**Annals of Science and Inquiry**
Vol. 2 No. 11 (2026)

# Candidate Retrieval and the Invisible Denominator of Alignment Evaluation

Dominic Jenkins[1] · Adam Tucker[2] · Paul Lewis[3]

[1] Algeria · [2] Bonaire, Sint Eustatius and Saba · [3] Palau

Published: November 02, 2026 | Article 953

## Abstract

Alignment accuracy can be conditioned on an invisible retrieval denominator. Candidate-level success should be read together with the entities excluded before scoring begins.

# Candidate Retrieval and the Invisible Denominator of Alignment Evaluation

## Abstract

Alignment accuracy can be conditioned on an invisible retrieval denominator. Candidate-level success should be read together with the entities excluded before scoring begins.

## Introduction

The pipeline discussed by Wu et al. (2026) can be evaluated on more than one population. Scoring only retrieved candidates describes a conditional decision problem, whereas evaluating all eligible entities includes retrieval failures. Neither denominator is inherently wrong, but they support different claims. A review should make the eligibility and filtering steps visible before interpreting a reported accuracy value. It should also retain unresolved entities rather than silently remove them from the alignment question. The central issue is comparability: a system with a narrow candidate pool and a system covering a broader population cannot be ranked fairly by treating their conditional summaries as the same measure.

## Confidence can grow through repeated inference

The repeated inference perspective raised by Sun et al. (2018) highlights that confidence grows when an inferred match is used as evidence for subsequent inferences. However, this does not equate to a newly observed correspondence. Rather, it suggests the uncertainty of the initial premise remains unaccounted for. As the pipeline relies on these uncertain starting points, it can become increasingly confident, yet the reliability of each step is not fully captured. To address this, future evaluations should scrutinize the premises and ensure that downstream conclusions are coherent, rather than averaging away contradictions as if all inputs were independent. This approach would assess the structure of evidential support, not merely the quantity of sources or stages involved.

## Knowledge-graph representation evidence

In the context of candidate retrieval and alignment evaluation, the leakage analysis by Kapoor and Narayanan (2023) highlights that a significant portion of information can inadvertently cross over from model development to evaluation, potentially compromising the integrity of the assessment process. This leakage is often overlooked because it operates behind the scenes, as an "invisible denominator" that influences the outcome without explicit consideration. To address this issue, it is crucial to scrutinize the entities excluded at each stage of candidate selection and scoring, ensuring that the decision-making process remains independent of the data used for model development.

The distinction between explaining and predicting, as articulated by Shmueli (2010), underscores that a model's predictive success does not automatically imply a clear causal or substantive explanation. When evaluating alignment accuracy, it is imperative to align the intended inferential goal with the evaluation process. This ensures that any assessment remains independent of the selection criteria and focuses on genuine predictive performance rather than post-hoc justifications for observed patterns.

In the knowledge-graph survey by Ji et al. (2020), representation, acquisition, and application are recognized as distinct stages in the development of alignment or generation systems. Each stage introduces its own set of uncertainties, which must be carefully managed to maintain robust and reliable evaluation metrics. By distinguishing these stages, researchers can better understand the sources of error and uncertainty in their models, thereby improving the overall quality and reliability of their findings.

## Agreement needs an account of independence

Agreement can arise from two distinct sources: independent observations or shared assumptions. While both contribute to a sense of consistency, they offer different types of support. For example, repeated observations can enhance reliability without introducing new evidence. In contrast, shared assumptions imply that multiple descriptions are inheriting the same underlying principle. This distinction is crucial for understanding how agreement is established and whether it genuinely challenges existing conclusions. Discordance becomes informative under this account because it highlights which assumption did not persist, thereby revealing gaps in the argument. The goal is not to dismiss all agreements, but rather to identify precisely what an agreement establishes and ensure that broader conclusions still require independent verification.

## Retrieval determines which identities remain possible

Karpukhin et al. (2020) and Lewis et al. (2020) highlight how retrieval determines which identities remain possible before scoring begins. The invisible denominator of alignment evaluation is a critical factor that should be visible throughout the process. If relevant evidence is excluded prior to scoring, it cannot be recovered solely through interpretation later on. This means that candidate-level success must be analyzed in conjunction with the entities excluded at the outset.

While both Karpukhin et al. and Lewis et al. offer valuable insights, their methodologies differ significantly. The uncertainty and error meanings associated with these approaches are distinct, requiring careful evaluation. Efficiency claims should not be isolated but rather considered within the broader context of cost versus identity quality. Incorporating this visibility into alignment evaluations would provide a more comprehensive understanding of resource savings and their implications on candidate quality.

## Candidate access and identity support

RotatE models relational structure using embeddings, highlighting how representational assumptions influence downstream tasks (Sun et al., 2019). The key issue here is the invisible retrieval denominator-how many entities are initially available for consideration before candidate-level success is assessed. If only a subset of these entities is considered, it affects alignment accuracy. For instance, consider a scenario where a model focuses on matching records that might or might not represent the same entity. If the model overlooks records that could be identical but are excluded from the initial pool due to data constraints or model design, this can lead to an inaccurate assessment of alignment success. The retrieval process determines which identities remain possible, and this exclusion is crucial for understanding how well a model performs in terms of identifying correct matches. In other words, the entities not included in the candidate set should be accounted for when evaluating alignment accuracy, as they could have provided valuable context that influenced the model's performance.

## Temporal evaluation reconstructs what was knowable

In Kapoor and Narayanan's (2023) work, the leakage perspective highlights a critical issue: when retrospective evaluations are considered, the alignment accuracy of a system can be significantly influenced by later knowledge not present at the time of initial evaluation. This temporal reconstruction can make earlier tasks seem easier than they were, leading to an overestimation of performance. To address this, future assessments should explicitly distinguish between correctness under the original information boundary and competence under later conditions. By preserving versions of the graph, retrieval collection, and candidate output, we can observe how later observations influence adaptation. This approach helps in separating temporal generalization from a revised system that has access to new evidence. It also ensures that the alignment claim is more narrow, reproducible, and informative for downstream use, avoiding automatic transfer of probabilities or similarities across different environments.

## Semantic models and evidence boundaries

In the context of semantic models and evidence boundaries, retrieval determines which identities remain possible. The passage emphasizes that candidate-level success must be considered alongside the entities excluded before scoring begins. This is critical because downstream reasoning stages cannot interpret evidence that retrieval has already excluded. Therefore, it's essential to retain the denominator of the initial information-access problem. For instance, a few-shot language-model study examines adaptation through examples in a prompt (Brown et al., 2020), where the coverage of these examples forms part of the evaluated information setting. Similarly, a transformer model's attention-based sequence modeling architecture (Vaswani et al., 2017) is distinct from its grounding, answerability, and factual verification capabilities. In an information task, the evidence claim must focus on what the model was given and what its output supports, not just the model's architectural capacity. Thus, maintaining the retrieval denominator ensures that the full scope of available evidence is considered, providing a comprehensive evaluation of candidate success.

## Identity and applicability are separate claims

An entity correspondence assesses whether two records refer to the same object, but it does not guarantee that every associated fact applies simultaneously or in the same context. As highlighted by Ji et al. (2020), distinguishing between representation and application is crucial. The knowledge-graph perspective suggests that a final resolved name might obscure earlier uncertainties within the pipeline. This distinction is vital because a correctly matched entity can carry outdated facts, which have different implications for downstream reasoning.

Structural similarity can identify plausible candidates without independently confirming semantic identity, thereby distinguishing cases of undifferentiated errors. A future evaluation should trace decisions through identity, temporal scope, and the applicability of evidence, clarifying why an alignment result is useful rather than merely reporting a successful match.

## A representation is not independent semantic verification

The embedding perspective in Sun et al. (2019) and relational graph learning in Schlichtkrull et al. (2017) focus on encoding relationships within a graph for a specific task, but their scores remain contingent upon the assumptions inherent in the represented graph. This means that the embedding or graph representation itself is not an independent semantic verification of real-world identity. It reflects only the data and structure available to the model. When considering candidate-level success, it is crucial to acknowledge which identities are excluded before scoring begins. To properly evaluate candidate access and identity support, one must account for the missingness and incompatibilities within the graph representation. This includes examining how neighborhood information influences the candidate's inclusion or exclusion, thereby preserving the incomplete input graph and its inherent limitations. By doing so, a more comprehensive understanding of the model’s alignment accuracy can be achieved, distinguishing it from mere structural resemblance.

## Language-model reasoning has an information boundary

The few-shot perspective in Brown et al. (2020) and the transformer architecture in Vaswani et al. (2017) focus on model capability within the supplied information, suggesting that a fluent semantic judgment can be valuable without introducing an independent fact. However, this does not imply that the model's final matching is sufficient to validate its informative nature. In other words, a disagreement between the candidate and premise does not necessarily indicate that one component is inherently unreliable. Rather, it highlights the need for careful consideration of which evidence was used at each stage. The key scientific question remains: how does the evidence at a given stage influence the identity decision, and whether this evidence improves the final outcome when defined by a clear temporal and source population?

## What the observation represents

In a hypothetical chart representing alignment accuracy, one entry signifies the success of a specific candidate retrieval. At a granular level, this observation captures the success rate of a single candidate against excluded entities before scoring begins. However, at a second level of aggregation, this unit shifts to represent the overall effectiveness of a retrieval system, encompassing not just the successes but also the identities that were excluded from consideration. This distinction highlights that alignment accuracy is contingent on the invisible retrieval denominator, which includes all potential candidates and excludes those deemed irrelevant or incorrect. Thus, candidate-level success must be understood in relation to the broader context of entities not included in the evaluation process.

## Sorting records into comparable groups

Sorting records into comparable groups can alter the observed association between retrieval and alignment accuracy. For instance, if we categorize candidates based on their relevance to a query, those deemed less relevant might be excluded from consideration before scoring begins. This exclusion process could skew our perception of alignment success. If we only score candidates that are highly relevant, the apparent alignment success will appear higher than it actually is when compared to a scenario where all candidates, regardless of relevance, are included in the evaluation. In this way, the inclusion or grouping of records can introduce an invisible denominator, affecting how we perceive the alignment accuracy and its relationship with retrieval.

## Temporal order and the record

To compare a cross-sectional record with a dated series of records, consider the temporal order they provide. In a cross-sectional record, we see a snapshot at a single point in time, capturing all entities and their alignment status as they exist then. By contrast, a dated series offers a more dynamic view, detailing changes over time. Each record in this series reflects an alignment status at a specific point in the timeline, allowing us to track shifts and variations in alignment accuracy. The cross-sectional record, while useful for understanding static conditions, lacks the temporal context that informs how entities were aligned at different points in time. Consequently, conclusions derived from these two types of records can diverge significantly, as they support different narratives about alignment dynamics and the invisible retrieval denominator.

## Separating changes within and between groups

In the context of candidate retrieval, let's consider an abstract pattern where group-level changes diverge from the individual records within those groups. Imagine a scenario where we analyze the alignment accuracy of a set of documents, each representing different entities. When we compare these groups as a whole, we might observe a higher overall success rate than what can be attributed to chance alone. This suggests that some entities are consistently retrieved or excluded across the group, creating an invisible denominator that influences alignment accuracy. However, if we look at individual records within these groups, we might find a more nuanced picture. Some records may consistently align with others, while others remain consistently excluded, indicating that certain entities are systematically favored or disfavored in retrieval. This pattern highlights the need to consider not just the success of individual candidates but also the entities that are systematically left out before scoring begins. In other words, the success of each record within a group is influenced by the entities that were not included in the evaluation process, creating an invisible denominator that affects overall alignment accuracy.

## Conclusion

One consequence of this invisible retrieval denominator is that it can lead to an overestimation of candidate retrieval success, as only relevant candidates are included in the evaluation. This can result in a perception of higher alignment accuracy than what would be observed if all potentially relevant candidates were considered. Another open question is how varying criteria for relevance or quality of candidates impact the alignment accuracy, as different groups of excluded entities can significantly alter the reported success rate of retrieval systems.

## References


Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. v. d., Titov, I., Welling, M. (2017). Modeling Relational Data with Graph Convolutional Networks. *arXiv preprint arXiv:1703.06103*. https://arxiv.org/abs/1703.06103

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

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

Kipf, T. N., Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907*. https://arxiv.org/abs/1609.02907

Sun, Z., Hu, W., Zhang, Q., Qu, Y. (2018). Bootstrapping Entity Alignment with Knowledge Graph Embedding. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 4396-4402*. https://doi.org/10.24963/ijcai.2018/611

Sun, Z., Deng, Z., Nie, J., Tang, J. (2019). RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. *arXiv preprint arXiv:1902.10197*. https://arxiv.org/abs/1902.10197

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

Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P. S. (2020). A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *arXiv preprint arXiv:2002.00388*. https://arxiv.org/abs/2002.00388

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

Wu, W., Xu, K., Zhao, R., et al. (2026). Adaptive retrieval-augmented evidence-fusion for highly heterogeneous temporal entity alignment. *Information Fusion*. doi:10.1016/j.inffus.2026.104516.