# FONDUEQA: A BENCHMARK FOR EVALUATING PARAMETRIC COMPENSATION IN MULTI-HOP QA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) show remarkable performance based on parametric knowledge, i.e., their ability to recall facts and associations internalized in a model's weights. However, studying this capability on established multi-hop benchmarks may be complicated by benchmark contamination, as many of these benchmarks have been publicly available long enough to plausibly appear in modern pretraining corpora. As a result, correct answers under missing evidence may reflect either general parametric knowledge or memorization of benchmark-specific instances. To study how models use internal knowledge when retrieval evidence is incomplete, we introduce **FONDUEQA**, in which every required multi-hop fact is uniquely localized to a designated evidence document. This enables controlled document-level evidence ablations to study parametric compensation under incomplete retrieval. Our empirical evaluation reveals that parametric compensation strongly depends on both entity popularity and the structural role of missing evidence. Generally, models compensate more successfully for facts involving popular entities. On **FONDUEQA**, we show that bridging evidence can be more useful than the answer-bearing document itself. Moreover, models may recall a missing fact in isolation but fail to use it under partial evidence, particularly when the context contains plausible competing entities, which can reduce Exact Match by up to 15 percentage points.

## 1 INTRODUCTION

Large Language Models (LLMs) are pre-trained on a large corpus of factual knowledge that is stored within their weights, referred to as parametric knowledge (Zhao et al., 2026). Whenever user queries and requests go beyond the internalized knowledge, LLMs are increasingly paired with retrieval systems that provide external evidence at inference time, improving factual accuracy (Ni et al., 2026; Gao et al., 2024; Arslan et al., 2024). Entity popularity is associated with differences in factual QA performance and the gains obtained from retrieval augmentation (Yang et al., 2024). Complementary theoretical work models the benefits and risks of retrieved information at the token level (Xu et al., 2025). At the same time, LLMs can sometimes answer correctly even when the retrieved context is insufficient to support the answer (Joren et al., 2025). However, in multi-hop QA, the utility of a retrieved passage depends on its relationship to the other passages in the context (Jain & Garimella, 2025).

Prior work studies this interaction for single-fact questions, showing that models recall popular facts more reliably and benefit more from retrieval for less popular information (Mallen et al., 2023; Yang et al., 2024). Multi-hop questions add structure to this interaction. Consider the query *"What is the nationality of the director of the film Titanic?"*. Answering it requires knowing the interdependent facts that *Titanic* was directed by *James Cameron*, and that *James Cameron* is *Canadian*. With only part of this evidence observed, the question becomes whether the model can recover the missing link, either from *Titanic* to *James Cameron* or



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from *James Cameron* to his nationality. How evidence structure affects compensation, and which other factors determine success under partial retrieval, remain largely unexplored.

We argue that existing multi-hop benchmarks provide important components for this analysis, but were not designed for a controlled removal of individual reasoning steps (Table 1). While POPQA(Mallen et al., 2023) provides entity-popularity measures and MINTQA (He et al., 2026) extends this to multi-hop, they do not provide context documents. Further, 2WIKIMULTIHOPQA and MUSIQUE-ANS (Ho et al., 2020; Trivedi et al., 2022) move towards stronger multi-hop supervision, but focus on reasoning paths or connected reasoning rather than explicit fact isolation required for controlled fact removal. Consequently, we observe (i) *Question leakage*, where the question reveals (parts of) the answer; (ii) *Cross-document leakage*, where another document reveals a questioned fact; (iii) *Distractor leakage*, where an allegedly irrelevant document contains required evidence.

To enable a controllable investigation over facts, we propose a novel multi-hop QA benchmark, called FONDUEQA. In a first step, we leverage a knowledge graph as the source of candidate facts. In a second step, we construct template-based multi-hop questions from these facts. In the third step, our pipeline ensures that each fact is assigned to exclusively one gold document, which we refer to as *Fact Isolation*. Subsequently, we conduct a thorough validation and filtering step, where it is ensured that the required facts are supported within their assigned gold document without any cross leakage of required evidence. Rather than altering the original documents, we discard invalid instances unless the generated question-answer pair can be adjusted without changing the intended evidence structure. As a result, each document-level ablation induces a controlled removal of its assigned fact, enabling direct comparisons across partial evidence conditions.

In our empirical study, we show that parametric compensation under partial evidence depends not only on the accessibility of missing facts, but also on their structural role in the reasoning path. On FONDUEQA, bridging evidence (e.g. 'James Cameron directed Titanic') can be more valuable than the answer-bearing document itself (e.g. 'James Cameron is Canadian'), while facts that are successfully recalled in isolation may still fail to contribute to the answer. Compensation is further shaped by the surrounding context, as plausible competing entities can reduce *Exact Match* by up to 15 percentage points when supporting alternative answers, but may improve it by more than 25 points when supporting the correct answer.

We make three main contributions:

- **Controlled benchmark construction.** We introduce a benchmark construction pipeline that enforces fact isolation and controls for evidence leakage. Using this pipeline, we construct FONDUEQA, covering different question types, evidence structures, answer formats, and role-specific popularity information. We also adapt established multi-hop benchmarks for the same controlled analysis.
- **Determinants of parametric compensation.** We show that compensation for missing evidence depends on entity popularity and evidence role. On FONDUEQA, bridge evidence can be more useful than the document containing the target although the utility of evidence roles varies across benchmarks.
- **Limits of parametric compensation.** We show that recalling a missing fact in isolation does not guarantee its use under partial evidence, and that plausible competing entities can substantially change whether parametric knowledge is successfully integrated into the reasoning path.

## 2 RELATED WORK

**Benchmarks.** The original HOTPOTQA (Yang et al., 2018) introduced Wikipedia-based QA pairs to benchmark multi-document reasoning. 2WIKIMULTIHOPQA (Ho et al., 2020) provides a complete reasoning path for the multi-hop questions aimed at measuring the true reasoning skill of models rather than relying on shortcuts. POPQA (Mallen et al., 2023) studies single factual question answering across entities with varying popularity, highlighting the difficulty of recalling unpopular knowledge from model parameters. MINTQA

Table 1: **Comparison of FondueQA to existing QA benchmarks.**

| Benchmark | Live Bench | Answer Aliases | Popularity | Context Documents | Subquestions | Evidence Control | |
|---|---|---|---|---|---|---|---|
| | | | | | | Distractor Checks | Fact Isolation |
| HotpotQA | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ |
| MINTQA | ✗ | ✗ | ✓ | ✗ | ✓ | – | – |
| 2WikiMultiHopQA | ✗ | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ |
| PopQA | ✗ | ✓ | ✓ | ✗ | – | – | – |
| MuSiQue-Ans | ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| **FondueQA** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

(He et al., 2026) extends this perspective to multi-hop questions with varying levels of knowledge freshness and entity popularity. MUSIQUE-ANS (Trivedi et al., 2022) was designed to make it harder to answer through disconnected reasoning. MUSIQUE-FULL (Trivedi et al., 2022) also constructs unanswerable variants by removing support for a selected subquestion. Our construction instead targets a unique fact-to-document assignment for controlled, role-specific evidence ablations. Taken together, existing benchmarks provide complementary components for our analysis (Table 1). POPQA and MINTQA provide entity popularity information, while several multi-hop benchmarks provide context documents or explicit subquestions. Some benchmarks additionally provide answer aliases for a more robust evaluation. FONDUEQA further follows a versioned live-benchmark setup in which new instances will be released over time.

**Parametric compensation.** Prior work investigates the effect of entity popularity and parametric compensation with a focus on improving retrieval augmentation, however they are constrained to single-hop questions (Mallen et al., 2023). Joren et al. (2025) demonstrate that LLMs can answer correctly even when the retrieved context is insufficient, highlighting that external evidence might not need to contain all information required for an answer. Other studies further investigate how models balance parametric and contextual knowledge in retrieval-augmented generation (Farahani & Johansson, 2024; Bi et al., 2026). Related work on answerability shows that recognizing unanswerable questions becomes harder when the context contains plausible same-type candidates (Sulem et al., 2021). Whether such competition affects parametric compensation under controlled multi-hop evidence gaps remains unexplored.

**Evaluation Bias.** Benchmark contamination occurs when evaluation items or information derived from them are included in model training, potentially inflating measured performance (Xu et al., 2024). Dong et al. (2024) inspect potential contamination from the concentration of sampled model outputs. Deng et al. (2024) show that commercial LLMs can reconstruct masked incorrect answer options and unlikely words from benchmark instances, providing evidence of potential benchmark contamination. Earlier studies show that high performance on multi-hop QA benchmarks can arise from dataset artifacts and disconnected reasoning rather than the intended composition of supporting facts (Guo et al., 2023; Chen & Durrett, 2019; Trivedi et al., 2020). More recently, Fang et al. (2025) identify behavior consistent with knowledge leakage through context-removal and question-perturbation tests.

## 3 BENCHMARK CONSTRUCTION PIPELINE

We construct FONDUEQA to support controlled removal of individual facts from multi-hop evidence. Our pipeline generates multi-hop questions from knowledge graph triples and assigns each required fact to one document. It then verifies the fact is absent from the question and all other documents, including distractors. Invalid instances are repaired or discarded. Thus, removing a gold document removes its assigned fact.

Our construction supports two interventions. An *evidence gap* removes one or more documents and, consequently, the required facts from the context. A *reasoning trap* keeps this gap fixed but adds plausible competing entities to the context. We formalize both interventions below and describe the resulting benchmark.

### 3.1 Formal Setting

Let $\mathcal{KG} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$ be a *Knowledge Graph* with entities $\mathcal{E}$, relations $\mathcal{R}$, and facts (triples) $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$. We use Wikidata as $\mathcal{KG}$ and link each entity $e$ to its Wikipedia article $\mathrm{doc}(e)$, which serves as the document that expresses facts about that entity. We assume a closed world, i.e., a triple that is not in $\mathcal{KG}$ is treated as unknown. We consider a fact $f = (s, r, o)$ is expressed by the associated document of its subject, whereas the support of $f$ in the object's document must not hold (cf. Section 3.2).

A question is constructed from a set of required facts $F = \{f_1, \ldots, f_k\} \subseteq \mathcal{T}$. The relations among these facts define the question's evidence structure. We refer to the induced subgraph of a selection of required facts as *Evidence Structure*. Given a specific evidence structure we define a reasoning path leveraging required facts:

**Definition 1** (Reasoning Path)**.** *A reasoning path from $e_0$ to $e_n$ is a sequence of facts $\pi = (f_1, \ldots, f_n)$ with $f_i = (e_{i-1}, r_i, e_i) \in \mathcal{T}$ and two consecutive facts are linked by a commonly shared entity.*

A QA instance is defined as $x = (q, D, A, F, \Gamma)$: a question $q$, a set of context documents $D := D^+ \cup D^-$ (gold and distractor documents, $D^+ \cap D^- = \emptyset$), a set $A$ of valid answer forms (canonical answer and aliases), the set $F = \{f_1, \ldots, f_k\} \subseteq \mathcal{T}$ of facts required to answer $q$, and an assignment $\Gamma : F \to D^+$ with $\Gamma(f) = \mathrm{doc}(s)$ of each required fact $f = (s, r, o)$ to its gold document. For a text $t$ and a fact $f$, let $\mathrm{supp}(t, f) \in \{0, 1\}$ indicate whether $f$ can be derived from $t$. Removing one or more required facts from the context results in an *Evidence Gap* defined as:

**Definition 2** (Evidence Gap)**.** *Let $G \subseteq F$ be a non-empty set of required facts. The evidence gap induced by $G$ removes the gold document $\Gamma(f)$ for every $f \in G$, yielding the context $D_{-G} = D \setminus \{\Gamma(f) \mid f \in G\}$. The facts in $G$ are missing and the facts in $F \setminus G$ are retained. The gap is* partial *if $G \neq F$ and* complete *if $G = F$.*

For this intervention to be controlled, each removed fact must appear only in its assigned document. We state the corresponding requirements in Section 3.2. Furthermore, replacing distractors with plausible competitors leads to a *Reasoning Trap*:

**Definition 3** (Reasoning Trap)**.** *Let $f^- = (s, r, o) \in G$ be a missing fact of an evidence gap $G \subseteq F$. A* competitor *for $f^-$ is a distractor $d \in D^-$ about an alternative entity $o' \neq o$ that is admissible for the same reasoning role and supports no required fact in $F$, i.e., $\mathrm{supp}(d, f) = 0$ for all $f \in F$. A* reasoning trap *replaces distractor documents in $D_{-G}$ with competitors while keeping the missing facts and context size fixed.*

For example, consider the question *"What is the nationality of the director of* Titanic*?"*. Its reasoning path consists of the bridge fact $f_1$=(*Titanic*, director, James Cameron) and the target fact $f_2$=(James Cameron, nationality, Canada). Removing the bridge fact corresponds to the partial evidence gap $G = \{f_1\}$. The target evidence $f_2$ remains, while the document identifying James Cameron as the director of *Titanic* is removed. Steven Spielberg or Ridley Scott serve as alternative directors while preserving the gap.

### 3.2 Requirements for Controlled Evidence Removal

Existing multi-hop benchmarks can admit reasoning shortcuts that bypass the intended use of supporting documents (Jiang & Bansal, 2019; Trivedi et al., 2022; 2020). In our validation, we distinguish question leakage, cross-document leakage and required evidence appearing in distractor documents. We impose the following conditions.

**Fact isolation.** Each required fact must be supported by its assigned gold document and absent from the remaining context:

(C1) **Gold support:** $\Gamma$ is injective and $\mathrm{supp}(\Gamma(f), f) = 1$ for every $f \in F$.

(C2) **Isolation from other documents:** $\mathrm{supp}(d, f) = 0$ for every $f \in F$ and every $d \in D \setminus \{\Gamma(f)\}$.

(C3) **Absence from the question:** $\mathrm{supp}(q, f) = 0$ for every $f \in F$, and $q$ alone is insufficient.

Table 2: **Evidence structures.** $v_1, v_2$ denote the property values and $v^*$ a fixed reference value.

| Structure | $\lvert F\rvert$ | Question $q$ asks for |
|---|---|---|
| Bridge | 2-3 | Final entity $e_2$ of $(e_0, r_1, e_1), (e_1, r_2, e_2)$ with $e_1$ unnamed in $q$ |
| Comparison | 2 | Which of two entities has the smaller/larger $v$ (or whether $v_1 = v_2$) for the same relation |
| Cross-Property | 2 | Whether $v_1 = v_2$ for two different relations |
| Contrastive | 2 | Which of two entities satisfies $v = v^*$: A, B, both, or neither |
| Nested Bridge | 4 | Comparison of the end values of two bridge paths |

For required facts $F = \{f_1, \ldots, f_k\}$ and gold documents $D^+ = \{d_1, \ldots, d_k\}$, define

$$M_{ij} = \mathrm{supp}(d_j, f_i).$$

Conditions (C1) and (C2) imply $M = I_k$ for the gold documents, while (C2) additionally excludes support from distractor documents. Removing $\Gamma(f_i)$ therefore removes $f_i$ while leaving the other required facts available.

**Question identification.** The query $q$ must identify the entities needed to follow the intended evidence structure without revealing a required fact or the answer. Because entity names can be ambiguous (Zhu & Iglesias, 2018; Min et al., 2020; Lee et al., 2024), we add a disambiguating descriptor when needed for unique reference. The descriptor is retained only if the resulting question continues to satisfy (C3).

**Evidence structure diversity.** The benchmark covers evidence structures (cf. Section 3.1) that differ in fact dependencies, entities exposed in the question, and how the answer is determined. We use manually designed templates for five overarching structures summarized in Table 2. Additional templates can be added without changing the rest of the pipeline. Sample questions for each type are given in Appendix A.1.

### 3.3 Construction and Validation

**Instance generation.** For each template, we sample the required facts $F$ from $\mathcal{KG}$, generate the question $q$ and valid answers $A$, and then assign each fact $f_i$ to its Wikipedia document $d_i = \Gamma(f_i)$.

**Validation and repair.** We verify structural constraints rule-based and use GPT-OSS-120B for semantic assessments and leakage checks, to satisfy the requirements defined in Section 3.2. Invalid instances are repaired without modifying the source documents and then fully validated again. Candidates that still fail are discarded. Validation details and repair operations are provided in Section A.2.

**Distractors.** We retrieve topically related Wikipedia documents with sentence-transformers/all-MiniLM-L6-v2 and retain a candidate $d$ only if it supports none of the required facts:

$$\mathrm{supp}(d, f) = 0 \qquad \forall f \in F.$$

### 3.4 FondueQA Benchmark

Our final release comprises **13,063 questions** with 20 context documents each: 39.3% Bridge, 39.3% Comparison, 9.8% Contrastive, 9.8% Cross-Property, and 1.7% Nested Bridge. Answers are yes/no (36.5%), option selection (21.8%), or open (41.6%). Each required fact is annotated with the Wikipedia pageviews of its role entity (median 9,778, interquartile range 2,407–55,261), which we use in Section 5 as a proxy for parametric accessibility. Further details on the popularity measure and Wikipedia corpus are given in Section A.4. Table 4 and Section A.3 give the composition and the filtering breakdown. FondueQA is publicly available as a versioned benchmark at `https://anonymous.4open.science/r/FondueQA-7C6B`, with quarterly releases to provide new benchmark instances over time.

# 4 EXPERIMENTAL SETUP

**Language Models and Configurations.** We evaluate models from four families spanning different parameter scales, dense and mixture-of-experts architectures, and where supported, both standard and reasoning inference. For models supporting both modes, we use the model's native mechanism for activating reasoning. Table 5 in Section B.1 summarizes the evaluated configurations.

All models receive the same prompt, which explicitly permits to draw on parametric knowledge and instructs to return a concise final answer. We set the temperature to 0 and top-$p$ to 0.9, with a maximum reasoning budget of 15,000 tokens where applicable and a 150-token final-answer budget otherwise. Standard and reasoning runs differ only in the activation of reasoning mode. We report complete prompts, inference parameters, and serving details in Section B.1.

**Datasets and Context Conditions.** We use extended versions of established benchmarks, denoted by the -EXT suffix, which are subject to the same checks as FONDUEQA described in Section 3.3. We evaluate on FONDUEQA, 2WIKIMULTIHOPQA-EXT, HOTPOTQA-EXT, and MINTQA-EXT. Our primary analyses use FONDUEQA, whose construction and re-balancing is explicitly designed to fulfill the requirements introduced in Section 3.2. Applying these constraints post-hoc to the -EXT variants can alter the composition of the retained subsets. We therefore use -EXT datasets primarily to test whether the observed effects generalize beyond FONDUEQA. Conversion of the -EXT variants is reported in Section B.2, with results in Section C.3.

Following our definition of reasoning paths (Definition 1) and evidence gaps (Definition 2), we evaluate a complete reasoning path, a complete gap with all gold evidence removed, and targeted partial gaps obtained by removing individual facts. A separate question-only condition removes the retrieval context entirely. For each removed fact, we additionally query the atomic fact in isolation to test whether the fact is available from the model's parametric knowledge. At a fixed evidence gap, we further evaluate reasoning traps (Definition 3) by varying the number of type-compatible distractors that provide plausible alternative reasoning paths. Construction details and coverage are provided in Section B.5. In addition to evidence availability, we control for positional effects by keeping retained gold-document positions fixed for direct role comparisons. The resulting context conditions and positional controls are detailed in Section B.3.

**Evaluation Metrics.** We use *Exact Match* (EM) as metric across all question formats and additionally define token-level F1 for open-ended answers. For a model prediction $p$ and the set of valid answer forms $A$ defined in Section 3.1, including the canonical answer and permissible aliases, we compute

$$\mathrm{EM}(p, A) = \max_{a \in A} \mathrm{EM}(p, a), \qquad \mathrm{F1}(p, A) = \max_{a \in A} \mathrm{F1}(p, a).$$

We score the single-hop fact questions using the same answer normalization and alias matching as the corresponding multi-hop questions, as detailed in Section B.4.

# 5 FINDINGS

**Bridge Evidence Can Outweigh Answer-Bearing Evidence.** On FONDUEQA, retaining the bridge fact often outweighs retaining the answer-bearing target document, particularly for smaller models. This shows that under partial retrieval, parametric knowledge can compensate for the missing fact, but its effectiveness may depend on the structural role left out.

For bridge questions, we compare two evidence gaps for a reasoning path $\pi = (f_1, f_2)$ within the same question. In Bridge-only, the first hop $f_1 = (e_0, r_1, e_1)$ is retained, thereby identifying the intermediate entity $e_1$. In Target-only, the second hop $f_2 = (e_1, r_2, e_2)$ is retained, which directly contains the final answer $a{=}e_2$. Although Target-only directly contains the answer-bearing fact, Bridge-only often performs better, particularly for smaller models (Figure 1). The predominant asymmetry persists under position-matched conditions, as

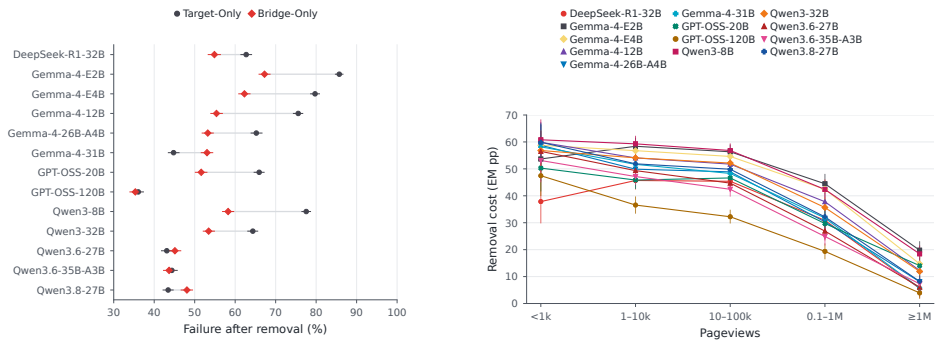



Figure 1: **Evidence structure and popularity shape compensation.** Left: Failure rates under role-specific evidence gaps for position-matched bridge questions. Right: Performance loss $\Delta_{\text{removal}}(f_i) = \text{EM}_{\text{full}} - \text{EM}_{-f_i}$ as a function of entity popularity. Error bars show 95% bootstrap confidence intervals.

detailed in Section C.1. We argue that bridge evidence is particularly valuable, as it explicitly identifies the relevant intermediate entity and thereby determines to which entity missing knowledge must be applied. Although an answer-bearing target document contains the final fact, it could be less immediately helpful if it first requires reconstructing which of the documents actually pertains to the original question.

**Popularity in Multi-Hop QA.** Across evidence gaps, models and question types, missing facts involving more popular entities incur lower removal costs, suggesting more successful parametric compensation.

We use Wikipedia page-views as a proxy for parametric accessibility and define the removal cost of a fact $f$ as $\Delta_{\text{removal}}(f) = \text{EM}_{\text{full}} - \text{EM}_{-f}$. Details on the popularity measure are provided in Section A.4. As shown in the right panel of Figure 1 missing evidence becomes less costly for more popular entities, indicating easier parametric compensation. Models can thus more frequently compensate for missing facts about popular entities, while external evidence remains more important for rarer entities. This extends the popularity effect observed for single-hop QA by Mallen et al. (2023) to multi-hop questions under partial retrieval.

**Atomic Recall Does Not Guarantee Compensation.** Atomic recall and compensation are related but not equivalent. Models can answer correctly without recalling the missing fact in isolation, while also failing to answer correctly despite recalling it in isolation.

The left panel of Figure 2 relates this atomic recall to the correctness of the corresponding multi-hop question under partial evidence. Cases in which the missing fact is recalled but the multi-hop answer remains incorrect occur only to a limited extent in most models, but are present across the board. Conversely, many models answer a significant proportion of the questions correctly even though the missing fact is not recalled atomically. Across several models, this proportion is approximately 32–45%. Therefore, a single failed recall does not preclude successful compensation. Since, by design, the missing fact is not contained in the remaining context, the context cannot provide it directly. However, it could, for example, highlight related entities and/or narrow down the plausible solution space, thereby guiding the model towards the correct answer.

Contrastive questions allow for a more detailed examination of this behavior. They are formulated as a choice between two explicitly named entities, although, depending on the underlying facts, *Both* or *None* may also be the correct answer. Under partial evidence, we refer to the entity whose relevant evidence is still supported in the context as a *visible entity*; the entity whose evidence has been removed as a *hidden entity*. The right panel of Figure 2 shows the response distribution for gold answers *Both* and *None* and additionally distinguishes whether the fact of the entities was recalled atomically.

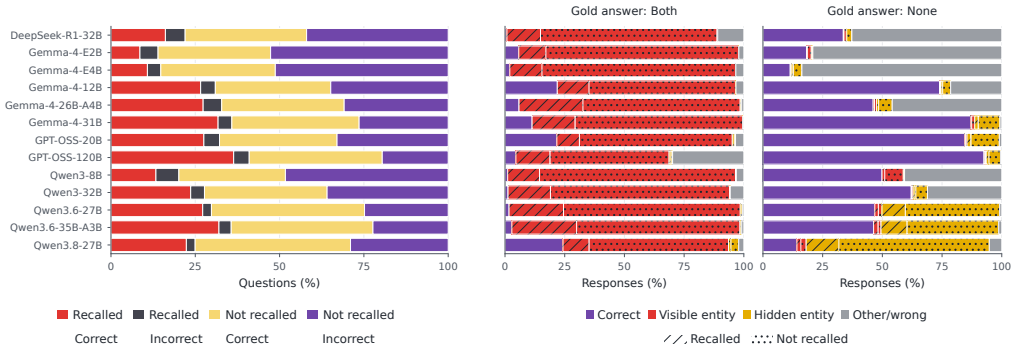



Figure 2: **Parametric availability and contextual behavior under partial evidence.** Left: Joint outcomes of atomic recall of the missing fact and correctness of the corresponding multi-hop question under partial evidence. Right: Contrastive questions with partial evidence systematically shift response behavior toward the evidence-supported entity.

When Gold=*Both*, the models often respond only with the visible entity. This is particularly plausible when the hidden fact is not recalled. It is important to note, however, that visible-entity responses also occur even when atomic recall is successful and account for up to about $27\%$ of the responses in some models. In these cases, the missing information can be recalled in isolation but is not combined with the visible evidence to produce the *Both* response. In contrast, when Gold=*None*, visible-entity responses are rare. When models select a single entity, the response more frequently falls on the hidden entity and accounts for about half of the predictions in some models. Both recalled and not-recalled cases occur here, so atomic recall only partially explains the response behavior in this scenario as well.

**Competing Evidence Can Shift Parametric Compensation.** Under a fixed evidence gap, plausible alternative entities can make it harder to use the given answer-bearing evidence. Moreover, this effect depends on what the competing evidence supports, with answer-consistent candidates strengthening the correct answer and consistently misleading documents shifting models away from it.

The right-hand plot in Figure 3 shows that, in reasoning-mode, additional plausible candidates make it more difficult for several models to utilize the answer-bearing evidence. The effect is heterogeneous across models, but can become substantial as competition increases, with drops of up to about 15 percentage points for some configurations. Candidate competition thus represents a specific reasoning trap: multiple semantically plausible bindings compete with the path intended for the question. To determine whether this effect is solely attributable to additional distractors or generally higher context complexity, we control the semantic direction of the candidate evidence. In a *Same-Gold* setting, the candidate documents support the same answer as the correct target path (e.g. other directors from 'Canada'), thereby generating additional answer-consistent paths. In the *Same-Other* setting, all documents consistently support another plausible but incorrect alternative (e.g. directors with a different nationality). The direction of this effect changes with candidate consistency. At $k{=}19$, several configurations show shifts of more than $+10$ percentage points, while non-reasoning modes can even reach roughly $+50$ points for *Same-Gold* and $-35$ points for *Same-Other*. The family-level trajectories in Figure 4 show the same tendency as competition increases for reasoning-mode configurations, although the magnitude varies across model families. Additional model-level results are provided in Section C.2.

**Implications for Retrieval-Augmented QA.** Our findings suggest that partial retrieval is not uniformly harmful. Retrieval should not only account for coverage, but also for how easily missing information can be compensated for and which role the retrieved evidence plays in the reasoning path. In particular, less popular facts

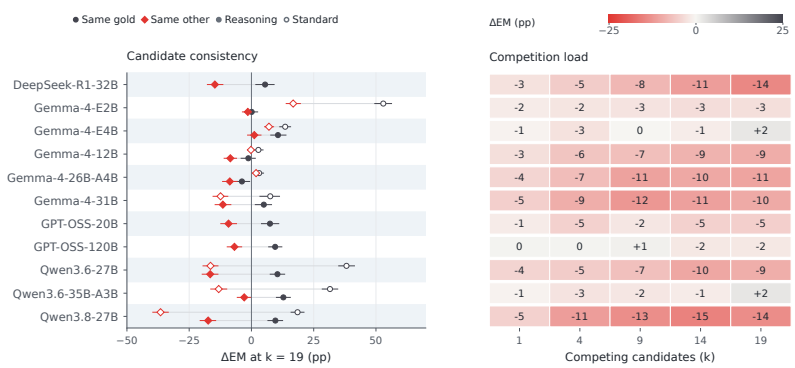



Figure 3: **Candidate competition.** Left: $\Delta$EM at $k = 19$ across candidate-consistency conditions. Right: $\Delta$EM with increasing numbers of random candidates in reasoning-mode.

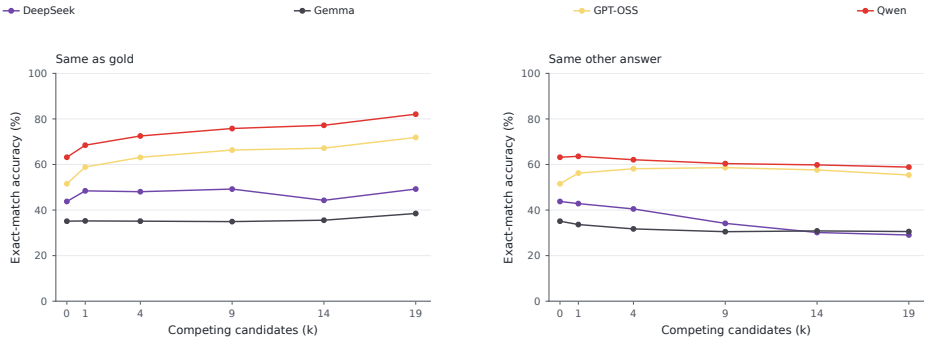



Figure 4: **Competing evidence.** Exact Match with increasing number of candidate documents. Left: *Same-Gold* candidates vs. Right: *Same-Other* candidates.

may require stronger retrieval support, while bridge evidence can sometimes be more useful than answer-bearing evidence. At the same time, retrieving many plausible alternative entities can be harmful when the bridge is missing, suggesting that candidate ambiguity matters alongside recall. Finally, models should be able to reliably integrate parametrically available facts when evidence is missing anyway.

## 6 Conclusion

We introduce FondueQA, a multi-hop QA benchmark in which every required fact is uniquely localized to a single gold document and verified to be absent from the remaining context. This enables controlled document-level evidence ablations to study parametric compensation under incomplete retrieval. Our results show that successful compensation depends not only on whether information is available parametrically, but also on the structure and ambiguity of the remaining evidence. FondueQA thereby provides a controlled setting for studying how parametric knowledge interacts with incomplete retrieval in multi-hop reasoning.

## AI USE STATEMENT

In this work generative AI tools were used to provide feedback on research methodology, assist with method implementation and code development, create and edit research artifacts and tables, summarize existing literature, identify potential research gaps and relevant literature, support brainstorming, improve the readability of the manuscript and suggest keywords.

All AI-assisted outputs were reviewed and verified by the authors. AI-assisted code and research artifacts were checked for correctness, literature-related suggestions were verified against the original sources, and all manuscript edits were reviewed before inclusion. The authors take full responsibility for the final content, claims, results and artifacts presented in this work.

## ETHICS STATEMENT

FondueQA is constructed from publicly available information in Wikipedia and Wikidata and therefore includes questions and facts concerning real persons. As both resources are collaboratively maintained and continuously editable, their contents may contain inaccuracies, outdated information, biases, or, in rare cases, deliberately misleading, discriminatory, offensive, or otherwise harmful statements. Such content is neither intended nor endorsed by the authors and should not be interpreted as independently verified claims. During benchmark construction, we apply validation and filtering procedures to remove inconsistent, unsupported, ambiguous, or otherwise unsuitable instances, including content that appears inappropriate for inclusion. However, these procedures cannot guarantee the factual correctness or harmlessness of every source statement. Consequently, FondueQA should be treated as a research benchmark derived from these sources rather than as an authoritative source of biographical or factual information about individuals, and users should consider the provenance and limitations of the underlying data when interpreting examples concerning real persons.

## REPRODUCIBILITY STATEMENT

We provide the benchmark, validated external benchmark variants, construction and evaluation code, validation prompts and experiment configurations in the anonymous repository at `https://anonymous.4open.science/r/FondueQA-7C6B`. The appendix documents the benchmark construction and validation procedure (Sections A.2 and A.3), models and serving configuration (Section B.1), context conditions (Section B.3) and external benchmark adaptation (Section B.2). All experiments used a fixed seed 42 and the released configurations reproduce the experimental conditions reported in the paper.

## REFERENCES


Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790, 2024. ISSN 1877-0509. doi: 10.1016/j.procs.2024.09.178. URL http://dx.doi.org/10.1016/j.procs.2024.09.178.

Baolong Bi, Shenghua Liu, Yiwei Wang, Yilong Xu, Junfeng Fang, Lingrui Mei, and Xueqi Cheng. Parameters vs. context: Fine-grained control of knowledge reliance in language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TJ3DqFiGau.

Jifan Chen and Greg Durrett. Understanding dataset design choices for multi-hop reasoning. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4026–4032, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1405. URL https://aclanthology.org/N19-1405/.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8706–8719, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.482. URL https://aclanthology.org/2024.naacl-long.482/.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12039–12050, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.716. URL https://aclanthology.org/2024.findings-acl.716/.

Yixiong Fang, Tianran Sun, Yuling Shi, Min Wang, and Xiaodong Gu. LastingBench: Defend benchmarks against knowledge leakage. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18304–18317, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.993. URL https://aclanthology.org/2025.findings-emnlp.993/.

Mehrdad Farahani and Richard Johansson. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16966–16977, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.943. URL https://aclanthology.org/2024.emnlp-main.943/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL https://arxiv.org/abs/2312.10997.

Wangzhen Guo, Qinkang Gong, Yanghui Rao, and Hanjiang Lai. Counterfactual multihop QA: A cause-effect approach for reducing disconnected reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4214–4226, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.231. URL https://aclanthology.org/2023.acl-long.231/.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. MINTQA: A multi-hop question answering benchmark for evaluating LLMs on new and long-tail knowledge. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 445–479, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.18. URL https://aclanthology.org/2026.acl-long.18/.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Akriti Jain and Aparna Garimella. Modeling contextual passage utility for multihop question answering. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 464–471, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-299-2. doi: 10.18653/v1/2025.ijcnlp-short.37. URL https://aclanthology.org/2025.ijcnlp-short.37/.

Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2726–2736, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1262. URL https://aclanthology.org/P19-1262/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Yoonsang Lee, Xi Ye, and Eunsol Choi. Ambigdocs: Reasoning across documents on different entities under the same name. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=mkYCfO822n.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.466. URL https://aclanthology.org/2020.emnlp-main.466/.

Bo Ni, Zheyuan Liu, Yongjia Lei, Leyao Wang, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Xin Dong, Yinglong Xia, Krishnaram Kenthapadi, Ryan Rossi, Franck Dernoncourt, Mehrab Tanjim, Nesreen Ahmed, Xiaorui Liu, Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. Towards trustworthy retrieval

augmented generation for large language models: A survey. *ACM Comput. Surv.*, 58(15), September 2026. ISSN 0360-0300. doi: 10.1145/3837074. URL https://doi.org/10.1145/3837074.

Elior Sulem, Jamaal Hay, and Dan Roth. Do we know what we don't know? studying unanswerable questions beyond SQuAD 2.0. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4543–4548, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.385. URL https://aclanthology.org/2021.findings-emnlp.385/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8846–8863, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.712. URL https://aclanthology.org/2020.emnlp-main.712/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Cheng Xu, Shuhao Guan, Derek Greene, and M-Tahar Kechadi. Benchmark data contamination of large language models: A survey, 2024. URL https://arxiv.org/abs/2406.04244.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. A theory for token-level harmonization in retrieval-augmented generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 36616–36642, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5a461bdff86cc07e976bb6c518810398-Paper-Conference.pdf.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. Crag - comprehensive rag benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10470–10490. Curran Associates, Inc., 2024. doi: 10.52202/079017-0335. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1435d2d0fca85a84d83ddcb754f58c29-Paper-Datasets_and_Benchmarks_Track.pdf.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Zican Dong, Yupeng Hou, Beichen Zhang, Yingqian Min, Junjie Zhang, Peiyu Liu, Xiaolei Wang, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Yiwen Hu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *Frontiers of Computer Science*, 20(12), May 2026. ISSN 2095-2236. doi: 10.1007/s11704-026-60308-3. URL http://dx.doi.org/10.1007/s11704-026-60308-3.

Ganggao Zhu and Carlos A. Iglesias. Exploiting semantic similarity for named entity disambiguation in knowledge graphs. *Expert Systems with Applications*, 101:8–24, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.02.011. URL https://www.sciencedirect.com/science/article/pii/S0957417418300897.

# A BENCHMARK CONSTRUCTION DETAILS

## A.1 QUESTION TEMPLATES

Table 3 summarizes the question families used in FONDUEQA. The examples illustrate the corresponding reasoning structures but are not drawn from the original dataset.

Table 3: **Question families in FondueQA.** Representative examples illustrating the reasoning structure and typical answer form of each question type.

| Type | Example question | Example answer |
|---|---|---|
| Bridge | In which country is the director of the movie *Titanic* born? | Canada |
| Comparison | Do Barack Obama and George W. Bush have the same occupation? | Yes |
| Contrastive | Who was born in Honolulu, Barack Obama or George W. Bush? | Barack Obama |
| Cross-property | Is the director of *Titanic* also the screenwriter of *Avatar*? | Yes |
| Nested Bridge | Who is older, the director of *Titanic* or the director of *Jaws*? | Steven Spielberg; Director of *Jaws* |

## A.2 VALIDATION AND REPAIR

Our validation pipeline operationalises the requirements for controlled evidence removal introduced in Section 3.2. An instance is retained only if every required fact can be uniquely assigned to its designated gold document and all applicable question-, evidence-, and context-level checks are satisfied.

**Gold support and fact isolation.** For every required fact $f_i$, we verify that its assigned gold document $\Gamma(f_i)$ supports the fact and that no other context document provides the same required evidence. This directly enforces conditions (C1) and (C2).

**Question and entity validation.** We verify that the question does not directly expose a required fact or reveal which answer candidate is correct, corresponding to condition (C3). We further check that the entities required to follow the intended reasoning structure are identified unambiguously. Questions that remain ambiguous or admit an unintended answer path are rejected or, where applicable, repaired.

**Answer and structural validation.** We verify that the designated evidence supports the composed answer and that the instantiated reasoning structure matches the intended template. In particular, required hops must contribute distinct evidence; instances with redundant hops or unintended single-document solution paths are rejected or repaired.

**Distractor and context validation.** Candidate distractors are retained only if they contain none of the required facts. After assembling the complete context, we perform a final document-level leakage check and reject contexts containing residual required evidence, duplicate documents, or other violations of evidence isolation. We additionally enforce the context-length constraints used during benchmark construction.

**Automatic adjudication.** We combine deterministic checks with semantic adjudication by GPT-OSS-120B. Deterministic checks are used whenever the relevant constraint can be established directly from the benchmark structure or annotations, while semantic checks are used for support and leakage judgments that require natural-language interpretation. The validation procedure is intentionally conservative: ambiguous

cases are rejected rather than retained when evidence is insufficient to establish the required isolation conditions. The complete validation prompts and implementation are provided in the anonymous repository at `https://anonymous.4open.science/r/FondueQA-7C6B`.

**Repair operations.** If a bridge-question instance $x = (q, D, A, F, \Gamma)$ violates the fact isolation, we attempt a conservative repair

$$\mathcal{T} : x \mapsto x' = (q', D, A', F', \Gamma'),$$

where the original Wikipedia documents $D$ remain unchanged. For a reasoning path $\pi = (f_1, \ldots, f_k)$, we consider four transformations:

$$\begin{aligned}
\mathcal{T}_{\text{retarget}} &: (f_1, \ldots, f_k) \mapsto (f_1, \ldots, f_{k-1}, f_k'), \\
\mathcal{T}_{\text{flip}} &: (f_1, \ldots, f_k; \Gamma) \mapsto (f_k', \ldots, f_1'; \Gamma'), \\
\mathcal{T}_{\text{compress}} &: (f_1, f_2, f_3) \mapsto (f_1', f_2'), \\
\mathcal{T}_{\text{reframe}} &: q(e_0) \mapsto q(\delta(e_0)).
\end{aligned}$$

Here, *retargeting* replaces the terminal fact with one supported exclusively by its designated gold document; *flipping* constructs a valid reasoning path with reversed gold-document roles and updates $\Gamma$ accordingly; *compression* replaces a path containing a redundant hop with a shorter valid path; and *reframing* replaces an explicitly named source entity $e_0$ in the question with a unique, non-answer-bearing description $\delta(e_0)$. Any resulting changes to the required facts or answer are reflected in $F'$ and $A'$. A repaired instance is retained only if it satisfies all validation conditions again.

**Final acceptance.** Repairs never modify the underlying source documents. After repair, the full instance, including the assembled context and distractors, is passed through the complete validation procedure again. Instances that still violate any required condition are discarded.

### A.3 Dataset Statistics and Filtering

**Balancing.** Starting from 198,375 generated candidates, 31,077 instances pass validation and repair (Figure 5). Because chained structures must satisfy more conditions jointly, the validated pool is skewed toward Comparison, with substantially fewer Bridge instances surviving validation. We therefore rebalance the four main question types by retaining all validated Bridge questions and stratified undersampling the remaining types to a 40/40/10/10 target distribution for Bridge, Comparison, Contrastive, and Cross-property questions, respectively. Finally, we add all validated Nested Bridge instances, which are too sparse for meaningful undersampling, yielding the 13,063-question paper release summarized in Table 4.

### A.4 Wikipedia Corpus and Entity Popularity

**Wikipedia corpus.** We use an English Wikipedia snapshot from October 2017 as the document corpus. The snapshot intentionally predates the knowledge cutoffs of modern language models considered in our experiments by a substantial margin. Since our analysis studies parametric compensation, this provides at least the possibility that the entities and facts contained in the benchmark could have occurred in pretraining data. Benchmark facts are sampled from current Wikidata, but an instance is retained only if the corresponding fact is supported by its assigned document in the 2017 Wikipedia corpus. Facts that are too recent, have changed over time, or are otherwise unsupported by the historical document therefore fail validation and are discarded. Such temporal mismatches may contribute to the *Missing gold evidence* removals reported in Figure 5.

**Entity popularity.** Following Mallen et al. (2023), we use Wikipedia pageviews as a proxy for entity popularity and, indirectly, for the frequency with which an entity may have appeared in pretraining data. For each

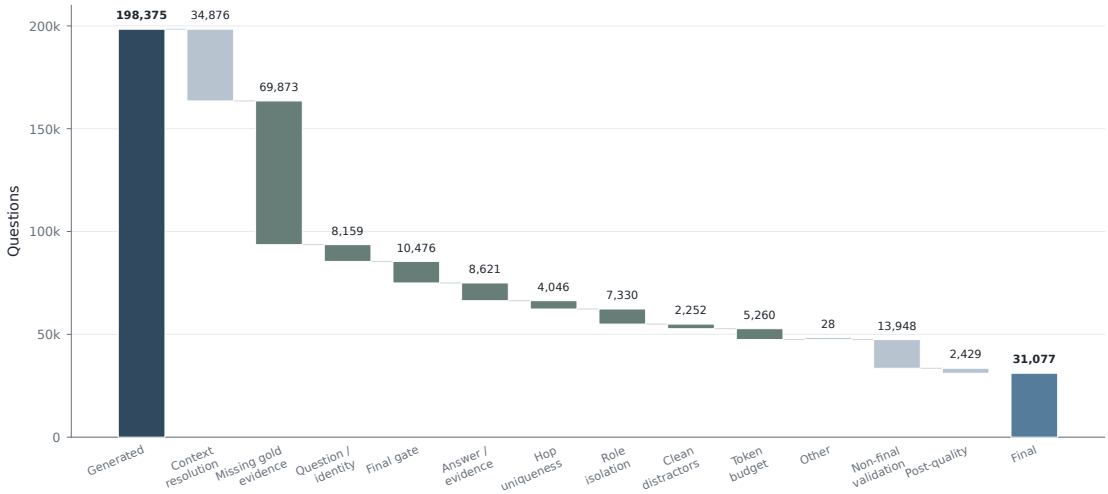



Figure 5: **FONDUEQA construction filtering.** Intermediate bars show the number of candidate questions removed by each validation or filtering step, while the first and final bars report the initial candidate pool and resulting validated dataset size.

Table 4: **FONDUEQA construction pool and rebalanced paper release.** Percentages and distributional statistics refer to the final 13,063-question release.

| Statistic | Value |
|---|---|
| Generated candidates | 198,375 |
| Candidates with complete gold context | 163,499 |
| Validated construction pool | 31,077 |
| Published paper release | 13,063 |
| Validated release backlog | 18,014 |
| Documents per question | 20 |
| Bridge | 5,138 (39.3%) |
| Comparison | 5,138 (39.3%) |
| Contrastive | 1,285 (9.8%) |
| Cross-property | 1,284 (9.8%) |
| Nested bridge | 218 (1.7%) |
| Yes/no answers | 4,771 (36.5%) |
| Option-selection answers | 2,854 (21.8%) |
| Open answers | 5,438 (41.6%) |
| Context tokens, median [IQR] | 9,142 [5,966–13,352] |
| Maximum context tokens | 29,734 |
| Role-entity pageviews, median [IQR] | 9,778 [2,407–55,261] |

entity, we sum its monthly Wikipedia pageviews over all months of 2023. We choose a period close to, but preceding the relevant training cutoffs of the evaluated models where such information is available. Compared with more recent pageview statistics, this potentially reduces the risk of assigning high popularity to entities that became prominent only after the model's training period. Pageviews remain only a proxy for parametric accessibility and should not be interpreted as direct evidence of pretraining exposure.

# B EXPERIMENTAL DETAILS

## B.1 MODEL SERVING

Table 5: **Language models evaluated in our experiments.** We evaluate standard and reasoning inference whenever supported by the model and our serving configuration. For mixture-of-experts models, parameter counts are reported as total / active parameters.

| Family | Model | Parameters | Architecture | Standard | Reasoning |
|---|---|---|---|---|---|
| *Qwen* | Qwen3.8-27B | 27B | Dense | ✓ | ✓ |
| | Qwen3.6-35B-A3B | 35B / 3B active | MoE | ✓ | ✓ |
| | Qwen3.6-27B | 27B | Dense | ✓ | ✓ |
| | Qwen3-8B | 8B | Dense | ✓ | ✓ |
| | Qwen3-32B | 32B | Dense | ✓ | ✓ |
| *Google* | Gemma-4-E2B-IT | 2B | Dense | ✓ | ✓ |
| | Gemma-4-E4B-IT | 4B | Dense | ✓ | ✓ |
| | Gemma-4-12B-IT | 12B | Dense | ✓ | ✓ |
| | Gemma-4-26B-A4B-IT | 26B / 4B active | MoE | ✓ | ✓ |
| | Gemma-4-31B-IT | 31B | Dense | ✓ | ✓ |
| *OpenAI* | gpt-oss-20b | 20B / 3.6B active | MoE | ✗ | ✓ |
| | gpt-oss-120b | 120B / 5.6B active | MoE | ✗ | ✓ |
| *DeepSeek* | DeepSeek-R1-Distill-Qwen-32B | 32B | Dense | ✗ | ✓ |

**General prompt.** All models receive the same evaluation system prompt across context conditions.

```
Answer the question based on the provided context paragraphs if
possible.  If an information is not present in the context you may
use your internal parametric knowledge.  You must perform multi-hop
reasoning by combining information from multiple documents if
necessary.

Instructions:
1.  Identify the supporting facts from the context (if possible)
that are required to answer the question.
2.  Synthesize the information to formulate the final answer.
3.  Keep your answer concise and factual.

Give a short, direct answer, usually one word or a short phrase.
For yes/no questions, answer only yes or no.
Do not explain your reasoning.
```

**Question only prompt.** When no context is provided, the model receives only the question together with the following system prompt.

```
You are a precise question-answering assistant.  Answer the
question.  Give a short, direct answer, usually one word or a short
phrase.  For yes/no questions, answer only yes or no.  Do not
explain your reasoning.
```

**Inference parameters.** We use temperature 0 with $p = 0.9$ for all model configurations. Standard inference is limited to 150 output tokens. For models with a native reasoning mode we allow a maximum reasoning budget of 15,000 tokens. All runs use a fixed random seed of 42. We use a maximum context budget of 40,000 tokens.

**Model serving.** We serve all models locally using vLLM (version: 0.26.0) on NVIDIA H100 GPUs. Runs use up to four GPUs with tensor parallelism of size 4, and pipeline parallelism of size 1. Model-specific reasoning modes are activated through their native inference mechanisms where supported. The models used are reported in Table 5.

### B.2 Extended Benchmark Construction

**External benchmark adaptation.** For HotpotQA and 2WikiMultiHopQA, we retain the original question, answer, and gold documents and apply our pipeline post-hoc to the provided contexts. MINTQA requires an additional context-construction step because the source instances do not provide a document context. We therefore resolve the annotated gold entities to their corresponding documents and construct the context before applying the same validation procedure. Instances for which a required gold document cannot be resolved are discarded. Since our analysis targets multi-hop reasoning, we additionally exclude MINTQA instances that contain only a single reasoning hop.

**Validation procedure.** For each remaining source instance, we resolve the annotated supporting evidence and map the required facts to their corresponding gold documents and reasoning roles. We then apply the same checks used for FondueQA: each required fact must be supported by its assigned gold document, must not be recoverable from another context document, and must not be directly exposed by the question. We additionally verify that the composed answer is supported, screen distractor documents for leakage of required facts, and enforce the applicable context-length constraints.

**No QA repair.** In contrast to the construction of FondueQA, we treat the original question-answer instances of the external benchmarks as immutable. We do not rewrite questions, answers, reasoning chains, or gold-document text to make an instance satisfy the requirements. If a content-level violation cannot be resolved without changing the original QA instance, the instance is discarded. The only modification permitted is the replacement of a distractor document that leaks required evidence with a clean distractor.

**Conservative filtering.** Our validation procedure is intentionally conservative. A rejection indicates that an instance did not satisfy a required check with sufficient confidence under our pipeline; it should not be interpreted as a definitive claim that the original benchmark instance is incorrect. Validation checks are applied sequentially, and an instance is removed at the first failing stage. Consequently, the reported rejection category denotes the first observed failure and does not imply that later checks would necessarily have been satisfied. In ambiguous cases, we prefer to discard an instance rather than retain one that may violate a check. The reported retention rates therefore characterize the subset for which our evidence-control requirements can be established reliably, rather than the overall quality of the source benchmarks. Table 6 presents the filtering outcomes for all benchmarks.

Applying the validation to MuSiQue-Ans leaves only a small number of valid instances. We therefore exclude MuSiQue-Ext from the main cross-benchmark analysis and focus on MINTQA-Ext, HotpotQA-Ext, and 2WikiMultiHopQA-Ext.

Table 6: **Filtering outcomes for the extended benchmarks.** MINTQA additionally requires exclusion of single-hop instances and document resolution before entering the common validation pipeline. Retention rates in parentheses are computed relative to the initial candidate pool.

| Outcome | HotpotQA -Ext | 2WikiMultiHopQA -Ext | MINTQA -Ext | MuSiQue -Ext |
|---|---|---|---|---|
| Initial candidates | 7,404 | 12,576 | 17,887 | 2,417 |
| Pre-validation: single-hop | – | – | 5,894 | – |
| Pre-validation: document resolution | – | – | 2,592 | – |
| Entered validation | 7,404 | 12,576 | 9,401 | 2,417 |
| Evidence mapping | 414 | 224 | 0 | 192 |
| Missing Evidence | 130 | 1,756 | 1,652 | 470 |
| Question / identity | 2,353 | 27 | 20 | 338 |
| Answer / evidence | 341 | 88 | 10 | 62 |
| Role isolation | 680 | 2,677 | 3,822 | 566 |
| Clean distractors | 11 | 2 | 80 | 1 |
| Token budget | 0 | 0 | 140 | 0 |
| Total rejected | 3,929 | 4,774 | 14,210 | 1,629 |
| Pending / technical | 51 | 62 | 490 | 257 |
| **Retained** | **3,424 (46.2%)** | **7,740 (61.5%)** | **3,187 (17.8%)** | **531 (22.0%)** |

### B.3 Context Conditions

Our evaluation contexts factorize two independent dimensions: which gold evidence is available and where the retained gold documents are placed. Table 7 summarizes the resulting conditions.

**Evidence availability.** We evaluate the complete gold context, contexts with no gold evidence, and partial-evidence conditions in which only selected reasoning roles are retained or individual gold documents are removed. A separate question-only condition removes the retrieval context entirely. These conditions induce the Evidence Gaps used throughout our analyses.

**Evidence placement.** Contexts in FONDUEQA contain 20 documents. Depending on the condition, retained gold documents are placed toward the front, middle, or end of the context. The exact positions are given in Table 8.

**Position-matched comparisons.** For direct comparisons between evidence roles, we keep the positions of the retained gold documents fixed across conditions. This ensures that differences between, for example, Bridge-only and Target-only contexts are not driven by systematic differences in document position.

### B.4 Evaluation Details

**Answer normalization and scoring.** Predictions and references are normalized using Unicode NFKC normalization, lowercasing, removal of ASCII punctuation and the English articles *a*, *an*, and *the*, and whitespace normalization. The reference set contains the gold answer and its stored aliases. Exact Match (EM) is one if the normalized prediction is nonempty and matches any normalized reference. Token-level F1 is computed

Table 7: **Factorized context construction.** Each evaluation context combines an evidence-availability setting with a compatible evidence-placement setting. Availability determines which gold documents are present; placement determines their positions in the 20-document context.

| Evidence availability | | Evidence placement | |
|---|---|---|---|
| **Setting** | **Gold evidence present** | **Setting** | **Placement** |
| Question only | None | Front | Starts at position 1 |
| No gold | None; distractors only | Middle | Centered at position 10 |
| All gold | All gold documents | End | Ends at position 19 |
| Role only | One selected role | | |
| Leave-one-out | All except one | | |
| Evaluation context = Evidence availability × Evidence placement | | | |

Table 8: **Gold-document positions.** Positions depend on the number of retained gold documents $k$.

| $k$ | **Front** | **Middle** | **End** |
|---|---|---|---|
| 1 | 1 | 10 | 19 |
| 2 | 1–2 | 10–11 | 19–20 |
| 3 | 1–3 | 9–11 | 18–20 |
| 4 | 1–4 | 9–12 | 17–20 |

over whitespace-separated normalized tokens as

$$\mathrm{F1}(p, a) = \frac{2c(p, a)}{|p| + |a|},$$

where $c(p, a)$ is the multiset token overlap and $|p|$ and $|a|$ denote the respective token counts. We report the maximum score over all permissible references. For `yes`, `no`, and `noanswer`, pairwise F1 reduces to exact agreement. Empty or failed predictions receive zero, while missing requests are excluded.

### B.5 Reasoning Trap Construction

**Target-only evidence gap.** We construct the Reasoning Trap experiment from two-hop Bridge questions in the final FondueQA release whose original reasoning path is preserved. Questions whose question or reasoning chain was modified through semantic repair are excluded. Each context uses the partial evidence gap $G = \{f_1\}$, corresponding to a Target-only condition: the bridge document identifying the target entity is removed, while the true target document and its terminal evidence remain available. The model must therefore recover the missing bridge parametrically while distinguishing the true target from plausible alternatives.

**Candidate construction.** Competitors must occupy the same functional position as the true target. They must be reachable through the same upstream relation, support the same terminal relation, and share at least one concrete `instance_of` value with the true target. For example, for

$$\textit{film} \rightarrow \textit{director} \rightarrow \textit{country of citizenship},$$

a competitor must likewise occur as the director of another work and have a supported citizenship value. Remaining positions are filled with *bridge-like* documents whose entities match the type of the original bridge entity but do not satisfy the target-candidate information.

**Evidence and leakage constraints.** Candidates are retrieved from the local knowledge graph and, where necessary, supplemented through targeted Wikidata queries. Every selected entity must have an article in the October 2017 English Wikipedia snapshot, and a competitor is retained only if its terminal value or a validated alias is supported by the article. Entities overlapping with the original reasoning path are excluded, and inserted articles may not mention original path entities. Bridge-like documents additionally may not mention selected competitors. Duplicate documents are removed, and all checks are applied to the complete retained article. These mention constraints use lexical matching over entity names and aliases.

**Competition conditions.** For every question, we construct contexts with

$$k \in \{0, 1, 4, 9, 14, 19\}$$

competitors. Each context contains exactly 20 documents and at most 30,000 `o200k_base` tokens. The true target is fixed at rank 1, competitors occupy ranks $2, \ldots, k+1$, and the remaining positions contain bridge-like documents:

$$[\text{gold target}, \underbrace{\text{competitors}}_{k}, \underbrace{\text{bridge-like documents}}_{19-k}].$$

Competitors therefore replace distractors rather than increasing context size. Candidate order is deterministic, and larger values of $k$ extend the competitor prefix used at smaller values.

For $k > 0$, we construct three conditions.

**SAME-GOLD.** All competitors support the gold terminal value.

**SAME-OTHER.** All competitors support the same non-gold terminal value. The value is selected from an answer group that supports the required competition level.

**RANDOM.** Competitors are sampled from the full eligible pool without constraining their terminal values to agree.

Thus, SAME-GOLD and SAME-OTHER control the semantic direction of competing evidence, whereas RANDOM varies candidate competition without enforcing a common answer.

**Coverage.** Of the 5,138 Bridge questions in the release, 2,594 preserve their original reasoning path and enter candidate construction. After type, bridge-like document, and competitor availability checks, 2,207 questions retain a valid $k = 0$ control. The paired cohorts contain 1,527, 1,230, 1,060, 918, and 967 questions for $k = 1, 4, 9, 14, 19$, respectively. Analyses requiring complete coverage across all positive competition levels use the intersection of these cohorts ($n = 871$). Candidate-competition experiments are reported for the subset of model configurations for which complete runs across all competition conditions are available.

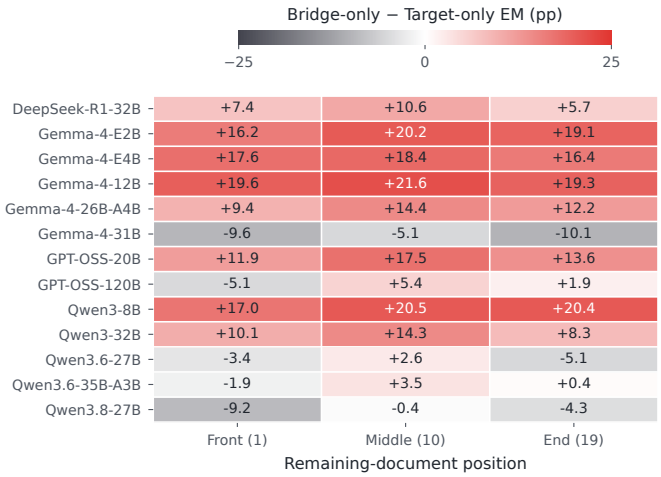



Figure 6: **Position-matched role comparison.** Difference in Exact Match between Bridge-only and Target-only when the retained gold document is placed at the same context position in both conditions. Positive values indicate higher performance with Bridge-only evidence.

# C Additional Results

## C.1 Position-Matched Comparisons

To control for document position, we compare Bridge-only and Target-only while placing the retained gold document at the same position in both conditions. We repeat this comparison with the retained document at the front, middle, and end of the context. Figure 6 reports the resulting difference $\mathrm{EM}_{\text{Bridge-only}} - \mathrm{EM}_{\text{Target-only}}$, such that positive values indicate an advantage for bridge evidence.

The predominant role asymmetry persists across document positions. In particular, several smaller models retain a clear Bridge-only advantage at the front, middle, and end of the context. The magnitude varies with position, and some larger models remain close to zero or favor Target-only evidence, indicating that position can modulate but does not generally account for the observed role effect.

## C.2 Additional Candidate Competition Results

We provide additional model-level results for the candidate competition experiment from Section 5. Figure 7 shows the change in Exact Match across increasing competition levels for the random-candidate condition. The effect is heterogeneous across model configurations, but additional competitors reduce performance for many configurations, with larger reductions often appearing at higher competition levels.

Figure 8 shows the complete trajectories for the *Same-Gold* and *Same-Other* conditions. *Same-Gold* competitors generally shift performance toward the correct answer as competition increases, whereas *Same-Other* competitors tend to shift performance in the opposite direction. The magnitude of both effects varies substantially across models and inference modes.

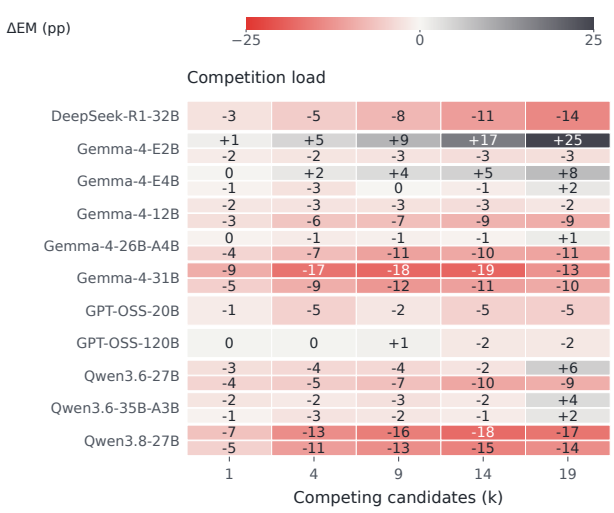



Figure 7: **Candidate competition across competition levels.** Change in Exact Match relative to the $k = 0$ condition as the number of random type-compatible competitors increases.

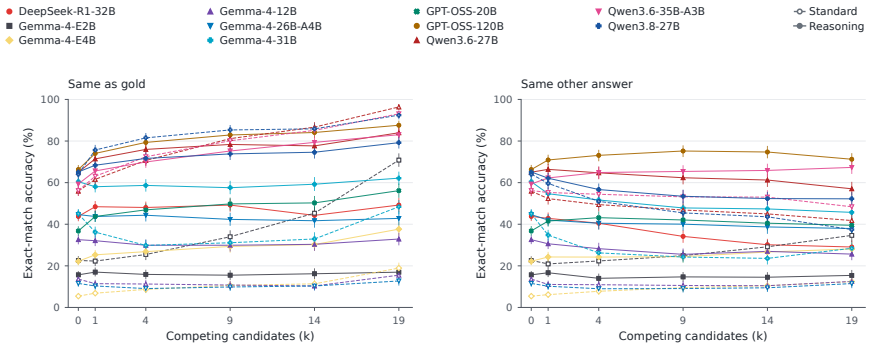



Figure 8: **Candidate-consistency effects across competition levels.** Exact Match with increasing numbers of competing candidates. Left: Same-Gold candidates support the correct terminal value. Right: Same-Other candidates consistently support an alternative value.

### C.3 Extended Benchmark Results

We use the validated external benchmark variants to examine if effects observed on FondueQA transfer to independently constructed multi-hop QA benchmarks. We evaluate HotpotQA-Ext, 2WikiMultiHopQA-Ext, and MINTQA-Ext using the same available model configurations, scoring procedure, and evidence gap definitions wherever available. The resulting role analysis contains up to 5,123 FondueQA questions, 2,252 HotpotQA-Ext questions, 3,545 2WikiMultiHopQA-Ext questions, and 2,184 MINTQA-Ext questions per model. For HotpotQA, we additionally report the corresponding standard benchmark setting as a reference. We exclude MuSiQue-Ext from the main cross-benchmark analysis because only 531 in-

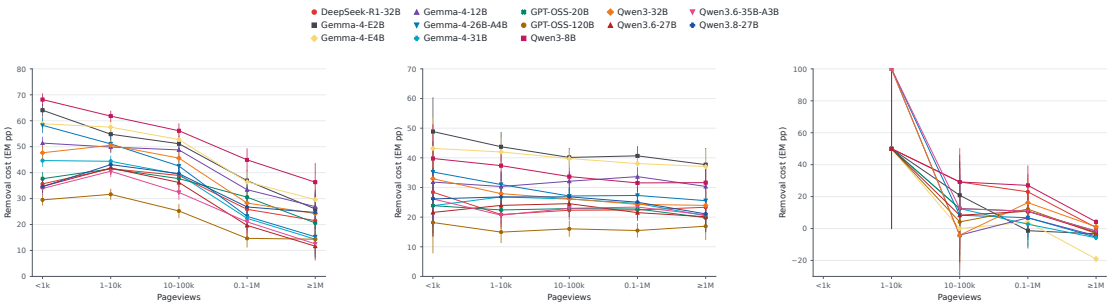



Figure 9: **Entity popularity and parametric compensation on the extended benchmarks.** Left: 2WIKIMULTIHOPQA-EXT, Middle: HOTPOTQA-EXT, Right: MINTQA-EXT

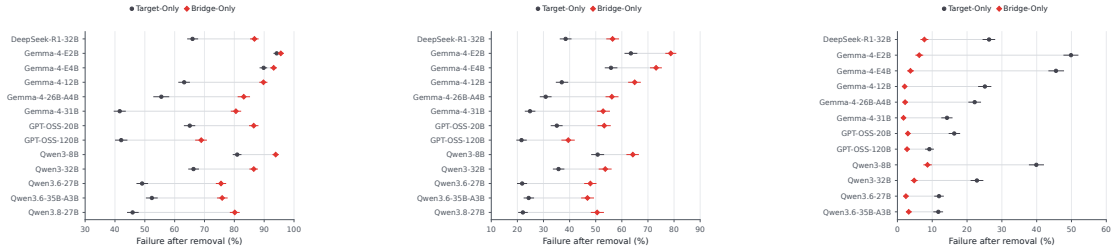



Figure 10: **Role-specific Evidence Gaps on the external benchmarks.** Failure after removing bridge or target evidence, conditioned on full-context success. 2WIKIMULTIHOPQA-EXT (left) and HOTPOTQA-EXT (middle) predominantly favor Target-only evidence, whereas MINTQA-EXT (right) strongly favors Bridge-only evidence.

stances remain after applying our validation pipeline. Where applicable, we also include results on standard HOTPOTQA as a reference.

**Popularity effects generalize.** We first examine whether the relationship between entity popularity and parametric compensation transfers beyond FONDUEQA. As in the main analysis, we group all questions by the popularity of the entity associated with the removed fact and measure the resulting removal cost. The same general trend is visible on the external benchmarks, as missing evidence associated with more popular entities is typically less costly to remove, indicating that parametric compensation becomes easier as the relevant entity becomes more popular. However, the magnitude differs across datasets and models as can be seen in Figure 9.

**Evidence-role effects are benchmark-dependent.** As shown in Figure 10, the structural role effect transfers less uniformly. On FONDUEQA, Bridge-only evidence outperforms Target-only evidence for most models, while MINTQA-EXT shows the same direction and generally an even larger Bridge-only advantage. In contrast, 2WIKIMULTIHOPQA-EXT and HOTPOTQA-EXT predominantly favor Target-only evidence. The corresponding standard HOTPOTQA setting shows the same tendency, but with a weaker asymmetry.

To identify where these differences arise, we separately compare Bridge-only and Target-only accuracy across datasets (Figure 11). The largest discrepancies appear in the Bridge-only condition. Models retain substan-

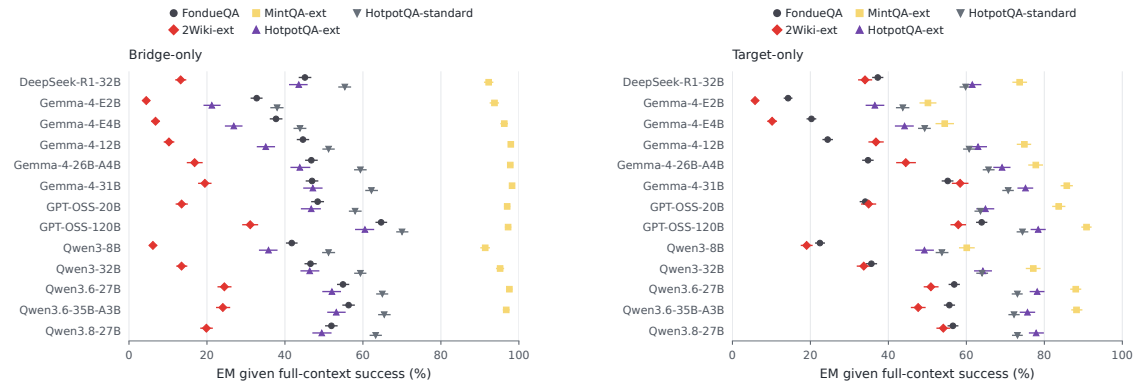



Figure 11: **Decomposition of the cross-benchmark role effect.** Partial-context EM conditioned on full-context success for Bridge-only and Target-only evidence. Differences across benchmarks are particularly pronounced in the Bridge-only condition.

tially more of their full-context successes from bridge evidence on FONDUEQA and especially MINTQA-EXT, whereas Bridge-only accuracy is markedly lower on 2WIKIMULTIHOPQA-EXT.

**Popularity does not explain the 2WIKIMULTIHOPQA-EXT reversal.** A possible explanation for the different role effect is that FONDUEQA and 2WIKIMULTIHOPQA-EXT contain entities with different levels of popularity. We therefore align their bridge- and target-entity popularity distributions (Figure 12).

First, one-to-one matching without replacement yields 1,892 questions from each benchmark. The resulting bridge medians are 3,479 pageviews for FONDUEQA and 3,466 for 2WIKIMULTIHOPQA-EXT, while the target medians are 10,534 and 10,412, respectively. Despite this alignment, the role-effect reversal remains.

We additionally reweight FONDUEQA toward the complete 2WIKIMULTIHOPQA-EXT popularity profile. The resulting bridge medians are 2,659 and 2,630 pageviews, respectively, and the target medians are 7,569 and 7,428. The weighted FONDUEQA sample contains 1,289 unique questions with an effective sample size of approximately 846, while the reference 2WIKIMULTIHOPQA-EXT profile contains 2,124 questions. The qualitative difference again remains. Differences in role-entity popularity alone therefore do not explain the 2WIKIMULTIHOPQA-EXT reversal.

**The reversal persists on easier questions.** We next test whether differences in overall question difficulty account for the role effect. For each evaluated model, we retain questions for which the other models achieve at least 90% mean full-context EM (Figure 13), excluding the evaluated model from the difficulty estimate. This leaves 2,921–3,260 FONDUEQA, 1,035–1,093 HOTPOTQA-EXT, 836–928 2WIKIMULTIHOPQA-EXT, and 1,747–1,791 MINTQA-EXT questions per model. For the standard HOTPOTQA reference, 2,601–2,740 questions remain.

The benchmark-specific pattern remains under this restriction. 2WIKIMULTIHOPQA-EXT and HOTPOTQA-EXT continue to favor Target-only evidence for most models, whereas MINTQA-EXT remains strongly Bridge-only-favoring. The standard HOTPOTQA setting shows a weaker Target-only tendency. FONDUEQA also remains Bridge-only-favoring for most models, although several effects move closer to zero. Hence, restricting the analysis to questions that are reliably answered from complete evidence does not remove the benchmark-specific role differences.

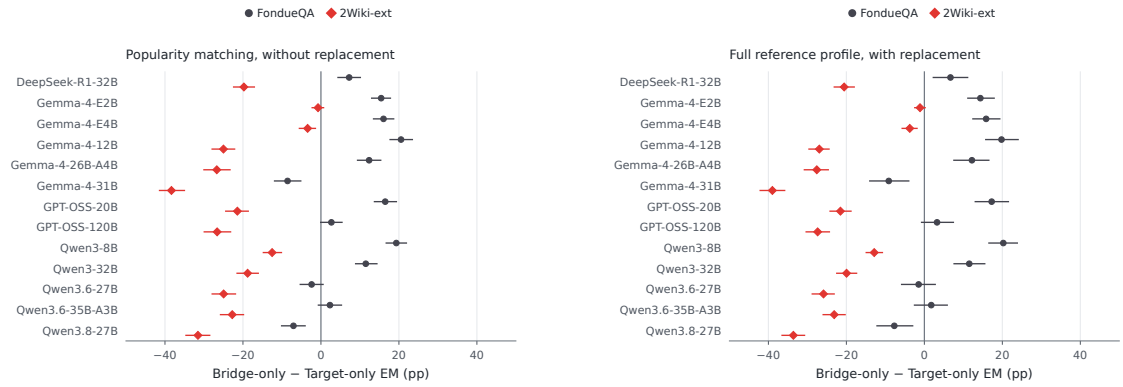



Figure 12: **Role effects after controlling for entity popularity.** Bridge-only minus Target-only EM after one-to-one popularity matching without replacement (left) and reweighting FONDUEQA toward the full 2WIKIMULTIHOPQA-EXT popularity profile (right). The cross-benchmark reversal persists under both controls.

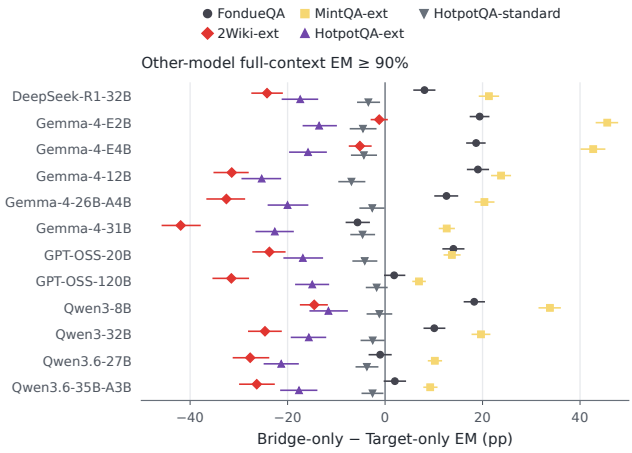



Figure 13: **Role effects on questions reliably solved by other models.** Bridge-only minus Target-only EM for questions on which the remaining evaluated models achieve at least 90% mean full-context EM.

**Discussion.** The external results show different degrees of generalization for the two determinants studied in Section 5. Entity popularity transfers comparatively well, as across benchmarks, missing information about more popular entities is generally easier to compensate for parametrically. The relative utility of bridge and target evidence, however, is benchmark-dependent. Bridge evidence can be more useful than the answer-bearing document, as observed on FONDUEQA and MINTQA-EXT. In contrast, 2WIKIMULTIHOPQA-EXT and HOTPOTQA-EXT predominantly favor Target-only evidence, while the standard HOTPOTQA setting shows a weaker Target-only tendency.

Table 9: **Context characteristics for the role analysis.** Token counts denote median inference-time context lengths.

| Dataset | Documents | Tokens |
|---|---|---|
| FONDUEQA | 20 | 7,243 |
| HOTPOTQA-EXT | 10 | 1,268 |
| HOTPOTQA | 10 | 1,383 |
| 2WIKIMULTIHOPQA-EXT | 10 | 732 |
| MINTQA-EXT | 10 | 952 |

Neither aligning entity popularity nor restricting the analysis to easier questions removes the difference between FONDUEQA and 2WIKIMULTIHOPQA-EXT. The datasets nevertheless differ substantially in their context construction (Table 9).

Context length alone is unlikely to explain the benchmark-specific role differences. All external settings considered here use ten-document contexts and are substantially shorter than FONDUEQA, yet their role effects point in different directions. A possible difference instead lies in the composition of the remaining context and in how many plausible entities or bindings remain after an evidence gap is introduced. This would be consistent with the candidate competition experiment (Section 5), where plausible alternative targets change the usefulness of otherwise available evidence.

The external comparison does not manipulate these properties directly. Context composition and candidate ambiguity should therefore be interpreted as possible explanations rather than established causes of the cross-benchmark differences. Overall, the results suggest that parametric accessibility through entity popularity generalizes comparatively robustly, whereas evidence roles do not admit a universal ordering of importance. Their usefulness depends on the broader question–context configuration in which the missing reasoning step must be recovered.