# RECOR: Reasoning-focused Multi-turn Conversational Retrieval Benchmark

**Mohammed Ali**[1*] **Abdelrahman Abdallah**[1] **Amit Agarwal**[2]
**Hitesh Laxmichand Patel**[2] **Adam Jatowt**[1]

[1]University of Innsbruck, Austria [2]Oracle AI, USA

## Abstract

Existing benchmarks treat multi-turn conversation and reasoning-intensive retrieval separately, yet real-world information seeking requires both. To bridge this gap, we present a benchmark for reasoning-based conversational information retrieval comprising 707 conversations (2,971 turns) across eleven domains. To ensure quality, our Decomposition-and-Verification framework transforms complex queries into fact-grounded multi-turn dialogues through multi-level validation, where atomic facts are verified against sources and explicit retrieval reasoning is generated for each turn. Comprehensive evaluation reveals that combining conversation history with reasoning doubles retrieval performance (Baseline .236 $\rightarrow$ History+Reasoning .479 nDCG@10), while reasoning-specialized models substantially outperform dense encoders. Despite these gains, further analysis highlights that implicit reasoning remains challenging, particularly when logical connections are not explicitly stated in the text. [1]

## 1 Introduction

Information seeking is inherently conversational. Exploring complex topics, users rarely employ single, comprehensive queries, instead engaging in multi-turn dialogue, progressively refining their understanding through iterative exchanges. This motivates Conversational Information Retrieval (CIR) systems that support ongoing interaction over isolated sessions.

Retrieval-Augmented Generation (RAG) has emerged as a critical approach for grounding LLM responses in external knowledge (Lewis et al., 2020; Chen et al., 2024b). The primary focus of RAG benchmarks has been on single-turn interactions (Friel et al., 2024; Yang et al., 2024; Pattnayak et al., 2025). Multi-turn RAG, where each turn depends on preceding context, presents additional challenges that remain underexplored (Katsis et al., 2025).

A parallel challenge exists in retrieval itself. The BRIGHT benchmark (Su et al., 2024) demonstrated that current retrieval systems struggle significantly with reasoning-intensive queries. These queries require multi-step inference, connecting disparate pieces of information, and drawing non-obvious conclusions. Standard lexical and semantic matching fails when relevance depends on logical reasoning rather than surface similarity.

These two challenges, multi-turn conversation and reasoning-intensive retrieval, have been studied separately. However, no existing benchmark combines them. Yet real information needs often require both. A researcher investigating climate policy might ask why carbon prices worked in some countries but not others, then whether the successful cases shared common features, then how those lessons apply to developing nations, etc. Each turn requires reasoning to identify relevant documents, and each answer builds on previous understanding.

Bridging this gap requires high-quality multi-turn conversations, yet these are difficult to obtain. Human annotation is expensive and does not scale. Synthetic generation is fast but often produces shallow conversations with repetitive content and limited document diversity. Without principled methods that incorporate explicit reasoning, one cannot create the benchmarks needed to advance conversational search systems.

We address this challenge through a Decomposition-and-Verification framework that transforms complex single-turn queries into grounded multi-turn dialogues. Given a query with its gold answer and supporting documents, the system validates document-answer alignment, then decomposes the answer into granular aspects representing distinct information facets. From

*Corresponding author: mohammed.ali@uibk.ac.at
[1]https://github.com/RECOR-Benchmark/RECOR



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 2688–2723
July 2-7, 2026 

each aspect, atomic facts are extracted and verified against sources. For each verified aspect, the system generates a focused sub-question, scores relevant documents, and produces a grounded conversational turn.

Unlike approaches relying solely on semantic matching, our framework generates explicit retrieval reasoning for each turn, specifying what information documents should contain and what signals indicate relevance. This reasoning-guided selection produces natural document diversity, with different turns drawing upon different passages.

Our contributions: (1) A Decomposition-and-Verification framework that transforms complex queries into fact-grounded multi-turn dialogues, reusable for conversational data synthesis; (2) Using this framework, a benchmark for reasoning-intensive conversational retrieval across eleven domains (707 conversations, 2,971 turns); (3) Extensive experiments on retrieval and generation showing that history + reasoning doubles retrieval (.236 $\rightarrow$ .479 nDCG@10), reasoning-specialized models substantially outperform dense encoders, and implicit reasoning remains challenging.

## 2 Related Work

We summarize key related work below while Appendix A provides a comprehensive review.

**Conversational Search.** Researchers have developed several benchmarks for multi-turn information seeking. QReCC (Anantha et al., 2021) contains 14K open-domain conversations where users ask follow-up questions that require resolving references to previous turns. TREC iKAT (Aliannejadi et al., 2024) extends this setting by incorporating user personas across 20 topics, requiring systems to personalize responses based on user background. MTRAG (Katsis et al., 2025) evaluates multi-turn RAG systems on 110 conversations with challenges like unanswerable questions across four domains. While capturing conversational complexity, these benchmarks assume document relevance relies primarily on semantic similarity.

**Reasoning for Retrieval.** BRIGHT (Su et al., 2024) showed that current retrievers fail on queries requiring reasoning. Systems achieving 59.0 nDCG@10 on standard benchmarks score only 18.3 on BRIGHT. The benchmark also showed that providing chain-of-thought reasoning improves retrieval by up to 12.2 points. However, BRIGHT only covers single turn queries (Table 1).

| Benchmark | MT | RI | FG | RR | Dom. |
|---|---|---|---|---|---|
| QReCC | ✓ | × | × | × | — |
| TREC iKAT | ✓ | × | × | × | 20* |
| MTRAG | ✓ | × | × | × | 4 |
| BRIGHT | × | ✓ | × | ✓ | 12 |
| **Ours** | ✓ | ✓ | ✓ | ✓ | **11** |

Table 1: Benchmark comparison. MT: Multi-turn. RI: Reasoning-intensive. FG: Fact-grounded. RR: Retrieval reasoning. Dom.: Number of domains. QReCC is open-domain. *TREC iKAT has 20 topics, not domains.

## 3 Dataset Construction and Analysis

Figure 1 illustrates our complete pipeline, which we detail in the following subsections. See Appendix C for a detailed example. (Pipeline prompts: see Appendix F)

### 3.1 Data Sources and Collection

Our benchmark draws from two complementary sources: the existing BRIGHT benchmark and the newly collected StackExchange data. This combination provides both established reasoning-intensive queries and domain-specific technical questions, enabling comprehensive evaluation across eleven domains.

#### 3.1.1 BRIGHT Benchmark

We leverage the BRIGHT dataset (Su et al., 2024) as a foundational component, utilizing its complex queries and human-annotated reasoning across six domains: biology, psychology, economics, earth science, sustainable living, and robotics. These queries require logical inference rather than surface similarity, and the original positive and negative passages, curated for multi-step reasoning, are preserved.

#### 3.1.2 StackExchange Collection

To extend domain coverage, we collected additional complex queries from StackExchange across five complementary domains: drones, hardware, law, medical sciences, and politics. Our collection methodology ensures data quality comparable to BRIGHT.

**Question-Answer Extraction.** We scraped StackExchange domain-specific forums, applying strict selection criteria. Questions must be sufficiently complex and detailed, filtering out simple

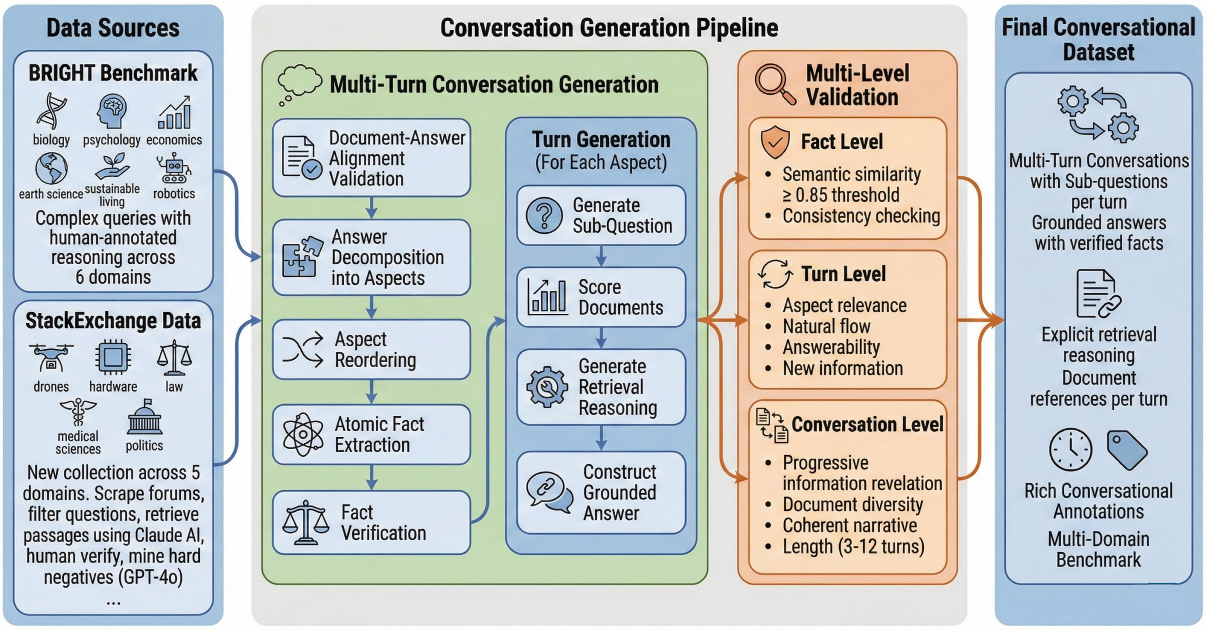


Figure 1: Overview of Decomposition-and-Verification framework for generating grounded multi-turn conversations.

requests or yes/no questions. Only questions with verified (accepted) answers were retained, and answer completeness was verified to ensure meaningful responses requiring reasoning.

**Positive Passage Retrieval.** For each question-answer pair, we employed a two-stage process. First, Claude AI Assistant retrieved candidate passages from diverse sources (Wikipedia, technical blogs, research articles). The assistant was prompted to find passages providing information necessary to answer the question, even if not directly stating the answer. Second, human annotators verified each retrieved passage for relevance and quality consistency across domains.

**Hard Negative Mining.** Standard negative sampling (e.g., BM25 top-k) produces passages easily distinguishable from positives. We developed a query-based approach where GPT-4o (OpenAI, 2024) generates topically related queries focused on unhelpful aspects. We retrieved web pages via Google Custom Search API, excluding original positive sources, and segmented them into comparable-length passages. Automated verification removed duplicates and passages with high lexical overlap. This method produces hard negatives that share domain vocabulary but lack the specific information needed for answering, creating a genuine challenge for reasoning-based retrieval.

### 3.2 Multi-Turn Conversation Generation

Transforming single-turn QA pairs into multi-turn conversations requires balancing document grounding with natural dialogue flow. Our Decomposition-and-Verification framework achieves this through systematic answer decomposition, fact-level verification, and explicit retrieval reasoning.

#### 3.2.1 Answer Decomposition and Fact Extraction

Given a question $Q$, answer $A$, and supporting documents $D = \{d_1, d_2, \ldots, d_n\}$, our pipeline first validates document-answer alignment, filtering irrelevant or weakly related passages.

The validated answer is decomposed into granular aspects: distinct components representing coherent subtopics or reasoning steps. Aspect extraction is iterative, checking each candidate for meaningful content, non-overlapping, and sufficient document coverage. For example, a climate policy answer might yield aspects covering carbon pricing, implementation challenges, and economic trade-offs.

Aspects are reordered to ensure natural conversational flow (e.g., progressing from concepts to implications; see Appendix F for prompts). From each aspect, we extract **atomic facts**: minimal statements that cannot be divided further. Each fact is verified against source documents via semantic similarity and logical consistency; facts without clear support are discarded.

#### 3.2.2 Turn Generation with Retrieval Reasoning

For each verified aspect, the system generates a focused conversational turn. A natural follow-up question targets the specific aspect, building upon preceding context while introducing new information needs. Relevant documents are scored on multiple criteria: coverage of atomic facts, answer completeness, explanation clarity, and penalties for misleading content. Unlike semantic similarity alone, this scoring considers whether documents contain the specific information needed to address the sub-question.

For each selected document, we generate explicit reasoning explaining its relevance: what information it contains and what signals indicate its usefulness. The turn's answer is constructed from verified facts with direct connections to source passages. This reasoning-guided approach produces natural document diversity across turns, as different aspects draw upon different information subsets.

#### 3.2.3 Multi-Level Validation

Quality assurance operates at three levels. **At the fact level**, each atomic fact is verified against source documents, and facts are accepted if they are explicitly stated, clearly implied, or are reasonable paraphrases. **At the turn level**, generated sub-questions are validated for aspect relevance, natural flow, answerability from available documents, and distinctiveness from previous turns. **At the conversation level**, complete dialogues are evaluated for logical progression, document diversity across turns, coherent narrative structure, and appropriate length (3–12 turns). (See Appendix F for prompts).

### 3.3 Dataset Statistics and Characteristics

Our benchmark comprises 707 conversations with 2,971 turns across eleven domains (Table 2). The StackExchange subset contributes 220 conversations (956 turns) while BRIGHT provides 487 conversations (2,015 turns). Conversations average 4.20 turns (range 3–12) with 5.87 semantic aspects and 2.01 supporting documents per turn. The benchmark contains 507,141 total documents (2,900 positive, 504,241 hard negatives), creating a challenging 174:1 negative-to-positive ratio. Queries average 18 words, answers average 70 words, and conversation history accumulates to approximately 155 words. The eleven domains cover diverse reasoning patterns: evidence-based inference (scientific), procedural reasoning (technical), and causal analysis (social). Detailed statistics are provided in Appendix B.

| Domain | #Conv. | #Turns | Avg #T | Avg #A | Avg #D |
|---|---|---|---|---|---|
| Drones | 37 | 142 | 3.84 | 4.76 | 2.36 |
| Hardware | 46 | 188 | 4.09 | 5.20 | 2.10 |
| Law | 50 | 230 | 4.60 | 7.00 | 2.55 |
| Medical Sci. | 44 | 183 | 4.16 | 6.18 | 2.23 |
| Politics | 43 | 213 | 4.95 | 7.42 | 2.49 |
| Biology | 85 | 362 | 4.26 | 5.89 | 1.56 |
| Earth Sci. | 98 | 454 | 4.63 | 6.23 | 1.58 |
| Economics | 74 | 288 | 3.89 | 5.62 | 2.28 |
| Psychology | 84 | 333 | 3.96 | 5.65 | 2.16 |
| Robotics | 68 | 259 | 3.81 | 4.94 | 1.76 |
| Sust. Living | 78 | 319 | 4.09 | 5.79 | 1.88 |
| **Total** | **707** | **2,971** | **4.20** | **5.87** | **2.01** |

Table 2: Per-domain statistics. Avg #T: average turns per conversation; Avg #A: average aspects per conversation; Avg #D: average documents per turn.

### 3.4 Human Evaluation

To validate conversation quality, three PhD students evaluated 200 randomly sampled conversations (balanced across domains) on four criteria using a 1–5 Likert scale: naturalness (dialogue flow), turn coherence (logical progression), question quality (meaningful sub-questions), and groundedness (answer-document alignment). Additionally, annotators manually revised the automatically selected documents and generated retrieval reasoning for each turn, correcting any misalignments or inaccuracies. Results show strong performance: naturalness 4.2 ($\kappa$=0.79), turn coherence 4.6 ($\kappa$=0.77), question quality 4.2 ($\kappa$=0.75), and groundedness 4.4 ($\kappa$=0.73), validating our Decomposition-and-Verification framework.

To extend this validation to the full dataset, we used GPT-4o to automatically assess all conversations on the same four criteria. LLM scores align closely with human judgments, with differences ranging from +0.14 to +0.35, all within acceptable thresholds. We also analyzed turn-level dependency types and question patterns. Full results and evaluation prompts are provided in Appendix G.

## 4 Experiments

We evaluate our benchmark through retrieval experiments that assess passage retrieval quality (§4.1) and generation experiments that judge answer quality from retrieved passages (§4.2).

### 4.1 Retrieval Evaluation

#### 4.1.1 Setup

**Query Processing Strategies.** Multi-turn queries challenge retrieval systems through coreferences and context dependencies. We evaluate five strategies: **Baseline** uses the raw current query. **Query Rewrite** employs GPT-4o to generate self-contained queries. **Reasoning** decomposes information needs via explicit search rationale. **History** prepends full conversation history. **History+Reasoning** combines context with reasoning.

**Retrieval Models.** We evaluate eight retrievers spanning three categories. Reasoning-specialized models include DIVER (Long et al., 2025) and ReasonIR (Shao et al., 2025), trained on multi-hop queries requiring inference beyond surface matching. General-purpose dense encoders include BGE (Chen et al., 2024a), E5 (Wang et al., 2024), Contriever (Izacard et al., 2021), SFR (Nguyen et al., 2024), and Qwen (Bai et al., 2023). For lexical retrieval, we include BM25 (Robertson and Zaragoza, 2009) as a sparse baseline.

**Metrics.** We report nDCG@10 (Järvelin and Kekäläinen, 2002) as our primary metric. For each retriever and strategy combination, we compute nDCG@10 per turn, an average within each domain, then macro-average across all 11 domains to ensure equal weight regardless of domain size. Statistical tests use paired comparisons across 23,768 turn-level observations. Full metrics appear in Appendix D.2.

#### 4.1.2 Results

Table 3 presents retrieval performance across all configurations. The average Baseline score (.236) confirms that retrieving relevant passages is difficult without context, as later-turn queries contain implicit references requiring prior conversational knowledge. History+Reasoning achieves .479, doubling Baseline performance (paired $t = 119.67$, $p < 0.001$).

Reasoning-specialized retrievers consistently outperform other architectures. DIVER achieves the highest score of .584 with History+Reasoning (95% CI: .520 to .649). ReasonIR reaches .552 (95% CI: .501 to .603). Both substantially exceed the best dense encoder, SFR at .464, confirming that models trained for multi-hop reasoning transfer effectively to conversational retrieval. Complete statistical details are provided in Appendix D.1.

#### 4.1.3 Ablation Study

A central question is how much each component contributes to retrieval improvement. Table 4 presents an ablation isolating the individual and combined effects of History and Reasoning.

| Retriever | Base | QR | Reas | Hist | H+R |
|---|---|---|---|---|---|
| | | *Reasoning-specialized* | | | |
| **DIVER** | .347 | .430 | .496 | .545 | **.584** |
| **ReasonIR** | .266 | .357 | .494 | .496 | .552 |
| | | *Dense encoders* | | | |
| **Qwen** | .269 | .345 | .399 | .425 | .461 |
| **SFR** | .240 | .324 | .396 | .429 | .464 |
| **BGE** | .230 | .328 | .347 | .411 | .445 |
| **E5** | .183 | .272 | .352 | .404 | .429 |
| **Contriever** | .168 | .232 | .303 | .366 | .409 |
| | | *Lexical* | | | |
| **BM25** | .185 | .288 | .360 | .446 | .489 |
| **Average** | .236 | .322 | .393 | .440 | **.479** |
| **95% CI** | ±.004 | ±.004 | ±.004 | ±.004 | ±.004 |

Table 3: Retrieval performance (nDCG@10, macro-averaged across 11 domains). Base=Baseline, QR=Query Rewrite, Reas=Reasoning, Hist=History, H+R=History+Reasoning. All strategy differences are significant ($p < 0.001$, paired $t$-test, $n$=23,768).

Reasoning alone improves retrieval by 67% over Baseline, helping retrievers identify passages requiring inference rather than simple keyword matching. History provides larger gains (+86%) by directly resolving coreferences and ellipsis that make later-turn queries incomplete. Combined, History+Reasoning yields +103% improvement, larger than either component alone but smaller than their sum (153%), confirming that both address some overlapping challenges while each contributes unique value.

The pattern varies across architectures. ReasonIR benefits equally from both components (+86% each), consistent with its reasoning-intensive training. BM25 benefits most from History (+141%) because lexical matching needs additional query terms. Dense encoders show balanced gains, with SFR improving +65% from Reasoning and +79% from History. Notably, all retrievers benefit when both components are combined.

#### 4.1.4 Turn Position Effects

Figure 2 illustrates retrieval performance evolution across conversation turns, aggregated across all retrievers and domains. At Turn 1, queries are self-contained. Baseline, Query Rewrite, and History

| Retriever | Base | +Reas | +Hist | H+R |
|---|---|---|---|---|
| DIVER | .347 | .496 (+43%) | .545 (+57%) | .584 (+68%) |
| ReasonIR | .266 | .494 (+86%) | .496 (+86%) | .552 (+108%) |
| BM25 | .185 | .360 (+95%) | .446 (+141%) | .489 (+164%) |
| Qwen | .269 | .399 (+48%) | .425 (+58%) | .461 (+71%) |
| SFR | .240 | .396 (+65%) | .429 (+79%) | .464 (+93%) |
| **Average** | .236 | .393 (+67%) | .440 (+86%) | .479 (+103%) |

Table 4: Ablation of Reasoning and History components (nDCG@10, 11 domains). Parentheses: % improvement over Baseline.

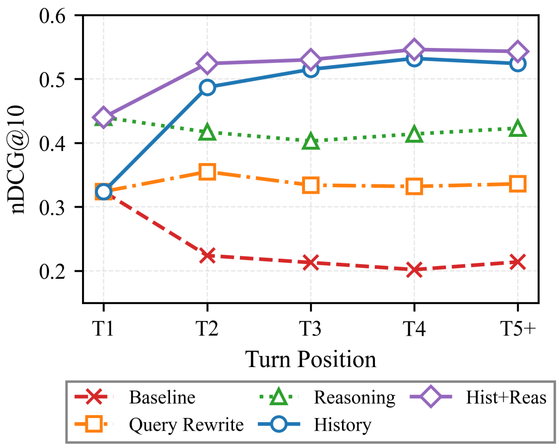


Figure 2: Retrieval performance (nDCG@10) by turn position (avg. 8 retrievers, 11 domains). Non-reasoning strategies share equal T1 performance; divergence emerges from Turn 2 as context dependence grows.

produce identical scores (.324) since no history exists. However, Reasoning-based strategies show higher T1 performance (.440) because reasoning is appended even at Turn 1, providing additional retrieval signals. From Turn 2 onward, strategy differences emerge as queries become context-dependent. Baseline performance drops sharply to .224 at Turn 2 (31% decline) and stabilizes around .20, confirming that later-turn queries cannot be interpreted in isolation.

In contrast, History-based strategies improve as conversations progress, rising from .487 at Turn 2 to .532 at Turn 4. The performance gap over Baseline widens from 117% to 163%, confirming that the benchmark captures genuine conversational complexity where accumulated context is increasingly critical. Per-retriever analysis is provided in Appendix D.3.

#### 4.1.5 Domain Variation

Retrieval difficulty varies across subject areas (Table 5). Life science domains (Psychology, Biology) achieve the highest scores (.733, .732) due to standardized terminology, while technical domains (Robotics, Hardware) show lower absolute performance but larger relative improvements (+201%, +80%), suggesting specialized vocabulary creates baseline difficulty that History+Reasoning addresses.

| Domain | Base | H+R | Δ | Best |
|---|---|---|---|---|
| Psychology | .391 | .733 | +87% | DIVER |
| Biology | .450 | .732 | +63% | DIVER |
| Earth Science | .428 | .694 | +62% | DIVER |
| Sust. Living | .405 | .682 | +68% | DIVER |
| Robotics | .195 | .587 | +201% | BM25 |
| Politics | .362 | .582 | +61% | DIVER |
| Economics | .283 | .572 | +102% | DIVER |
| Law | .331 | .540 | +63% | ReasonIR |
| Drones | .325 | .499 | +53% | ReasonIR |
| Medical Sci. | .395 | .488 | +24% | DIVER |
| Hardware | .253 | .455 | +80% | ReasonIR |

Table 5: Domain results (nDCG@10). Base: DIVER Baseline (strongest retriever). H+R: best retriever with History+Reasoning. Δ: relative improvement. Best: top retriever per domain.

| Complexity | Turns | Base | H+R | Δ |
|---|---|---|---|---|
| Low (3–4 aspects) | 463 | .428 | .813 | +90% |
| Medium (5–6 aspects) | 734 | .454 | .810 | +78% |
| High (7+ aspects) | 1,067 | .509 | .810 | +59% |

Table 6: Retrieval performance by conversation complexity (queries from Turn 2 onward, maximum nDCG@10 across all retrievers). Low-complexity conversations benefit more from History+Reasoning than high-complexity ones.

DIVER performs best in 7 of 11 domains, while ReasonIR excels in technical and legal domains (Law, Hardware, Drones) requiring procedural reasoning. BM25 achieves the highest Robotics score (.587) as specific technical terms favor exact matching. See Appendix D.4 for the complete retriever-by-domain matrix.

#### 4.1.6 Conversation Complexity

Each conversation in our benchmark covers multiple semantic aspects derived from decomposing the source answer into distinct subtopics. For example, a sustainability conversation with five aspects might discuss carbon footprint calculation, renewable energy alternatives, policy incentives, implementation barriers, and long-term environmental impact as separate threads. We examined whether complex conversations with more aspects would benefit most from History+Reasoning (Table 6).

The results reveal a counterintuitive pattern: low-complexity conversations benefit most from History+Reasoning (+90%) while high-complexity ones benefit least (+59%). Complex conversations contain richer content across turns (more entities, terminology, and concepts), giving Baseline more signals to work with, even without history. This explains why Baseline performs better on complex conversations (.509 vs. .428 for simple ones), while History+Reasoning reaches a similar ceiling ($\sim$.81) regardless of complexity. It also aligns with prior finding that context augmentation helps less when queries already contain rich information (Krasakis et al., 2023).

#### 4.1.7 Failure Analysis

We categorize all 2,971 queries by the maximum nDCG@10 achieved by any retriever with History+Reasoning. Results show 63.2% of queries achieve nDCG@10 $\geq$ 0.8, demonstrating that combining history with reasoning addresses most conversational retrieval challenges. Another 22.4% score between 0.5 and 0.8, while 8.1% score between 0.3 and 0.5. The remaining 6.3% (187 queries) are persistent failures where no configuration reaches 0.3.

Among these failures, Turn 1 queries appear more often than expected (37% of failures vs. 24% of all queries). This reveals two failure types: Turn 1 failures are inherently difficult questions where context cannot help, while later-turn failures (63%) occur when history contains relevant information but retrieval still fails.

Manual inspection reveals two primary failure modes. First, informal phrasing creates vocabulary mismatch. For example, “Is extra cooling really necessary for the 10700K?” uses colloquial terms while documents discuss “Intel Core i7-10700K thermal design power specifications.” Second, implicit domain knowledge is required. For example, “Does that shift have something to do with how the red and green cones balance each other?” requires understanding that “that shift” refers to the Purkinje effect from earlier turns. Even with complete history, retrievers fail because the reasoning chain connecting these concepts is never explicitly stated. These failures suggest the need for retrieval systems capable of implicit reasoning.

### 4.2 Generation Evaluation

#### 4.2.1 Setup

**Generators.** We evaluate seven instruction-tuned language models spanning three size categories: large (Llama-3.3-70B, Qwen2.5-72B), medium (Gemma-3-12B, Qwen2.5-14B), and small (Llama-3.1-8B, Qwen2.5-7B, Mistral-7B).

**Retrieval Modes.** We evaluate three conditions. **Oracle** provides gold-standard passages containing the correct answer, establishing an upper bound. **Retrieved** uses passages from our best retrieval configuration (DIVER with History+Reasoning, $k$=5). **NoRetrieval** generates answers without retrieved passages, relying solely on the model's parametric knowledge.

**Evaluation.** We use GPT-4o as a judge to score five dimensions on a 1–5 scale: Correctness, Completeness, Relevance, Coherence, and Faithfulness. Faithfulness evaluates whether answers are grounded in retrieved passages and applies only to Oracle and Retrieved modes. Following §4.1.1, scores are normalized to 0–1 and macro-averaged across domains. Lexical metrics in Appendix E.3.

#### 4.2.2 Results

Table 7 presents generation quality across domains and models using Retrieved mode. Complete per-domain results for all modes appear in Appendix E.1.

Llama-3.3-70B achieves the highest average score (.880) and leads in 9 of 11 domains. This dominance reflects its larger training corpus and stronger instruction-following capabilities for synthesizing retrieved information.

Life Sciences domains achieve the highest scores (.887–.900), as these domains feature well-structured factual content with clear answer patterns. In contrast, Earth Science proves most challenging (.678–.785), where complex spatiotemporal reasoning about geological processes remains difficult for current LLMs.

Small models (7–8B) score only 5–7 points below 70B+ models on average. This gap is notably smaller than in retrieval tasks (§4.1.2), suggesting that retrieved passages reduce the dependency on model scale by providing external knowledge that compensates for smaller models' limited parametric capacity.

| Domain | Llama-70B | Qwen-72B | Gemma-12B | Qwen-14B | Llama-8B | Qwen-7B | Mistral-7B |
|---|---|---|---|---|---|---|---|
| | | | *Life Sciences* | | | | |
| **Psychology** | **.900** | .839 | .852 | .862 | .843 | .809 | .841 |
| **Biology** | **.897** | .858 | .840 | .793 | .839 | .847 | .820 |
| **Medical Sci.** | **.887** | .843 | .844 | .851 | .840 | .824 | .811 |
| | | | *Physical Sciences* | | | | |
| **Earth Sci.** | .767 | **.785** | .726 | .678 | .756 | .754 | .713 |
| | | | *Technical* | | | | |
| **Hardware** | **.890** | .843 | .842 | .833 | .819 | .823 | .808 |
| **Robotics** | **.868** | .847 | .822 | .842 | .809 | .827 | .785 |
| **Drones** | .873 | .836 | **.864** | .835 | .842 | .818 | .806 |
| | | | *Social Sciences* | | | | |
| **Economics** | **.895** | .871 | .861 | .852 | .845 | .849 | .839 |
| **Law** | **.879** | .845 | .842 | .844 | .825 | .812 | .811 |
| **Politics** | **.858** | .841 | .845 | .834 | .813 | .805 | .804 |
| | | | *Applied* | | | | |
| **Sust. Living** | **.888** | .846 | .849 | .845 | .838 | .823 | .817 |
| **Average** | **.880** | .850 | .835 | .824 | .833 | .817 | .805 |

Table 7: Generation quality by domain (Retrieved mode). LLM-Judge average score. Bold indicates best model per domain.

| Model | Oracle | Retrieved | No.Retr. |
|---|---|---|---|
| | *Large* | | |
| **Llama-3.3-70B** | **.920** | **.888** | **.859** |
| **Qwen2.5-72B** | .876 | .860 | .837 |
| | *Medium* | | |
| **Gemma-3-12B** | .890 | .854 | .824 |
| **Qwen2.5-14B** | .874 | .852 | .824 |
| | *Small* | | |
| **Llama-3.1-8B** | .862 | .840 | .789 |
| **Qwen2.5-7B** | .849 | .835 | .803 |
| **Mistral-7B** | .849 | .819 | .779 |
| **Average** | **.874** | **.850** | **.816** |

Table 8: Generation quality by retrieval mode. LLM-Judge scores macro-averaged across 11 domains. NoRetrieval excludes Faithfulness.

#### 4.2.3 Retrieval Mode Analysis

Table 8 compares generation quality across retrieval conditions.

Oracle achieves .874 average, confirming that LLMs generate high-quality answers when given relevant context. The gap between Oracle and Retrieved (2.4 points) represents the cost of imperfect retrieval, while the gap between Retrieved and NoRetrieval (3.4 points) confirms that RAG improves answer quality. The RAG benefit varies by model size: smaller models gain more (Llama-3.1-8B: +5.1 points) than larger models (Llama-3.3-70B: +2.9 points), because smaller models have weaker parametric knowledge and thus rely more on external retrieval.

#### 4.2.4 Fine-grained Analysis

**Retrieval-generation correlation.** Table 9 examines whether better retrieval leads to better generation. Results show moderate correlation ($r$ = .42): high-retrieval domains (Psychology, Biology) achieve high generation scores, but the relationship is not deterministic. Hardware achieves .851 generation despite low retrieval (.455), while Earth Science scores only .740 despite high retrieval (.694). This aligns with findings from BRIGHT (Su et al., 2024), where generation quality does not always reflect retrieval performance due to the generator's varying ability to integrate retrieved content across domains.

**Turn position effects.** As shown in Figure 3, generation quality drops at later turns: in Retrieved mode, Llama-3.3-70B declines 4.1 points from T1 to T5+. In contrast, passage retrieval with History+Reasoning slightly improves across turns (§4.1.4), suggesting that accumulated context benefits retrievers more than generators.

**Dimension analysis.** Across all models, Coherence scores highest (.895–.940), showing that LLMs produce fluent responses. Completeness scores lowest (.750–.846), indicating that covering all aspects of complex questions remains challenging. (See Appendix E.2 for details)

## 5 Conclusion

This work introduces the first benchmark combining multi-turn dialogue with reasoning-intensive retrieval. Our Decomposition and Verification framework generates fact grounded conversations through multi-level validation. Evaluation confirms conversational context is critical: History+Reasoning doubles baseline performance (.479 vs .236 nDCG@10). reasoning-specialized models outperform dense retrievers (DIVER .584 vs SFR .464), showing that multi-hop reasoning generalizes to conversation. However, implicit rea-

| Domain | nDCG@10 | LLM-Judge |
|---|---|---|
| Psychology | .733 | .856 |
| Biology | .732 | .856 |
| Economics | .572 | .867 |
| Sust. Living | .682 | .858 |
| Medical Sci. | .488 | .857 |
| Hardware | .455 | .851 |
| Law | .540 | .851 |
| Drones | .499 | .848 |
| Robotics | .587 | .843 |
| Politics | .582 | .843 |
| Earth Science | .694 | .740 |

Table 9: Retrieval-generation correlation by domain. nDCG@10 from DIVER H+R. LLM-Judge averaged across all 7 generators. Pearson $r$ = .42.

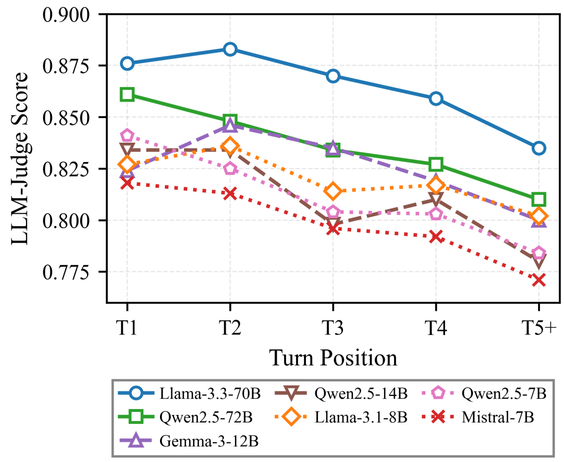


Figure 3: Generation quality by turn position across 7 models (Retrieved mode). Large models (solid) outperform small (dotted). All show significant decline ($p < 0.01$).

soning dependent on unstated domain knowledge remains challenging. These findings highlight current capabilities and limitations, directing future research in reasoning-guided conversational retrieval.

## Limitations

**(1) Language coverage.** The benchmark covers English only. Extending to other languages would require adapting the verification pipeline and collecting multilingual domain corpora.
**(2) Domain scope.** We cover eleven domains spanning scientific, technical, and social topics, but areas such as mathematics and commonsense reasoning remain unexplored and may exhibit different performance patterns.
**(3) Source dependency.** Our framework requires existing question-answer pairs with supporting documents. Domains lacking such structured resources would require additional annotation.

## Acknowledgements

The authors would like to acknowledge the financial support provided by the Austrian Research Agency (FFG) for the project "AI Enabled Sustainability Jurisdiction Demonstrator" (project No. 915229).

## References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, Tao Sheng, Sujith Ravi, and Dan Roth. 2025. Aligning LLMs for multilingual consistency in enterprise applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 117–137, Suzhou (China). Association for Computational Linguistics.

Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. 2024. Trec ikat 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 819–829.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pages 2318–2335.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of* the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025. Coral: Benchmarking multi-turn conversational retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1308–1330.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2174–2184.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1985–1988.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Antonios Minas Krasakis, Andrew Yates, and Evangelos Kanoulas. 2023. Contextualizing and expanding conversational queries without supervision. *ACM Transactions on Information Systems*, 42(3):1–30.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yecheng Luo, Yue Shen, Jian Wang, Hualei Zhou, Chunxiao Guo, Peng Wei, and 1 others. 2025. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Xuan-Phi Nguyen, Shrey Pandit, Senthil Purushwalkam, Austin Xu, Hailin Chen, Yifei Ming, Zixuan Ke, Silvio Savarese, Caiming Xong, and Shafiq Joty. 2024. Sfr-rag: Towards contextually faithful llms. *arXiv preprint arXiv:2409.09916*.

OpenAI. 2024. Gpt-4o system card.

OpenAI. 2025. GPT-4.1.

Priyaranjan Pattnayak, Amit Agarwal, Hansa Meghwani, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hybrid ai for responsive multi-turn online conversations with novel dynamic routing and feedback adaptation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 215–229.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and 1 others. 2025. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, and 1 others. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze D Gui, Ziran W Jiang, Ziyu Jiang, and 1 others. 2024. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on research and development in information retrieval*, pages 829–838.

## A Extended Related Work

**Conversational Information Retrieval.** Conversational search benchmarks have established foundations for multi-turn retrieval evaluation. QReCC (Anantha et al., 2021) provides 14K conversations emphasizing question rewriting to resolve coreferences and ellipses. TREC iKAT (Aliannejadi et al., 2024) adds personalization through user-specific knowledge bases over 116 million passages. MTRAG (Katsis et al., 2025) offers 110 human-generated conversations with challenges including unanswerable questions. The TREC Conversational Assistance Track (CAsT) (Dalton et al., 2020) introduced conversational passage retrieval with artifacts such as ellipsis and anaphora, driving advances in conversational dense retrieval (Yu et al., 2021). These benchmarks evaluate conversational dynamics but assume relevance depends on semantic similarity rather than reasoning.

**Multi-turn RAG and Conversational QA.** Additional benchmarks target multi-turn retrieval-augmented generation. CORAL (Cheng et al., 2025) provides large-scale conversations from Wikipedia supporting retrieval, generation, and citation tasks. Doc2Dial (Feng et al., 2020) creates goal-oriented dialogues grounded in government documents. OR-QuAC (Qu et al., 2020) extends QuAC to open-retrieval settings.

Foundational conversational QA datasets include CoQA (Reddy et al., 2019) with 127K questions featuring coreference and pragmatic reasoning, QuAC (Choi et al., 2018) with 14K information-seeking dialogues, and TopiOCQA (Adlakha et al., 2022) introducing topic switching. These datasets evaluate context resolution but do not require reasoning-intensive retrieval.

**Reasoning-Intensive Retrieval.** BRIGHT (Su et al., 2024) revealed that retrievers achieving 59.0 nDCG@10 on standard benchmarks drop to 18.3 on reasoning-intensive queries where relevant documents share no surface terms with queries. Chain-of-thought augmentation improves performance by up to 12.2 points. However, BRIGHT evaluates only single-turn queries, leaving multi-turn reasoning-intensive retrieval unexplored.

**RAG Evaluation.** RAG benchmarks assess grounded generation quality or answer correctness (Agarwal et al., 2025). RGB (Chen et al., 2024b) tests noise robustness and information integration. RAGBench (Friel et al., 2024) provides fine-grained annotations. CRAG (Yang et al., 2024) covers diverse question types across domains. These primarily evaluate single-turn interactions.

**Atomic Fact Verification.** Our verification pipeline relates to factuality evaluation methods. FActScore (Min et al., 2023) decomposes generations into atomic facts validated against Wikipedia. SAFE (Wei et al., 2024) extends this with search-augmented verification. SelfCheckGPT (Manakul et al., 2023) detects hallucinations through sampling consistency. Unlike these post-hoc methods, we use atomic fact extraction during construction to ensure each turn is grounded in source documents.

**Multi-hop Reasoning.** Multi-hop QA benchmarks require synthesizing information across documents. HotpotQA (Yang et al., 2018) provides 113K questions with sentence-level supporting facts. 2WikiMultiHopQA (Ho et al., 2020) constructs 2-hop and 4-hop questions via shared entities. MuSiQue (Trivedi et al., 2022) addresses shortcut exploitation through bottom-up composition. These focus on comprehension given retrieved passages; BRIGHT and our work evaluate whether retrievers can identify relevant passages when relevance requires reasoning.

**Positioning Our Work.** Existing benchmarks address subsets of the challenges in complex information seeking. Conversational benchmarks (QReCC, CAsT, MTRAG, CORAL) evaluate multi-turn dynamics without reasoning-intensive retrieval. BRIGHT requires reasoning but operates single-turn. Multi-hop QA benchmarks (HotpotQA, MuSiQue) require reasoning but assume gold retrieval.

Our benchmark uniquely combines: (1) multi-turn conversation with natural dialogue flow, (2) reasoning-intensive retrieval where relevance requires inference, (3) fact-grounded responses verified against source documents, and (4) explicit retrieval reasoning extending BRIGHT's insights to conversational settings. This combination tests whether systems can maintain both conversational coherence and reasoning-based retrieval across extended interactions—essential for supporting complex real-world information needs.

## B Detailed Dataset Statistics

### B.1 Turn Distribution

Conversation lengths range from 3 to 12 turns, with the distribution shown in Table 10. Most conversations contain 3–4 turns, while complex topics requiring extensive multi-step reasoning extend to longer dialogues.

| Turns | Count | Percentage |
|---|---|---|
| 3 | 296 | 41.9% |
| 4 | 195 | 27.6% |
| 5 | 108 | 15.3% |
| 6 | 50 | 7.1% |
| 7+ | 58 | 8.2% |

Table 10: Turn distribution across all 707 conversations (aggregated from 11 domains).

### B.2 Aspect Distribution

Each conversation contains multiple distinct semantic aspects extracted from the original answer. Table 11 shows the distribution across conversations, demonstrating that most conversations contain 5–7 aspects, with variation based on answer complexity.

| Aspects | Count | Percentage |
|---|---|---|
| 3 | 82 | 11.6% |
| 4 | 121 | 17.1% |
| 5 | 161 | 22.8% |
| 6 | 100 | 14.1% |
| 7 | 93 | 13.2% |
| 8+ | 150 | 21.2% |

Table 11: Distribution of semantic aspects per conversation across all 707 conversations (aggregated from 11 domains).

### B.3 Document Diversity per Turn

Supporting documents per turn range from single focused sources to synthesis across multiple documents. Table 12 shows this natural distribution, where simpler aspects may be addressed with one document while complex reasoning requires multiple sources.

| Docs/Turn | Count | Percentage |
|---|---|---|
| 1 | 1,667 | 56.1% |
| 2 | 527 | 17.7% |
| 3 | 312 | 10.5% |
| 4 | 248 | 8.3% |
| 5+ | 217 | 7.3% |

Table 12: Distribution of supporting documents per turn across all 2,971 turns (aggregated from 11 domains)

## C Pipeline Illustration

This appendix illustrates our Decomposition-and-Verification framework through a complete example, tracing a single source query through all pipeline stages to produce a grounded multi-turn conversation.

### C.1 Source Material

**Original Query.**

> *"Why does evolution not make our life longer? Wouldn't evolution favour a long life?"*

**Reasoning Annotation.**

> *"The question requires understanding why natural selection does not optimize for longevity. Relevant documents must explain evolutionary theories of aging, including mutation accumulation, antagonistic pleiotropy, and life-history trade-offs."*

### C.2 Stage 1: Answer Decomposition

The expert answer addresses multiple evolutionary mechanisms. Our pipeline identifies six distinct aspects, each suitable for a conversational turn.

| # | Aspect | Type | Cov. |
|---|---|---|---|
| 1 | Post-reproductive selection | Mechanism | 0.85 |
| 2 | Mutation accumulation | Mechanism | 0.90 |
| 3 | Late-onset disease example | Example | 0.80 |
| 4 | Antagonistic pleiotropy | Mechanism | 0.85 |
| 5 | Disposable soma theory | Mechanism | 0.75 |
| 6 | Extrinsic mortality | Implication | 0.80 |

Table 13: Aspects identified from the source answer.

### C.3 Stage 2–3: Fact Extraction and Verification

From each aspect, we extract atomic facts and verify them against source documents using semantic similarity. Facts without clear document support are discarded. Aspects with no verified facts are excluded from conversation generation.

Tables 13, 14, and 15 illustrate the pipeline progression. Table 13 shows the six aspects decomposed from the source answer (Stage 1), along with a coverage score (Cov.) indicating how much of each aspect is topically addressed by source documents. Table 14 shows the atomic facts extracted from each aspect and their verification status (Stages 2–3).

| Aspect | Extracted Facts | Verified | Rate |
|---|---|---|---|
| Post-reproductive selection | (1) Selection acts primarily on pre-reproductive traits; (2) Post-reproductive mutations face weak selection pressure | 2/2 | 100% |
| Mutation accumulation | (1) Deleterious late-onset alleles accumulate over generations | 1/1 | 100% |
| Late-onset disease example | (1) Huntington's manifests between ages 30–50; (2) Carriers reproduce before experiencing symptoms | 2/2 | 100% |
| Antagonistic pleiotropy | (1) Certain alleles confer early-life benefits but late-life costs | 1/1 | 100% |
| Disposable soma theory | (1) Organisms allocate resources between reproduction and repair; (2) High repair investment yields longer lifespan | 0/2 | 0% |
| Extrinsic mortality | (1) High environmental mortality favors early reproduction | 1/1 | 100% |
| **Total** | | **7/9** | **78%** |

Table 14: Fact extraction and verification results. Each row shows the atomic facts extracted from an aspect and whether they were verified against source documents.

As shown in Table 14, the disposable soma theory aspect yielded no verified facts and is therefore excluded from the generated conversation. This illustrates an important distinction: high coverage does not guarantee fact verification. The disposable soma theory has a coverage score of 0.75 in Table 13, meaning the source documents discuss this topic. However, its verification rate is 0% because the two specific claims extracted from the gold answer could not be confirmed in those documents. The remaining five aspects form the basis for the five-turn dialogue.

Table 15 demonstrates document scoring for Turn 4, where candidate documents are evaluated based on their support for the verified facts (Stage 4).

## C.4 Stage 4: Document Scoring

Candidate documents are scored using:

$$\text{score} = 0.5S_\text{s} + 0.3S_\text{c} + 0.15S_\text{l} - 0.05S_\text{m}$$

where $S_\text{s}$, $S_\text{c}$, $S_\text{l}$, $S_\text{m}$ denote support, completeness, clarity, and misleading scores respectively.

| Document | $S_\text{s}$ | $S_\text{c}$ | $S_\text{l}$ | $S_\text{m}$ | Score | Sel. |
|---|---|---|---|---|---|---|
| Aging_theory_12.txt | 8 | 7 | 7 | 0 | 7.15 | ✓ |
| Evolution_38.txt | 6 | 5 | 6 | 0 | 5.40 | ✓ |
| Genetics_91.txt | 4 | 3 | 5 | 0 | 3.65 | — |

Table 15: Document scoring example (Turn 4). Documents above threshold are selected.

## C.5 Generated Conversation

The pipeline produces the following five-turn conversation from the five aspects with verified facts.

**Turn 1** *Post-reproductive selection*
**Q:** In evolution, why doesn't natural selection care much about traits that show up after we've had kids?
**A:** Natural selection acts primarily on traits affecting survival and reproduction before the end of the reproductive period. Mutations with effects manifesting only after reproduction face substantially weaker selection pressure, since individuals have already transmitted their genes.

**Turn 2** *Mutation accumulation*
**Q:** Is that why some harmful genes stick around, just because they show up later in life?
**A:** Precisely. This is the basis of mutation accumulation theory. Deleterious alleles with late-onset effects are not efficiently removed by selection because their harmful consequences occur after reproduction. Over generations, such mutations accumulate in the population.

**Turn 3** *Late-onset example*
**Q:** Are there examples of genetic diseases that stick around because they only appear after people have children?
**A:** Huntington's disease is a classic example. Symptoms typically manifest between ages 30 and 50, often after individuals have had children. Because carriers reproduce before experiencing symptoms, selection against the causative allele is weak.

**Turn 4** *Antagonistic pleiotropy*
**Q:** What about genes that actually help us when we're young but cause problems when we're older?
**A:** This describes antagonistic pleiotropy, proposed by George Williams in 1957. Certain alleles confer fitness benefits early in life but produce deleterious effects later. Selection favours these alleles because early-life benefits outweigh late-life costs.

**Turn 5** *Extrinsic mortality*
**Q:** Why wouldn't living longer be advantageous if there's a lot of danger in the environment?
**A:** When extrinsic mortality is high, few individuals survive to benefit from extended longevity. Under such conditions, selection favours early reproduction over investment in long-term survival, explaining why high-mortality species typically have shorter lifespans.

## C.6 Reasoning Requirements

Each turn requires reasoning to connect the question to relevant documents. Table 16 shows the inference needed for each turn.

These reasoning requirements are inherent to the domain. A retrieval system must recognize,

| Turn | Question | Document Topic | Reasoning Required |
|---|---|---|---|
| 1 | Traits appearing after having children | Post-reproductive selection dynamics | Recognize that timing of trait expression determines selection intensity |
| 2 | Harmful genes persisting in populations | Late-acting mutation accumulation | Infer that weak selection permits mutation buildup over generations |
| 3 | Diseases appearing after reproduction | Huntington's disease onset patterns | Identify specific disease as instance of the general pattern |
| 4 | Genes helping early, harming later | Antagonistic pleiotropy | Map everyday description to formal evolutionary concept |
| 5 | Longevity despite environmental danger | Extrinsic mortality effects | Connect environmental risk to selection on lifespan |

Table 16: Reasoning required for retrieval. Finding relevant documents requires understanding the conceptual relationship between what the user asks and what documents discuss.

for example, that documents about "antagonistic pleiotropy" are relevant to a question about "genes that help early but cause problems later." This requires understanding the conceptual relationship between the question and the document topic—a process that goes beyond surface-level matching.

### C.7 Retrieval Reasoning Annotation

Each turn includes annotations specifying what information relevant documents should contain. Example for Turn 4:

```
{
  "target": "Genes with age-dependent
             opposing fitness effects",
  "relevance_signals": [
    "antagonistic pleiotropy",
    "early-life benefits with late-life
         costs",
    "Williams 1957 aging theory"
  ],
  "irrelevance_signals": [
    "neutral mutations",
    "single-effect alleles"
  ]
}
```

## D Additional Retrieval Results

### D.1 Statistical Details

Table 17 presents detailed statistical comparisons referenced in §4.1.2.

| Comparison | Mean Δ | SE | $t$ | $p$ |
|---|---|---|---|---|
| QR vs. Baseline | +0.086 | 0.0016 | 61.1 | <0.001 |
| Reas vs. Baseline | +0.157 | 0.0019 | 92.9 | <0.001 |
| Hist vs. Baseline | +0.204 | 0.0022 | 102.2 | <0.001 |
| H+R vs. Baseline | +0.243 | 0.0022 | 119.7 | <0.001 |
| H+R vs. History | +0.039 | 0.0013 | 30.2 | <0.001 |

Table 17: Paired t-tests ($n = 23,768$). Mean Δ: macro-averaged nDCG@10 difference across domains. SE: std. error.

### D.2 Additional Metrics

Table 18 presents MAP@10, Recall@10, and MRR for History+Reasoning.

| Retriever | nDCG@10 | MAP@10 | Recall@10 | MRR |
|---|---|---|---|---|
| DIVER | .584 | .488 | .753 | .592 |
| ReasonIR | .552 | .459 | .713 | .566 |
| BM25 | .489 | .406 | .627 | .508 |
| SFR | .464 | .374 | .627 | .470 |
| Qwen | .461 | .368 | .633 | .464 |
| BGE | .445 | .356 | .595 | .463 |
| E5 | .429 | .350 | .573 | .437 |
| Contriever | .409 | .325 | .552 | .428 |

Table 18: Full metrics for History+Reasoning (all @10, macro-averaged across 11 domains).

### D.3 Turn Position by Retriever

Table 19 shows per-retriever turn position effects referenced in §4.1.4.

| Retriever | T2 | T3 | T4 | T5+ | Trend |
|---|---|---|---|---|---|
| *History+Reasoning* | | | | | |
| DIVER | .629 | .638 | .641 | .632 | +2% |
| ReasonIR | .593 | .593 | .615 | .600 | +4% |
| BM25 | .580 | .574 | .587 | .572 | +1% |
| Qwen | .503 | .488 | .529 | .528 | +5% |
| *Baseline* | | | | | |
| DIVER | .339 | .327 | .320 | .333 | −6% |
| ReasonIR | .244 | .240 | .234 | .237 | −4% |
| BM25 | .192 | .170 | .151 | .162 | −21% |
| Qwen | .256 | .250 | .252 | .267 | −2% |

Table 19: Turn position by retriever. Trend shows T2→T4 change. With History+Reasoning, all retrievers maintain or improve at later turns. With Baseline, all retrievers degrade.

### D.4 Complete Domain Results

Table 20 presents the full domain-by-retriever matrix referenced in §4.1.5.

| Domain | BGE | BM25 | Contr. | DIVER | E5 | Qwen | ReasIR | SFR |
|---|---|---|---|---|---|---|---|---|
| Biology | .463 | .722 | .459 | **.732** | .589 | .636 | .636 | .554 |
| Drones | .423 | .300 | .366 | .491 | .296 | .340 | **.499** | .359 |
| Earth Sci. | .531 | .687 | .480 | **.694** | .595 | .545 | .621 | .558 |
| Economics | .419 | .526 | .402 | **.572** | .496 | .383 | .527 | .507 |
| Hardware | .356 | .345 | .280 | .433 | .336 | .362 | **.455** | .358 |
| Law | .373 | .337 | .376 | .490 | .282 | .416 | **.540** | .373 |
| Medical | .421 | .203 | .380 | **.488** | .187 | .392 | .400 | .352 |
| Politics | .463 | .406 | .459 | **.582** | .309 | .478 | .557 | .391 |
| Psychology | .520 | .649 | .478 | **.733** | .588 | .601 | .695 | .591 |
| Robotics | .381 | **.587** | .358 | .529 | .489 | .355 | .526 | .469 |
| Sust. Living | .541 | .612 | .458 | **.682** | .552 | .568 | .618 | .586 |

Table 20: Domain-retriever matrix (nDCG@10, History+Reasoning). Bold indicates best retriever per domain.

| Domain | Llama-70B | Qwen-72B | Gemma-12B | Qwen-14B | Llama-8B | Qwen-7B | Mistral-7B |
|---|---|---|---|---|---|---|---|
| Psychology | **.932** | .878 | .895 | .889 | .870 | .856 | .862 |
| Biology | **.925** | .882 | .891 | .867 | .865 | .860 | .855 |
| Medical Sci. | **.918** | .871 | .886 | .875 | .862 | .848 | .845 |
| Earth Science | .812 | **.824** | .778 | .752 | .795 | .788 | .762 |
| Hardware | **.922** | .869 | .880 | .862 | .851 | .849 | .840 |
| Robotics | **.905** | .873 | .865 | .868 | .845 | .852 | .822 |
| Drones | **.912** | .862 | .898 | .865 | .871 | .846 | .838 |
| Economics | **.928** | .895 | .902 | .882 | .873 | .871 | .865 |
| Law | **.915** | .872 | .881 | .870 | .856 | .842 | .844 |
| Politics | **.896** | .868 | .879 | .862 | .848 | .835 | .838 |
| Sust. Living | **.919** | .870 | .885 | .872 | .862 | .849 | .848 |
| **Average** | **.920** | .876 | .890 | .874 | .862 | .849 | .849 |

Table 21: Generation quality by domain (Oracle mode). LLM-Judge average score. Bold indicates best model per domain.

# E Additional Generation Results

## E.1 Per-Domain Generation Quality by Mode

Table 21 presents generation quality for Oracle mode across all domains.

Table 22 presents generation quality for NoRetrieval mode across all domains.

## E.2 Dimension Breakdown

Table 23 presents dimension-level scores for all models using Retrieved mode.

Table 24 presents dimension scores by domain for Llama-3.3-70B.

## E.3 Lexical Metrics

Table 25 presents lexical evaluation metrics for Retrieved mode.

Lexical metrics show lower absolute values than LLM-Judge scores due to the open-ended nature of conversational answers. METEOR correlates most strongly with LLM-Judge ($r$ = .72), while ROUGE-L shows weaker correlation ($r$ = .54).

## E.4 Model Size Comparison

Table 26 compares models within the same family across all three modes.

Within Qwen2.5, scaling from 7B to 72B improves Retrieved scores by 2.5 points. Llama shows larger scaling effects: 70B outperforms 8B by 4.8 points. The RAG benefit decreases with model size, confirming that larger models' parametric knowledge partially substitutes for retrieval.

| Domain | Llama-70B | Qwen-72B | Gemma-12B | Qwen-14B | Llama-8B | Qwen-7B | Mistral-7B |
|---|---|---|---|---|---|---|---|
| Psychology | **.872** | .815 | .822 | .828 | .762 | .778 | .758 |
| Biology | **.868** | .832 | .810 | .765 | .758 | .812 | .742 |
| Medical Sci. | **.862** | .818 | .815 | .822 | .765 | .795 | .738 |
| Earth Science | .698 | **.725** | .658 | .612 | .682 | .688 | .645 |
| Hardware | **.865** | .820 | .812 | .805 | .752 | .792 | .745 |
| Robotics | **.842** | .822 | .792 | .815 | .745 | .798 | .722 |
| Drones | **.848** | .812 | .835 | .808 | .778 | .788 | .742 |
| Economics | **.872** | .848 | .832 | .825 | .782 | .818 | .778 |
| Law | **.855** | .822 | .815 | .818 | .762 | .785 | .752 |
| Politics | **.835** | .818 | .818 | .808 | .752 | .778 | .748 |
| Sust. Living | **.862** | .822 | .822 | .818 | .772 | .795 | .758 |
| **Average** | **.859** | .837 | .824 | .824 | .789 | .803 | .779 |

Table 22: Generation quality by domain (NoRetrieval mode). LLM-Judge average score. Bold indicates best model per domain. Faithfulness excluded.

| Model | Corr | Comp | Rel | Coh | Faith | Avg |
|---|---|---|---|---|---|---|
| Llama-3.3-70B | .870 | .846 | .909 | .940 | .875 | **.888** |
| Qwen2.5-72B | .848 | .781 | .896 | .929 | .846 | .860 |
| Gemma-3-12B | .840 | .784 | .879 | .916 | .849 | .854 |
| Qwen2.5-14B | .840 | .776 | .888 | .920 | .838 | .852 |
| Llama-3.1-8B | .816 | .787 | .864 | .911 | .823 | .840 |
| Qwen2.5-7B | .816 | .760 | .868 | .912 | .820 | .835 |
| Mistral-7B | .799 | .750 | .847 | .895 | .806 | .819 |
| **Average** | .833 | .783 | .879 | .918 | .837 | .850 |

Table 23: Generation quality by dimension (Retrieved mode). Corr=Correctness, Comp=Completeness, Rel=Relevance, Coh=Coherence, Faith=Faithfulness. Scores macro-averaged across 11 domains.

| Domain | Corr | Comp | Rel | Coh | Faith |
|---|---|---|---|---|---|
| Psychology | .885 | .862 | .918 | .948 | .888 |
| Biology | .882 | .858 | .915 | .945 | .885 |
| Medical Sci. | .872 | .848 | .905 | .942 | .872 |
| Earth Science | .752 | .728 | .795 | .885 | .778 |
| Hardware | .875 | .852 | .908 | .942 | .878 |
| Robotics | .855 | .832 | .888 | .935 | .865 |
| Drones | .858 | .835 | .892 | .938 | .868 |
| Economics | .880 | .858 | .912 | .945 | .882 |
| Law | .865 | .842 | .898 | .938 | .872 |
| Politics | .845 | .822 | .878 | .932 | .858 |
| Sust. Living | .872 | .850 | .905 | .940 | .875 |
| **Average** | .870 | .846 | .909 | .940 | .875 |

Table 24: Dimension scores by domain (Llama-3.3-70B, Retrieved). Corr=Correctness, Comp=Completeness, Rel=Relevance, Coh=Coherence, Faith=Faithfulness.

| Model | ROUGE-L | METEOR | BERT-F1 |
|---|---|---|---|
| Llama-3.3-70B | .274 | **.345** | .365 |
| Qwen2.5-72B | **.283** | .264 | **.381** |
| Gemma-3-12B | .258 | .270 | .351 |
| Qwen2.5-14B | .283 | .269 | .375 |
| Llama-3.1-8B | .276 | .326 | .355 |
| Qwen2.5-7B | .275 | .271 | .367 |
| Mistral-7B | .272 | .286 | .352 |

Table 25: Lexical metrics (Retrieved mode). Macro-averaged across 11 domains. Bold indicates best per metric.

| Model | Oracle | Retr. | No.Retr. | RAG$\Delta$ |
|---|---|---|---|---|
| *Qwen2.5 family* | | | | |
| Qwen2.5-72B | .876 | .860 | .837 | +2.3 |
| Qwen2.5-14B | .874 | .852 | .824 | +2.8 |
| Qwen2.5-7B | .849 | .835 | .803 | +3.2 |
| *Llama family* | | | | |
| Llama-3.3-70B | .920 | .888 | .859 | +2.9 |
| Llama-3.1-8B | .862 | .840 | .789 | +5.1 |

Table 26: Model size comparison by family. RAG$\Delta$ = Retrieved − NoRetrieval ($\times$100). Scores macro-averaged across 11 domains.

## F Pipeline Prompts

We use GPT-4.1 (OpenAI, 2025) for all conversation generation steps. This appendix provides the complete prompts used in our Decomposition-and-Verification framework.

**Prompt for Document-Answer Alignment Validation**

```
You are evaluating whether documents can support an answer.
GOLD ANSWER:
{gold_answer}
DOCUMENTS:
{documents}
IMPORTANT:

  • Documents don't need ALL details

  • Core facts are enough

  • Related information counts as support

TASK:

  1. Identify KEY CLAIMS (focus on main ideas, not minor details, max 3-5 claims)

  2. Check if documents contain supporting information

  3. Return coverage percentage

Return JSON:
{
"key_claims": ["main claim 1", "main claim 2", ...],
"supported_claims": ["claim1", ...],
"unsupported_claims": ["claim2", ...],
"coverage_percentage": 0.0-1.0,
"is_sufficient": boolean
}
```

Figure 4: Prompt for validating document-answer alignment.

Prompt for Granular Aspect Extraction

```
Extract {num_aspects} distinct, GRANULAR aspects from the gold answer.
QUERY: {query}
REASONING (why this answer is relevant):
{reasoning}
GOLD ANSWER:
{gold_answer}
ALREADY EXTRACTED ASPECTS (do NOT duplicate these):
{existing_aspects}
WHAT IS A GRANULAR ASPECT?
An aspect does NOT need to be a broad topic. It can be a specific detail, a single step in a
process, or a distinct implication.

  • Broad (AVOID): "How Photosynthesis Works" (Too big, covers everything)

  • Granular (GOOD): "Role of Chlorophyll in Light Absorption"

  • Granular (GOOD): "The Calvin Cycle's Carbon Fixation Step"

ASPECT REQUIREMENTS:

  1. SPECIFICITY: Drill down into details. A single sentence with a verified fact can be an
     aspect.

  2. VERBATIM EXCERPT: Copy exact text from gold answer.

  3. SUBSTANTIVE: Must contain facts, mechanisms, or examples.

  4. DISTINCT ANGLE: If a topic is already covered, look for a specific implication, limitation,
     or counter-example.

ASPECT TYPES:
"detail" | "step" | "implication" | "distinction" | "definition" | "mechanism" | "example" |
"comparison" | "history" | "application"
Return JSON:
{
"aspects": [
{
"aspect_name": "Specific Name (3-6 words)",
"aspect_type": "detail|step|implication|...",
"excerpt": "Exact verbatim text from gold answer",
"distinct_from_existing": "How this differs from existing aspects"
}
],
"extraction_notes": "Notes on finding distinct details"
}
```

Figure 5: Prompt for extracting granular aspects from gold answers.

Prompt for Atomic Fact Extraction and Verification

```
Extract key facts from the EXCERPT, then verify each against the DOCUMENTS.
ASPECT EXCERPT:
{aspect_excerpt}
DOCUMENTS:
{documents}
TASK:

  1. Extract 1 to 5 key facts from the excerpt (depending on length).

       • If the excerpt is a single detail, extract just 1 fact.
       • Do NOT invent facts to fill a quota.

  2. For each fact, check if it's supported by the documents.

  3. Return ONLY facts that are supported.

A fact is "supported" if:

   • Explicitly stated in documents, OR

   • Clearly implied by combining document information.

Return JSON:
{
"extracted_facts": [
{
"fact": "the extracted fact (5-15 words)",
"is_supported": boolean,
"supporting_doc_id": "doc_X" or null,
"reason": "brief reason"
}
],
"supported_facts": ["fact1", "fact2", ...],
"unsupported_facts": ["fact3", ...],
"summary": {"total_extracted": integer, "supported_count": integer}
}
```

Figure 6: Prompt for extracting and verifying atomic facts against source documents.

## Prompt for Aspect Overlap Detection

Does the NEW ASPECT cover content already covered by EXISTING ASPECTS?
**NEW ASPECT:** {aspect_name} ({aspect_type})
**CONTENT:** {excerpt}
**EXISTING ASPECTS:**
{existing_aspects_text}
**WHAT COUNTS AS OVERLAP?**
Two aspects OVERLAP if they:

- Make the EXACT SAME factual claims.
- Explain the EXACT SAME step of a process.
- Are merely rephrasing the same information.

**WHAT IS NOT OVERLAP? (DISTINCT CONTENT):**

- Different steps of the same mechanism (e.g., "Step 1: Input" vs "Step 2: Processing")
- Different examples of the same concept
- Specific details vs General definitions
- Implications vs Mechanisms
- "What it is" vs "How it works" vs "Why it matters"

**DECISION:** Does the new aspect contain specific details, steps, or implications NOT in the existing aspects?
**Return JSON:**

```
{
"has_overlap": boolean,
"overlaps_with": "name of overlapping aspect" or null,
"overlap_type": "same_claims" | "same_examples" | "same_step" | "no_overlap",
"reasoning": "one sentence explanation"
}
```

Figure 7: Prompt for detecting content overlap between aspects.

## Prompt for Aspect Ordering by Logical Progression

```
Order these topic aspects for natural conversation progression.
CONVERSATION STRATEGY: {conversation_strategy}
ASPECTS (format: [type] name):
{aspect_summaries}
ORDERING PRINCIPLES:
1. Foundational/definitional aspects first (what something IS)
2. Then mechanisms/processes (how it WORKS)
3. Then historical context or causes (WHY/WHEN)
4. Then specific examples/cases (INSTANCES)
5. End with implications/modern relevance (SO WHAT)
EXAMPLE GOOD ORDER:
1. [factual] "Definition of nation vs state"
   → first (foundational)
2. [explanatory] "Distinction between political/cultural"
   → second
3. [historical] "Colonial impact on borders"
   → third
4. [factual] "Stateless nations examples"
   → fourth
5. [explanatory] "Modern nation-state challenges"
   → last (implications)
Return JSON:
{
  "ordered_indices": [0, 2, 1, 3],
  "reasoning": "Brief explanation of ordering logic"
}
```

Figure 8: Prompt for reordering aspects to ensure natural conversational flow.

## Prompt for Sub-question Generation

```
Generate a focused sub-question for this aspect.
QUERY: {query}
OVERALL REASONING: {overall_reasoning}
ASPECT: {aspect_name}
TYPE: {aspect_type}
KEY FACTS:
{semantic_facts}
PREVIOUS SUB-QUESTIONS:
{previous_subquestions}
Generate a sub-question that:

  1. Aligns with the overall reasoning and query intent

  2. Specifically targets this aspect

  3. Is distinct from previous sub-questions

  4. Can be answered using the facts listed

Return JSON:
{
"sub_question": "The focused sub-question?",
"confidence": 0.0-1.0,
"reasoning": "Why this targets this aspect AND aligns with overall reasoning"
}
```

Figure 9: Prompt for generating focused sub-questions targeting specific aspects.

Prompt for Document Scoring

```
Score how well each document helps answer the sub-question.
SUB-QUESTION: {sub_question}
RETRIEVAL REASONING: {retrieval_reasoning}
KEY FACTS TO COVER:
{semantic_facts}
CANDIDATE DOCUMENTS:
{candidate_docs}
SCORING CRITERIA (each 0-10):
```

- **support_score**: How well does the document support answering the sub-question?
- **completeness_score**: Does the document cover the key facts listed above?
- **clarity_score**: Is the information clearly presented and easy to extract?
- **misleading_score**: Does the document contain potentially misleading information? (PENALTY)

**Final score formula:**
score $= 0.5 \times S_s + 0.3 \times S_c + 0.15 \times S_l - 0.05 \times S_m$
**Return JSON:**

```
{
"document_scores": [
{
"doc_id": "doc_0",
"support_score": 8,
"completeness_score": 7,
"clarity_score": 9,
"misleading_score": 0,
"final_score": 7.85,
"reasoning": "Brief justification"
}
]
}
```

Figure 10: Prompt for scoring candidate documents based on multiple criteria.

**Prompt for Turn 1 Query Generation**

Transform the technical sub-question into a natural opening query for a conversation.
**ORIGINAL QUERY TOPIC:** {original_query}
**TECHNICAL SUB-QUESTION:** {sub_question}
**REQUIREMENTS:**

1. This is Turn 1—there is NO conversation history yet.
2. The question must stand alone and introduce the topic naturally.
3. Keep the same technical content but make it sound like a curious person asking.
4. Briefly mention the topic area so the question has context.

**GOOD EXAMPLES:**

- Technical: "What is the mechanism of phototaxis in insects?"
- Natural: "I've noticed bugs always fly toward lights at night. What's actually pulling them there?"
- Technical: "How does the nasal cycle regulate airflow?"
- Natural: "In smell perception, are there basic smells that combine like RGB colors?"

**Return JSON:**
```
{
"conversational_query": "The natural opening question WITH topic intro",
"kept_technical_content": true/false,
"natural_language_used": true/false
}
```

Figure 11: Prompt for generating the first turn query with topic introduction.

## Prompt for Follow-up Query Generation

Transform this into a natural follow-up question.
**TOPIC:** {original_query}
**RECENT CONVERSATION:**
{history}
**TECHNICAL QUESTION TO ASK:** {sub_question}
**PREVIOUS OPENERS USED:** {previous_starters}
**CORE REQUIREMENTS:**

1. Keep the SAME content/intent as the technical question
2. Sound like a real curious person, not a student or interviewer
3. VARY your opener—never repeat a starting word/phrase from "previous openers"
4. VARY your structure—don't use the same question pattern as recent turns

**NATURAL CONVERSATION PATTERNS:**

- **Direct:** "What makes...", "How does...", "Why do..."
- **Curious:** "I wonder if...", "What about...", "How come..."
- **Confirming:** "Does that mean...", "Is that why...", "Would that..."
- **Probing:** "But what if...", "Even when...", "What happens if..."
- **Connecting:** "And does that...", "Then how...", "Which would mean..."

**LANGUAGE RULES:**

- Use simple, everyday words; keep under 20 words
- Use pronouns (it, that, this, they) to connect to previous content
- Use contractions naturally (don't, isn't, wouldn't)
- Do NOT use academic language or filler transitions

**Return JSON:**
```
{
"conversational_query": "The natural follow-up question",
"transition_type": "A/B/C/D/E",
"uses_natural_language": true/false,
"references_previous_content": true/false
}
```

Figure 12: Prompt for generating natural follow-up queries in multi-turn conversations.

Prompt for Grounded Answer Generation

```
Answer this question naturally using the information provided.
CONVERSATION SO FAR:
{history}
CURRENT QUESTION:
{query}
AVAILABLE INFORMATION:
{documents}
HOW TO ANSWER:

  1. Use ONLY information from the provided text above

  2. Speak naturally—explain as if talking to a curious friend

  3. Build on what was discussed earlier in the conversation

  4. Focus on WHAT YOU CAN EXPLAIN, not what you can't

  5. If the full answer isn't available, explain what IS known about the topic

  6. Keep answer conversational: 2-4 sentences, clear and direct

IMPORTANT RULES:

  ✓ State facts directly and naturally

  ✓ Connect to previous conversation smoothly

  ✓ Explain using everyday language

  × Do NOT mention "documents", "sources", "the text"

  × Do NOT end with "However, [gaps in information]"

  × Do NOT add invented examples, names, or details

Return JSON:
{
"answer": "Your natural, conversational answer",
"uses_natural_language": true/false,
"avoids_meta_references": true/false,
"focuses_on_available_info": true/false
}
```

Figure 13: Prompt for generating grounded answers from retrieved documents.

## Prompt for Turn Diversity Validation

```
Does this new answer add value to the conversation?
RECENT CONVERSATION:
{previous_content}
NEW ANSWER:
{new_answer}
EVALUATION CRITERIA:
✓ ACCEPT if answer does ANY of these:

  1. Introduces NEW factual information not previously stated

  2. Drills deeper into a specific detail (e.g., mechanism → sub-mechanism)

  3. Provides a concrete example of something described generally before

  4. Explains a consequence/implication not yet discussed

  5. Answers a different aspect of the same topic

× REJECT only if answer does BOTH:

  1. Contains the SAME factual claims as previous turns (not just same topic)

  2. Uses similar phrasing/wording to express those facts

IMPORTANT DISTINCTIONS:

   • Same TOPIC but different ANGLE = ACCEPT

   • Same FACTS with different WORDS = REJECT

   • General → Specific = ACCEPT

Return JSON:
{
"adds_value": boolean,
"value_type": "new_facts" | "deepening" | "example" | "implication" | "different_angle" |
"repetitive",
"reason": "brief explanation"
}
```

Figure 14: Prompt for validating that new turns add value and avoid repetition.

## Prompt for Aspect Suitability Classification

```
Should this aspect become a conversation sub-question?
ASPECT: {aspect_name}
TYPE: {aspect_type}
EXCERPT: {excerpt}
CLASSIFICATION RULES:

  1. "substantive" → YES: Contains facts, mechanisms, explanations, examples, or concrete
     information.

       • NOTE: Short excerpts ARE ACCEPTABLE if they contain a clear fact.

  2. "meta" → NO: Contains disclaimers, advice to "consult professional", caveats, or
     meta-commentary.

  3. "insufficient" → NO: Content is meaningless or purely structural (e.g., "Here is a list:").

Return JSON:
{
"should_generate": boolean,
"reason": "explanation",
"aspect_category": "substantive" | "meta" | "insufficient"
}
```

Figure 15: Prompt for filtering aspects based on suitability for conversation generation.

## G Automatic Quality Validation

### G.1 Overview

To complement the manual evaluation conducted on 200 randomly sampled conversations (Section 3.4), we employed GPT-4o to automatically assess all conversations in the benchmark. This large-scale validation serves two purposes: (1) confirming that the quality observed in the manual sample generalizes to the full dataset, and (2) providing detailed turn-level linguistic analysis that would be impractical to obtain manually.

The automatic evaluation operates at two levels. At the conversation level, we assess the same four quality dimensions used in manual evaluation: naturalness, turn coherence, question quality, and groundedness. At the turn level, we analyze how each question depends on prior conversation history (dependency type) and what type of information each question seeks (question pattern).

### G.2 Conversation-Level Quality Assessment

We evaluated all 707 conversations on four quality dimensions using separate focused prompts for each dimension. Each dimension receives a score from 1 to 5, matching the Likert scale used in manual evaluation.

**Naturalness** measures whether the conversation sounds like natural human speech rather than robotic or overly formal language. The prompt checks for casual word choices, natural conversation starters, and appropriate sentence style.

**Turn Coherence** evaluates whether conversation turns connect logically to each other. The prompt examines clear references (pronouns pointing to previously mentioned content), logical connections between questions and preceding answers, and absence of abrupt topic jumps.

**Question Quality** assesses whether questions cover different useful aspects of the topic without repetition. The prompt checks that each question asks about something distinct, questions are specific enough to elicit clear answers, and together the questions provide good coverage of the topic.

**Groundedness** determines whether answers are supported by the source documents. The prompt verifies that claims match document content, no information is fabricated, and no contradictions exist between answers and sources.

Table 27 presents the comparison between LLM and human scores across all four dimensions.

| Dim. | LLM | $\sigma$ | Hum. | $\Delta$ |
|---|---|---|---|---|
| Naturalness | 4.34 | 0.58 | 4.20 | +0.14 |
| Turn Coherence | 4.97 | 0.18 | 4.62 | +0.35 |
| Quest. Quality | 4.42 | 0.54 | 4.21 | +0.21 |
| Groundedness | 4.73 | 0.53 | 4.45 | +0.28 |

Table 27: Comparison of automatic (LLM) and Human scores. **Dim.**: Quality Dimension; **LLM**: GPT-4o mean; $\sigma$: Standard deviation; **Hum.**: Human mean; $\Delta$: Difference (LLM − Human). All differences are within the acceptable 0.5 threshold.

The results demonstrate strong alignment between automatic and manual evaluation. All four dimensions show LLM scores within 0.35 points of human scores, well below the 0.5 threshold we consider acceptable. The consistently positive differences indicate that LLM evaluation tends to be slightly more lenient than human evaluation, a pattern commonly observed in LLM-as-judge settings. The low standard deviations, particularly for turn coherence (0.18), reflect high consistency in the generated conversations.

Table 28 presents the score distribution for each dimension.

| Score | Nat. | Coh. | Q. Qual. | Grnd. |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 3 |
| 3 | 39 | 2 | 15 | 19 |
| 4 | 387 | 15 | 381 | 146 |
| 5 | 281 | 690 | 311 | 539 |

Table 28: Score distribution across four quality dimensions. The concentration of scores at 4 and 5 indicates consistently high quality across the dataset. (Nat.: Naturalness, Coh.: Turn Coherence, Q. Qual.: Question Quality, Grnd.: Groundedness).

### G.3 Turn-Level Dependency Analysis

Beyond conversation-level quality, we analyzed how each of the 2,971 turns depends on the preceding conversation context. Understanding these dependency patterns is important for conversational retrieval systems, as different dependency types pose different challenges for query understanding and document retrieval.

We classified each turn into one of six dependency types:

**Coreference**: The question uses pronouns (it, this, that, they, these, those) that point back to entities or concepts mentioned in previous turns. Example: After discussing photosynthesis, asking "Does it also occur at night?"

**Ellipsis**: The question is grammatically incomplete and requires prior context to be understood. Example: "And the second reason?" or "What about in winter?"

**Substitution**: The question uses a general term to refer to something specific mentioned earlier. Example: After explaining a chemical reaction, asking "How efficient is this process?"

**Continuation**: The question is grammatically complete and addresses the same topic as previous turns, but without explicit backward references. Example: After discussing photosynthesis mechanisms, asking "What role does chlorophyll play?"

**Topic Shift**: The question moves to a new aspect or subtopic of the overall conversation theme. Example: After discussing scientific aspects, asking "What are the economic implications?"

**Self-contained**: The question is fully independent and can be understood without any prior context. This typically applies to first-turn questions.

Table 29 presents the distribution of dependency types across all turns.

| Dependency Type | Count | Percentage |
|---|---|---|
| Continuation | 1,028 | 34.6% |
| Coreference | 859 | 28.9% |
| Self-contained | 707 | 23.8% |
| Topic Shift | 326 | 11.0% |
| Substitution | 41 | 1.4% |
| Ellipsis | 10 | 0.3% |
| **Total** | **2,971** | **100%** |

Table 29: Distribution of turn dependency types. The diversity of dependency patterns reflects natural conversational behavior.

The dependency type entropy is 0.77 (normalized to 0-1 scale), indicating balanced variety in how turns connect to prior context. The high proportion of continuation (34.6%) and coreference (28.9%) turns demonstrates that the generated conversations maintain topical coherence while progressively building on shared context. The self-contained category (23.8%) aligns closely with the proportion of first turns in the dataset, as expected. Topic shifts (11.0%) appear at appropriate rates, allowing conversations to explore multiple aspects of complex topics.

### G.4 Turn-Level Question Pattern Analysis

We also classified each turn by the type of information the question seeks. This analysis reveals the diversity of information-seeking behaviors in the benchmark and ensures that conversations explore topics through varied questioning strategies.

We identified nine question patterns:

**Why**: Questions asking for reasons or causes. Example: "Why does this happen?" or "What causes this effect?"

**How**: Questions asking how something works or happens. Example: "How does the process unfold?" or "What is the mechanism?"

**What**: Questions asking for factual information such as definitions, names, dates, or places. Example: "What is the definition of X?" or "When did this occur?"

**Compare**: Questions asking about differences or similarities between concepts. Example: "How does X differ from Y?" or "Which approach is more effective?"

**What-if**: Questions exploring possibilities or hypothetical scenarios. Example: "What would happen if X changed?" or "Could this work under different conditions?"

**Confirm**: Questions seeking to verify understanding. Example: "Is it true that X causes Y?" or "Does that mean Z?"

**More-detail**: Questions requesting deeper explanation of something already mentioned. Example: "Can you explain more about X?" or "What specifically happens during Y?"

**Example**: Questions asking for concrete instances or cases. Example: "What are some examples of this?" or "Can you give a specific case?"

**Effect**: Questions asking about results, outcomes, or implications. Example: "What are the effects of X?" or "What does this mean for Y?"

Table 30 presents the distribution of question patterns across all turns.

The question pattern entropy is 0.91 (normalized to 0-1 scale), indicating high diversity in question types. No single pattern dominates the dataset, with the six most frequent patterns (how, effect, confirm, why, compare, what-if) each contributing between 11% and 19% of turns. This distribution reflects the reasoning-intensive nature of the benchmark, where users explore topics through causal reasoning (why), procedural understanding (how), consequence analysis (effect), and comparative evaluation (compare), rather than simple factual queries

| Question Pattern | Count | Percentage |
|---|---|---|
| How | 571 | 19.2% |
| Effect | 486 | 16.4% |
| Confirm | 465 | 15.6% |
| Why | 432 | 14.5% |
| Compare | 411 | 13.8% |
| What-if | 329 | 11.1% |
| What | 206 | 6.9% |
| Example | 50 | 1.7% |
| More-detail | 21 | 0.7% |
| **Total** | **2,971** | **100%** |

Table 30: Distribution of question patterns. The high entropy indicates diverse information-seeking behavior across conversations.

| Dom | #C | #T | Nat | Coh | QQ | Grd | DE |
|---|---|---|---|---|---|---|---|
| Biology | 85 | 362 | 4.38 | 4.98 | 4.47 | 4.74 | 0.93 |
| Drones | 37 | 142 | 4.59 | 4.97 | 4.51 | 4.76 | 0.96 |
| Earth Science | 98 | 454 | 4.33 | 5.00 | 4.54 | 4.67 | 0.77 |
| Economics | 74 | 288 | 4.12 | 4.97 | 4.41 | 4.62 | 0.86 |
| Hardware | 46 | 188 | 4.57 | 4.80 | 4.44 | 4.74 | 0.87 |
| Law | 50 | 230 | 4.34 | 5.00 | 4.47 | 4.74 | 0.74 |
| Med. Sciences | 44 | 183 | 4.42 | 4.98 | 4.36 | 4.82 | 0.81 |
| Politics | 43 | 213 | 4.39 | 4.95 | 4.30 | 4.81 | 0.88 |
| Psychology | 84 | 333 | 4.45 | 5.00 | 4.16 | 4.69 | 0.76 |
| Robotics | 68 | 259 | 4.07 | 4.97 | 4.49 | 4.68 | 0.80 |
| Sustain. Living | 78 | 319 | 4.32 | 5.00 | 4.46 | 4.83 | 0.86 |

Table 31: Per-domain quality scores and dependency entropy. **Dom**: Domain, **#C**: No. Conversations, **#T**: No. Turns, **Nat**: Naturalness, **Coh**: Coherence, **QQ**: Question Quality, **Grd**: Grounding, **DE**: Dependency Entropy.

(what, 6.9%).

### G.5 Per-Domain Results

Table 31 presents quality scores and turn-level metrics broken down by domain.

Quality scores remain consistently high across all eleven domains, with no domain falling below 4.0 on any dimension. Dependency entropy varies from 0.74 (Law) to 0.96 (Drones), reflecting domain-specific conversational patterns. Technical domains such as Drones and Biology exhibit higher dependency diversity, while structured domains such as Law show more uniform turn progression patterns.

### G.6 Evaluation Prompts

This section provides the complete prompts used for automatic evaluation.

**Naturalness Evaluation Prompt**

Evaluate whether this conversation sounds like NATURAL HUMAN SPEECH.
**CONVERSATION:**
{conversation}
**TASK:** Rate how natural and human-like the language is (not robotic or overly formal).
**CHECK ONLY these language features:**
**1. CASUAL WORD CHOICES:**

- Contractions used: "don’t", "isn’t", "what’s" (vs formal "do not", "is not")
- Casual words: "got", "stuff", "things", "pretty much" (vs formal "obtained", "materials")
- Softening words: "kind of", "sort of", "maybe", "I think"

**2. CONVERSATION STARTERS:**

- Natural openers: "So", "Well", "Oh", "Yeah", "I mean", "Actually"
- Response words: "Right", "Okay", "I see", "Got it"

**3. SENTENCE STYLE:**

- Natural incomplete sentences (OK in speech)
- Casual question phrasing (not stiff templates)

**DO NOT CHECK (other prompts handle these):**

- Whether turns connect logically
- Whether questions cover the topic well
- Whether answers are factually correct

**SCORING:**
5 - Sounds completely natural, like real human speech
4 - Mostly natural, few formal spots
3 - Mix of natural and formal/stiff language
2 - Mostly formal or robotic sounding
1 - Entirely artificial, would never occur in real speech
**Return JSON:**
{
"score": <1-5>,
"natural_phrases": ["quote 2-3 natural-sounding phrases found"],
"unnatural_phrases": ["quote any stiff or robotic phrases found"],
"justification": "1-2 sentences explaining your score"
}

Figure 16: Prompt for evaluating conversation naturalness.

**Turn Coherence Evaluation Prompt**

Evaluate whether conversation turns CONNECT LOGICALLY to each other.
**CONVERSATION:**
{conversation}
**TASK:** Rate how well each question follows from the previous answer.
**CHECK ONLY these connection features:**
**1. CLEAR REFERENCES:**

- When "it", "this", "that", "they" are used, is it clear what they refer to?
- When "the problem" or "the process" is mentioned, was it introduced earlier?

**2. LOGICAL CONNECTIONS (how does each question relate to the previous answer?):**

- ASKS FOR DETAILS: Question wants more information about something just mentioned
- ASKS TO CLARIFY: Question wants to understand something from the answer better
- ASKS ABOUT EFFECTS: Question asks what happens as a result of what was described
- ASKS FOR CONTRAST: Question asks about an alternative or opposite view
- NARROWS FOCUS: Question zooms in on one specific part mentioned

**3. NO GAPS:**

- Each question should make sense given what came before
- No sudden jumps to unrelated topics without transition

**DO NOT CHECK (other prompts handle these):**

- Whether language sounds natural
- Whether questions are about different topics
- Whether answers are factually correct

**SCORING:**
5 - Every question clearly connects to the previous answer
4 - Most questions connect well, 1-2 slightly unclear links
3 - Some questions connect, but several feel disconnected
2 - Many questions don't clearly follow from previous answers
1 - Questions seem random, no logical flow
**Return JSON:**

```
{
"score": <1-5>,
"good_connections": ["describe 1-2 turns that connect well"],
"weak_connections": ["describe any turns that don't connect clearly"],
"unclear_references": ["list any 'it/this/that' without clear meaning"],
"justification": "1-2 sentences explaining your score"
}
```

Figure 17: Prompt for evaluating turn coherence.

### Question Quality Evaluation Prompt

Evaluate whether the questions COVER DIFFERENT USEFUL ASPECTS of the topic.
**ORIGINAL TOPIC:**
{original_query}
**QUESTIONS:**
{questions}
**TASK:** Rate whether the questions explore the topic well without repeating.
**CHECK ONLY these coverage features:**
**1. DIFFERENT ASPECTS:**

- Does each question ask about something DIFFERENT?
- Are any two questions basically asking the same thing in different words?

**2. USEFUL QUESTIONS:**

- Are questions specific enough to get clear answers?
- Do questions go beyond obvious or trivial information?
- Bad examples: "What is X?" (too basic), "Tell me about X" (too vague)

**3. GOOD COVERAGE:**

- Together, do the questions explore important parts of the topic?
- Are there major aspects of the topic that are missed?

**DO NOT CHECK (other prompts handle these):**

- Whether questions sound natural
- Whether questions follow logically from answers
- Whether the answers are correct

**SCORING:**
5 - All questions distinct, specific, and together cover the topic well
4 - Most questions distinct and useful, maybe 1 similar pair
3 - Some overlap between questions, some too vague, partial coverage
2 - Several questions overlap or are too vague/trivial
1 - Heavy repetition, mostly vague or useless questions
**Return JSON:**
{
"score": <1-5>,
"aspects_covered": ["list the different aspects/subtopics the questions address"],
"repeated_questions": ["list any question pairs that ask about the same thing"],
"weak_questions": ["list any questions that are too vague or trivial"],
"justification": "1-2 sentences explaining your score"
}

Figure 18: Prompt for evaluating question quality.

Groundedness Evaluation Prompt

Evaluate whether the answers are SUPPORTED BY THE DOCUMENTS.
**SOURCE DOCUMENTS:**
{documents}
**ANSWERS TO CHECK:**
{answers}
**TASK:** Rate whether the answer content can be found in or reasonably inferred from the documents.
**CHECK ONLY these accuracy features:**
**1. CLAIMS MATCH DOCUMENTS:**

- Can each fact in the answers be found in the source documents?
- Are the facts stated correctly (not twisted or misrepresented)?

**2. NOTHING MADE UP:**

- Are there any invented names, places, or organizations?
- Are there any made-up numbers, dates, or statistics?
- Are there any fake technical terms?

**3. NO CONTRADICTIONS:**

- Do any claims directly contradict what the documents say?
- Are there claims that go way beyond what the documents support?

**ACCEPTABLE:**

- Rewording information from documents
- Drawing obvious conclusions from stated facts

**NOT ACCEPTABLE:**

- Adding facts not found in documents
- Stating guesses as if they were facts

**DO NOT CHECK (other prompts handle these):**

- Whether answers sound natural
- Whether answers connect to each other
- Whether the questions were good

**SCORING:**
5 - All claims supported by documents, nothing made up
4 - Nearly all claims supported (95%+), only minor inferences
3 - Most claims supported (80-95%), some unsupported but plausible
2 - Many claims unsupported (50-80%), some things made up
1 - Mostly unsupported (<50%), major fabrications
**Return JSON:**
{
"score": <1-5>,
"supported_claims": ["list 2-3 claims that ARE in the documents"],
"unsupported_claims": ["list claims NOT found in documents"],
"made_up_content": ["list any invented facts, names, or numbers"],
"justification": "1-2 sentences explaining your score"
}

Figure 19: Prompt for evaluating answer groundedness.

Turn Dependency Classification Prompt

Classify how this question DEPENDS on the previous conversation.
**PREVIOUS CONVERSATION:**
{prior_context}
**CURRENT QUESTION:**
{current_question}
**TASK:** Pick ONE dependency type that best describes how this question relates to what came before.
**DEPENDENCY TYPES (pick ONE):**

1. **"coreference"** - Uses pronouns pointing back to something mentioned before
   - Look for: it, this, that, they, these, those
   - Example: "Does IT also affect..." / "What about THAT?"
2. **"ellipsis"** - Incomplete sentence that needs context to understand
   - Look for: Missing words, fragments like "And...?", "How about...?"
   - Example: "And the second reason?" / "What about in winter?"
3. **"substitution"** - Uses a general term for something specific mentioned before
   - Look for: "this process", "that method", "the problem", "such cases"
   - Example: After "photosynthesis" → "How efficient is this process?"
4. **"continuation"** - Complete question on the same topic, but no explicit links
   - Look for: Full sentence, related topic, no pronouns pointing back
   - Example: After photosynthesis → "What role does chlorophyll play?"
5. **"topic_shift"** - Moves to a new aspect or subtopic
   - Look for: "What about...", "Regarding...", "Moving to...", new direction
   - Example: After science → "What are the economic effects?"
6. **"self_contained"** - Fully independent, makes sense alone (usually Turn 1)
   - Look for: Could be understood without any prior context
   - Example: First question of the conversation

**Return JSON:**
```
{
"dependency_type": "one of the 6 types above",
"evidence": "quote the specific words that show this type",
"explanation": "one sentence explaining your choice"
}
```

Figure 20: Prompt for classifying turn dependency types.

**Question Pattern Classification Prompt**

```
Classify what TYPE of answer this question is looking for.
QUESTION:
{question}
TASK: Pick ONE pattern that best describes what kind of information this question wants.
QUESTION TYPES (pick ONE):
1. "why" - Asks for REASONS or CAUSES

   • Look for: "Why...?", "What causes...?", "What leads to...?"

   • Wants: Explanations, reasons

2. "how" - Asks HOW something WORKS or HAPPENS

   • Look for: "How does...?", "What is the process...?", "What happens...?"

   • Wants: Steps, process descriptions

3. "what" - Asks for FACTS (definitions, names, dates, places)

   • Look for: "What is...?", "Who...?", "When...?", "Where...?"

   • Wants: Specific facts, definitions

4. "compare" - Asks about DIFFERENCES or SIMILARITIES

   • Look for: "How does X differ from Y?", "Is X similar to Y?", "Which is better?"

   • Wants: Comparisons

5. "what_if" - Asks about POSSIBILITIES or HYPOTHETICALS

   • Look for: "What would happen if...?", "Could X...?", "What if...?"

   • Wants: Speculation, possibilities

6. "confirm" - Asks to VERIFY understanding

   • Look for: "Is it true that...?", "Does that mean...?", "So...right?"

   • Wants: Yes/no confirmation

7. "more_detail" - Asks for MORE INFORMATION on something

   • Look for: "Can you explain more...?", "What specifically...?", "Tell me more..."

   • Wants: Deeper explanation

8. "example" - Asks for EXAMPLES or INSTANCES

   • Look for: "What are examples of...?", "Can you give an instance?", "Like what?"

   • Wants: Concrete cases

9. "effect" - Asks about RESULTS or IMPLICATIONS

   • Look for: "What does this mean for...?", "What are the effects...?", "What happens as a
     result?"

   • Wants: Outcomes, impacts

Return JSON:
{
"question_pattern": "one of the 9 types above",
"evidence": "quote the specific words that show this type",
"explanation": "one sentence explaining your choice"
}
```

Figure 21: Prompt for classifying question patterns.