching a final phrase does not equate to visual evidence driving continued search. Thus, separating test construction from model development would make benchmark performance more interpretable and prevent uncertainty in generated items from being misinterpreted as definitive judgments about the agent's capabilities.

A visual observation must have an evidential role

The visual observation's role is crucial for a search query that remains unanswered. In Selvaraju et al.'s (2016) localization perspective, an image serves to explain model responses, which is distinct from the multimodal perspectives presented by Liu et al. (2023) and Li et al. (2023), where images facilitate understanding through representations and training. However, these perspectives alone do not guarantee that a subsequent decision will be influenced by the image. To assess this dependency, future research should focus on evaluating whether the availability or absence of a relevant image impacts the next decision in the search process. This evaluation would examine changes in the evidence pathway when an image is present, absent, or contradictory to the query. The objective here is to understand evidence use rather than preferences for visually rich traces, ensuring that both correct and incorrect final answers can coexist alongside unnecessary images. Preserving these cases enhances the scientific assessment of the visual contribution's necessity in resolving search uncertainties.

What the observation represents

In the hypothetical chart, each entry represents a specific instance of a multimodal search question. At the first level of aggregation, each entry captures a discrete observation or piece of evidence that does not provide enough information to fully answer the question. This unit is valuable as it marks the point where the search must either continue or terminate. At a second level of aggregation, these entries aggregate into a larger pattern or trend, representing the state of the search at that moment. Here, the interpretation shifts from focusing on individual observations to understanding the broader context and progress of the search. This change in perspective allows for a more nuanced evaluation of the search's effectiveness and progress, rather than merely assessing each step individually.

Sorting records into comparable groups

Sorting records into comparable groups can alter the appearance of an observed association, making it seem stronger or weaker than it actually is. For instance, imagine a scenario where researchers are searching for patterns in data related to climate change impacts. If they group records by geographical region, the trend might appear more pronounced due to varying regional factors. Conversely, grouping by time period could reveal distinct seasonal patterns that weren't evident when examining data chronologically. Thus, the way records are sorted can influence our perception of the strength and significance of an observed association, highlighting the importance of careful groupings in search outcomes.

Conclusion

One consequence of implementing an item-validity check is that it prevents ambiguous items from being incorrectly marked as answered, thus improving the accuracy of the search outcome. An open question is how the validity of the search outcome is influenced by the specific context or conditions under which the search is conducted, such as the complexity of the task or the agent's prior knowledge.

References

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

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

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

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

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

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

Kapoor, S., Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4(9), 100804*. https://doi.org/10.1016/j.patter.2023.100804

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

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

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

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

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*.