# Tree-of-Quote Prompting Improves Factuality and Attribution in Multi-Hop and Medical Reasoning

**Justin Xu**[1,*], **Yiming Li**[1,*], **Zizheng Zhang**[1], **Augustine YH Luk**[1], **Mayank Jobanputra**[2], **Samarth Oza**[3], **Ashley Murray**[1], **Meghana R Kasula**[4], **Andrew Parker**[5], **David W Eyre**[1,4],
[1]University of Oxford, [2]Saarland University, [3]AIIMS Rajkot,
[4]Oxford University Hospitals NHS Foundation Trust, [5]Independent

## Abstract

Large language models (LLMs) can produce fluent but factually incorrect outputs and often have limited ability to attribute their claims to source material. This undermines their reliability, particularly in multi-hop and high-stakes domains such as medicine. We propose Tree-of-Quote (ToQ) prompting, a structured framework that decomposes complex questions into subquestions, generates quotes to support each step without retrieval, and selectively advances reasoning based on quote quality. We also introduce FQ-Score, a unified metric that captures answer correctness, attribution fidelity, and reasoning quality. Experiments on StrategyQA, 2WikiMultiHopQA, MuSiQue, MoreHopQA, and MedQA demonstrate that ToQ improves factuality and attribution over standard prompting baselines. To validate FQ-Score as a proxy for human judgment, we conduct two reader studies with clinicians on medical questions, and observe strong correlations. Both clinician scores and FQ-Scores also indicate a preference for ToQ over baselines due to a combination of greater correctness, completeness, and logical flow. Our results suggest ToQ is a promising approach for building more trustworthy and auditable LLM systems.

## 1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide range of natural language tasks, including question answering (QA), summarization, and reasoning (Singhal et al., 2023, 2025; Jahan et al., 2024; Naveed et al., 2024; Xu et al., 2025). Despite these advances, LLMs frequently generate outputs that are fluent yet factually incorrect — a phenomenon commonly referred to as hallucination (Augenstein et al., 2023; Huang et al., 2025). This issue is particularly critical in domains that require high factual precision and traceability, such as in medical QA or multi-hop reasoning tasks (Farquhar et al., 2024; Anjum et al., 2024). Moreover, the black-box nature of some LLMs complicates the ability to attribute the origins of model claims despite past efforts (Liu et al., 2025), making it difficult to discern whether a model response is grounded in credible evidence or is fabricated.

Existing prompting strategies such as Chain-of-Thought (CoT) prompting (Wei et al., 2023) encourage intermediate reasoning steps to improve accuracy, while Chain-of-Verification (CoVe) (Dhuliawala et al., 2023) introduces a post hoc check to validate steps. However, neither of these methods adequately addresses attribution. Even with well-reasoned answers, it is often uncertain whether the reasoning is based on actual knowledge from pre-training data or merely plausible fabrication. Prior work has shown that LLMs are sensitive to verbatim quotes from their training corpora and that quoting can improve both factuality and verifiability (Weller et al., 2024). However, quoting has largely been studied in isolation or used only in final answer justification — not as a tool integrated throughout the reasoning process.

To address this gap, we propose Tree-of-Quote Prompting (ToQ), a novel method that unifies quoting and reasoning in an iterative, modular pipeline. Rather than treating quoting as a post-processing step, ToQ introduces a structured workflow in which the model generates subquestions, attributes subanswers to non-verbatim quotes, and scores the quote quality before progressing. Inspired by Chain-of-Quote (CoQ) (Li et al., 2024) and Tree-of-Thoughts (ToT) (Yao et al., 2023), this structured prompting paradigm promotes robust multi-step reasoning with attributed quotes at every stage without an explicit retrieval pipeline.

Our contributions in this work are as follows:

1. We present empirical results of ToQ prompt-

[*]Co-first authors



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5623
November 4-9, 2025 

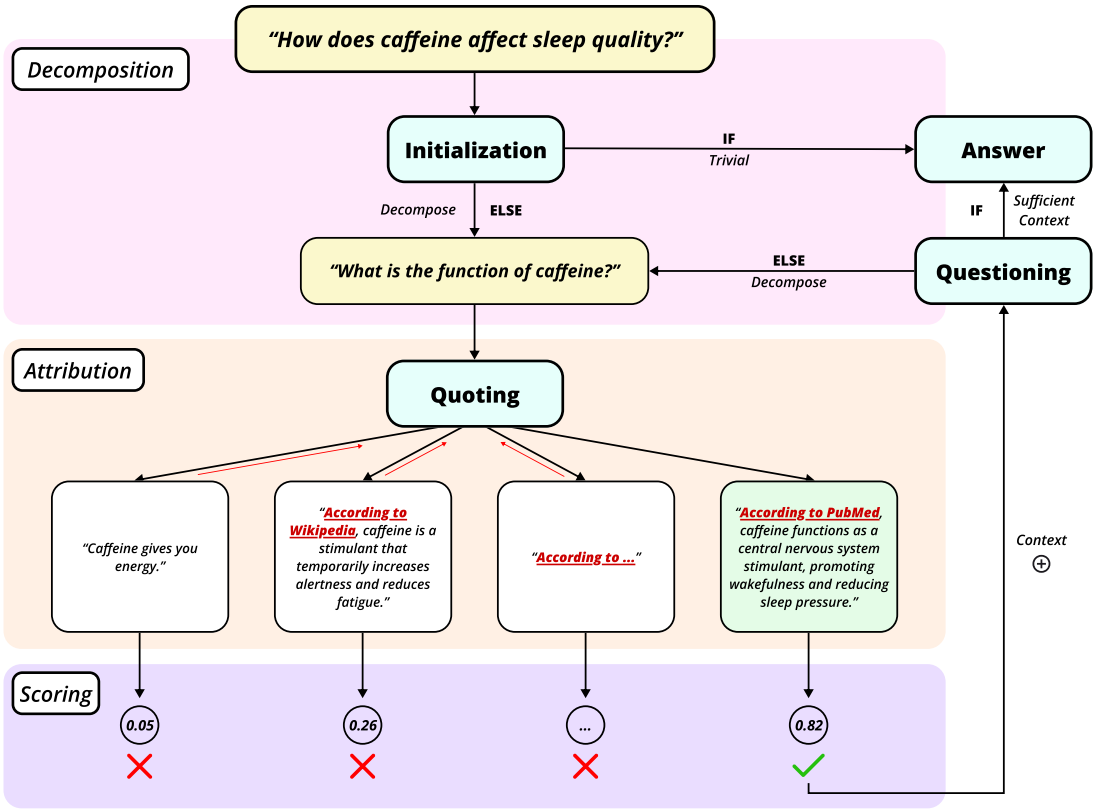


Figure 1: Schematic of the Tree-of-Quote (ToQ) prompting pipeline.

ing on five reasoning and QA datasets and compare it against several baselines, including zero-shot, CoT, CoVe, and CoQ prompting.

2. We introduce a new evaluation metric for reasoning chains with quotes — FQ-Score — that jointly measures answer correctness, quote attribution quality, and alignment between reasoning steps by leveraging a combination of existing metrics and LLM calls.

3. We conduct two reader studies to validate FQ-Score against human evaluation, demonstrating strong correlation between automated and clinician-assigned scores that capture correctness, completeness, and logical coherence.

## 2 Related Works

We build on a growing literature aiming to improve factuality and attribution in LLM outputs. Prior efforts like WebGPT (Nakano et al., 2022) and GopherCite (Menick et al., 2022) fine-tune LLMs to return answers with supporting links or quotes, often via reinforcement learning. More recent systems such as ALCE (Gao et al., 2023b) and LongCite (Zhang et al., 2024) introduce benchmarks and training pipelines to generate fine-grained, sentence-level citations in long-context QA. LongCite-8B/9B achieve state-of-the-art citation precision using a 45k-instance dataset. Additionally, HoT prompting (Nguyen et al., 2025) enhances attribution during reasoning by tagging facts in input and response. While helpful for verification, our method differs by explicitly conditioning reasoning advancement on quote quality.

Several works also explore post-hoc attribution. Sancheti et al. (2024) formalize this task for long documents and compare retrieval and entailment-based attribution methods. Ramu et al. (2024) decompose answers into factual units to improve attribution granularity. RARR (Gao et al., 2023a) retrofits answers by searching for missing evidence post-generation. In contrast, ToQ integrates grounding directly during generation via quote-supported subquestions. Other grounded reasoning strategies include "Attribute-First" (Slobodkin et al., 2024) and "Blueprint-before-Generation" (Fierro et al., 2024), which pre-select evidence before responding. ToQ similarly decomposes questions, but adds quote-based validation at each step.

LongCite (Zhang et al., 2024) and Coarse-Grained Answer Decomposition (Cao and Liu, 2021) emphasize source alignment in long-context

QA, as does our FQ-Score metric, which jointly evaluates correctness, attribution, and reasoning. Toolkits like CiteKit (Shen et al., 2024) and retrieval-augmented reasoning frameworks such as Self-RAG (Asai et al., 2023) show that modular pipelines improve verifiability. Hence, we design ToQ as a structured prompting and evaluation framework for auditable multi-hop reasoning.

## 3 Tree-of-Quote

Tree-of-Quote (ToQ) prompting is a modular and interpretable framework that integrates non-verbatim quoting into the reasoning process of LLMs. ToQ structures reasoning as a dynamic graph (via `LangGraph`) with four core nodes — initialization → quoting → scoring → questioning — that together support iterative, quote-grounded multi-hop inference (Figure 1).

The process begins with an "Initialization" step, where the system either decomposes a complex question into a simpler subquestion or, in the case of a trivial query, directly produces an answer. Next, during the "Quoting" phase, the system attempts to answer the subquestion using a quote from Wikipedia or PubMed, or from explicitly provided documents, by generating directly from its pretraining knowledge without retrieval. The exact quote source is intentionally left as a controllable design choice; however, only Wikipedia and PubMed are investigated in this paper. If the initial quote does not meet the required fidelity, as assessed by a scoring mechanism, it performs up to $N$ retries to improve the quote. In the "Scoring" stage, each quote is evaluated using the Quoted Information Precision (QUIP) metric (Weller et al., 2024). Quotes that score below a predefined threshold (which can be tuned to suit different goals) prompt additional retries as above; if retries are exhausted, the highest-scoring quote is selected. Finally, the "Questioning" phase determines whether the system has sufficient context to produce a final answer or whether it should generate another subquestion to continue the reasoning process. We leverage structured XML outputs for each node to define the next action for the pipeline. Prompts for each step are available in Appendices A.1–A.3.

### 3.1 Reasoning Graph and Tree Structure

ToQ operates as a tree-structured reasoning system. Each subquestion generated by the initialization or questioning node spawns a new quote-attempting branch. These branches are explored through the quoting and scoring nodes, where candidate quotes are evaluated via a scoring function (*i.e.*, QUIP). Branches that fail to meet a minimum quote quality threshold are pruned. The initialization node can also regenerate entirely new subquestions (varied based on LLM sampling parameters), forming alternate high-level reasoning paths. This branching behavior supports both exploration and selective trust in attributed evidence.

## 4 Experiments

### 4.1 Experimental Setup

We first evaluate ToQ on four datasets that span different kinds of QA and reasoning in the general domain. To probe implicit reasoning, we use the 687-question testing split of StrategyQA (Geva et al., 2021), a binary `yes|no` question set designed to measure the ability to infer a strategy for multi-hop reasoning. We further include two other multi-hop reasoning datasets with contextual support corpora: 2WikiMultiHopQA (2Wiki) (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022), and specifically evaluate using the 500 examples from their respective published subsampled test sets. These datasets include a fixed set of support documents with each question, allowing us to constrain the quoting process strictly to the provided corpus. Lastly, we leveraged MoreHopQA (Schnitzler et al., 2024), a recently proposed benchmark that emphasizes generative responses and integrates additional layers of complexity through commonsense, arithmetic, and symbolic reasoning.

For each dataset, we compare ToQ against four prompting baselines: zero-shot, where the model receives only the question; Chain-of-Thought (CoT) (Wei et al., 2023), which generates step-by-step rationales before the final answer; Chain-of-Verification (CoVe) (Dhuliawala et al., 2023), which appends a verification step to CoT; and Chain-of-Quote (CoQ) (Li et al., 2024), which produces reasoning chains that include evidence quotes but without attribution scoring or quality control. We evaluate all methods using two model backends, GPT-4o (OpenAI et al., 2024) and DeepSeek-Chat (V3) (DeepSeek-AI et al., 2024), to assess performance in different training recipes.

### 4.2 Metrics for Answer Evaluation

For answer-based metrics, we adapt the evaluation depending on the format of the dataset.

For datasets with text-based answers (*i.e.*, 2Wiki, MuSiQue, and MoreHopQA), we use a mix of lexical and semantic measures. Following the evaluation used in SQuAD (Rajpurkar et al., 2016), we report Exact Match (EM), defined as the percentage of predictions that match the reference answers exactly, after normalizing case, punctuation, whitespace, and removing articles (*i.e.*, "a", "an", "the"). We also report a macro-averaged F1 score that measures overlap between bags of tokens of the prediction and ground-truth answers. Additionally, we include Membership Match (MM), which evaluates whether the candidate answer is a subset or superset of any reference answer, ignoring stop words. Finally, to account for deeper semantic equivalence, we introduce a Semantic Match (SM) metric based on language model inference (see prompt in Appendix A.4), where an LLM judges whether the model's answer is meaning-equivalent to the reference.

We further include lexical similarity metrics — BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) — as well as semantic similarity metrics, including BERTScore (Zhang et al., 2020) and METEOR (Banerjee and Lavie, 2005), to give a fuller picture of surface-level and contextual overlap. We lastly compute AlignScore (Zha et al., 2023), a factual alignment metric that measures consistency between reference and candidate answers.

For the choice-based StrategyQA dataset, we compute simple accuracy, defined as the percentage of cases where the model-selected answer exactly matches the ground-truth label. We also report a conventional F1 score, which considers precision and recall for each of the answer choices.

### 4.3 Performance on Reasoning Datasets

Due to the stochastic nature of ToQ — particularly its quote-scoring and retry mechanism — we report all results as the mean of at least three independent replicates, alongside a range that captures the minimum and maximum observed scores.

Table 1 shows StrategyQA results, where we report accuracy and F1 for all methods. ToQ achieves the highest performance on both metrics across GPT-4o and DeepSeek-Chat, with CoT and CoQ performing competitively with GPT-4o and DeepSeek-Chat, respectively. Tables 2 and 3 provide comprehensive performance comparisons on 2Wiki and MuSiQue, covering the metrics described in Section 4.2. ToQ significantly outperforms all baselines on GPT-4o for 2Wiki, with consistent improvements across all lexical and semantic evaluation metrics. On DeepSeek-Chat, ToQ again leads overall performance across most metrics on 2Wiki, with the exception of MM where CoT and CoQ slightly outperform. Interestingly, zero-shot prompting on DeepSeek-Chat consistently ranks second across 7 out of 9 metrics, even outperforming CoT and CoQ, which is consistent with prior findings (Li et al., 2024). On MuSiQue, GPT-4o-ToQ again outperforms all baselines on every metric, with zero-shot emerging as the second-best approach in 6 of the 9 metrics. However, DeepSeek-Chat performs poorly, with zero-shot surprisingly outperforming all methods in 8 of 9 metrics, and ToQ coming in second in 6 of them. We hypothesize that this may be related to DeepSeek-Chat's pre-training or instruction tuning regime.

| **Method** | Accuracy ↑ | F1 ↑ |
|---|---|---|
| | *GPT-4o* | |
| Zero-Shot | 0.734 | 0.728 |
| CoT | 0.792 | 0.788 |
| CoVe | 0.764 | 0.759 |
| CoQ | 0.790 | 0.785 |
| ToQ (Ours) | **0.817** (0.803–0.833) | **0.815** (0.800–0.831) |
| | *DeepSeek-Chat* | |
| Zero-Shot | 0.659 | 0.624 |
| CoT | 0.780 | 0.777 |
| CoVe | 0.752 | 0.752 |
| CoQ | 0.785 | 0.779 |
| ToQ (Ours) | **0.812** (0.804–0.819) | **0.815** (0.803–0.815) |

Table 1: Accuracy and F1 on StrategyQA across prompting methods.

Table 4 presents results on MoreHopQA, where we report only the core answer-based metrics (EM, MM, SM, and F1). With GPT-4o, CoT slightly outperforms ToQ on 3 of the 4 metrics, but ToQ trails by negligible amounts (0.002–0.003), and notably surpasses CoT on SM by a sizable margin (+0.31). For DeepSeek-Chat, ToQ achieves the highest scores on 3 of the 4 metrics, trailing only on MM. CoT consistently ranks second.

We further compare ToQ against specialized reasoning-tuned models in zero-shot settings (GPT-4o with o4-mini, and DeepSeek-Chat with DeepSeek-R1). GPT-4o-ToQ consistently outperforms o4-mini across all metrics on 2Wiki by a sizable amount, and DeepSeek-Chat-ToQ also closely matches DeepSeek R1's performance on 2Wiki, though it still lags on MuSiQue.

| Method | 2WikiMultiHopQA ↑ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | EM | MM | SM | F1 (Macro) | BLEU-4 | ROUGE-L | BERTScore | METEOR | AlignScore |
| *GPT-4o* | | | | | | | | | |
| Zero-Shot | 0.542 | 0.778 | 0.784 | 0.672 | 0.042 | 0.681 | 0.604 | 0.588 | 0.607 |
| CoT | 0.468 | 0.820 | 0.866 | 0.663 | 0.042 | 0.669 | 0.579 | 0.615 | 0.551 |
| CoVe | 0.566 | 0.726 | 0.732 | 0.671 | 0.063 | 0.683 | 0.627 | 0.579 | 0.638 |
| CoQ | 0.312 | 0.814 | 0.850 | 0.561 | 0.041 | 0.569 | 0.501 | 0.574 | 0.389 |
| ToQ (Ours) | **0.686** (0.678–0.698) | **0.858** (0.840–0.872) | **0.890** (0.870–0.912) | **0.797** (0.781–0.810) | **0.185** (0.182–0.187) | **0.816** (0.813–0.818) | **0.819** (0.815–0.823) | **0.708** (0.705–0.713) | **0.761** (0.760–0.763) |
| *DeepSeek-Chat* | | | | | | | | | |
| Zero-Shot | 0.628 | 0.798 | 0.804 | 0.732 | 0.137 | 0.738 | 0.709 | 0.633 | 0.695 |
| CoT | 0.520 | **0.896** | 0.790 | 0.661 | 0.040 | 0.668 | 0.535 | 0.569 | 0.574 |
| CoVe | 0.562 | 0.738 | 0.752 | 0.670 | 0.074 | 0.676 | 0.611 | 0.576 | 0.635 |
| CoQ | 0.470 | 0.864 | 0.854 | 0.649 | 0.038 | 0.659 | 0.529 | 0.600 | 0.520 |
| ToQ (Ours) | **0.686** (0.680–0.686) | 0.856 (0.850–0.864) | **0.878** (0.870–0.888) | **0.793** (0.788–0.795) | **0.191** (0.184–0.196) | **0.798** (0.793–0.801) | **0.801** (0.794–0.808) | **0.699** (0.697–0.702) | **0.748** (0.743–0.753) |

Table 2: Performance on 2WikiMultiHopQA across prompting methods on GPT-4o and DeepSeek-Chat. Metrics include lexical (EM, MM, F1, BLEU, ROUGE), semantic (SM, BERTScore, METEOR), and factual consistency (AlignScore).

| Method | MuSiQue ↑ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | EM | MM | SM | F1 (Macro) | BLEU-4 | ROUGE-L | BERTScore | METEOR | AlignScore |
| *GPT-4o* | | | | | | | | | |
| Zero-Shot | 0.348 | 0.598 | 0.640 | 0.522 | 0.022 | 0.531 | 0.444 | 0.452 | 0.479 |
| CoT | 0.300 | 0.586 | 0.600 | 0.503 | 0.008 | 0.509 | 0.389 | 0.426 | 0.496 |
| CoVe | 0.316 | 0.514 | 0.530 | 0.462 | 0.030 | 0.458 | 0.411 | 0.396 | 0.432 |
| CoQ | 0.248 | 0.572 | 0.646 | 0.442 | 0.008 | 0.449 | 0.347 | 0.421 | 0.402 |
| ToQ (Ours) | **0.434** (0.407–0.477) | **0.625** (0.596–0.667) | **0.681** (0.645–0.738) | **0.594** (0.566–0.634) | **0.041** (0.031–0.044) | **0.598** (0.579–0.627) | **0.586** (0.562–0.618) | **0.500** (0.486–0.520) | **0.576** (0.560–0.615) |
| *DeepSeek-Chat* | | | | | | | | | |
| Zero-Shot | **0.436** | 0.662 | **0.724** | **0.604** | **0.046** | **0.613** | **0.570** | **0.521** | **0.556** |
| CoT | 0.380 | **0.696** | 0.678 | 0.528 | 0.006 | 0.582 | 0.439 | 0.452 | 0.522 |
| CoVe | 0.372 | 0.605 | 0.627 | 0.540 | 0.016 | 0.548 | 0.468 | 0.456 | 0.499 |
| CoQ | 0.344 | 0.648 | 0.686 | 0.528 | 0.011 | 0.540 | 0.373 | 0.442 | 0.487 |
| ToQ (Ours) | 0.385 (0.362–0.403) | 0.557 (0.522–0.578) | 0.633 (0.592–0.657) | 0.533 (0.505–0.555) | 0.036 (0.031–0.041) | 0.550 (0.516–0.588) | 0.516 (0.485–0.552) | 0.455 (0.423–0.490) | 0.527 (0.503–0.552) |

Table 3: Performance on MuSiQue across prompting methods on GPT-4o and DeepSeek-Chat. Metrics include lexical (EM, MM, F1, BLEU, ROUGE), semantic (SM, BERTScore, METEOR), and factual consistency (AlignScore).

In terms of attribution quality, we find that ToQ consistently produces higher-fidelity quotes compared to CoQ. For example, on StrategyQA, the mean QUIP score for ToQ is 0.725 compared to 0.369 for CoQ. This improvement stems from ToQ's quote-scoring and retry mechanism, which filters out hallucinated quotes and encourages the model to regenerate better-supported evidence. In open-domain quoting settings with Wikipedia or PubMed, quote fidelity can vary widely; however, ToQ attempts to mitigate the effects of this variability by enforcing quote quality thresholds and iteratively improving its evidence selection. As a result, ToQ achieves more reliable attribution, which translates to improved factual grounding throughout the reasoning process without the need for a retrieval-augmented generation pipeline.

| Method | MoreHopQA ↑ | | | |
|---|---|---|---|---|
| | EM | MM | SM | F1 (Macro) |
| *GPT-4o* | | | | |
| Zero-Shot | 0.184 | 0.227 | 0.378 | 0.195 |
| CoT | **0.287** | **0.325** | 0.491 | **0.301** |
| CoVe | 0.274 | 0.319 | 0.463 | 0.286 |
| CoQ | 0.263 | 0.320 | 0.481 | 0.279 |
| ToQ (Ours) | 0.285 (0.259–0.324) | 0.322 (0.299–0.355) | **0.522** (0.488–0.559) | 0.298 (0.274–0.339) |
| *DeepSeek-Chat* | | | | |
| Zero-Shot | 0.169 | 0.220 | 0.296 | 0.179 |
| CoT | 0.282 | **0.512** | 0.471 | 0.294 |
| CoVe | 0.253 | 0.314 | 0.417 | 0.271 |
| CoQ | 0.279 | 0.394 | 0.448 | 0.289 |
| ToQ (Ours) | **0.287** (0.253-0.308) | 0.324 (0.291-0.341) | **0.518** (0.492-0.538) | **0.297** (0.260-0.318) |

Table 4: Performance on MoreHopQA across prompting methods on GPT-4o and DeepSeek-Chat. EM, MM, SM, and F1 (Macro) are presented.

## 5 FQ-Score

### 5.1 Constituent Components

In this work, we also introduce Factuality with Quoting Score (FQ-Score), which comprises three evaluation dimensions: answer correctness (AC), quote attribution quality (QAQ), and step-wise attribution quality (SAQ). AC simply reflects a mean of the relevant metrics described in Section 4.2.

To capture the quality and integrity of the quoted evidence for QAQ, we use QUIP, which quantifies character n-gram overlap between a quoted span and the original pre-training corpus (*e.g.*, Wikipedia or PubMed). Higher QUIP scores indicate greater faithfulness and likelihood that the quote is verifiable. Additionally, if available, the number of retries required to achieve an acceptable quote (above a certain QUIP threshold), and the

length of the quote in words, also factor into QAQ (*i.e.*, excessive number of retries are penalized, and a greater QUIP score with larger quote content is rewarded).

To compute SAQ, we integrate two complementary signals: a language model–based evaluation of reasoning quality and a semantic similarity assessment across reasoning steps. First, motivated by prior studies using LLM-judges to evaluate reasoning (Hao et al., 2024; Wu and Cardie, 2025), we use GPT-4.1 to evaluate each step of the reasoning chain along four dimensions: (1) its relevance to the original question, (2) its usefulness in deriving the final answer, (3) the correctness of the subanswer, and (4) the completeness of reasoning. Each is scored on a 0–10 scale. Additionally, for each example as a whole, GPT-4.1 provides a binary judgment of whether the reasoning is sound and a 0–10 score assessing the overall logical flow across steps. These scores are combined into a single LLM-derived component, which is downweighted if the answer is judged to have been derived through unsound reasoning (see prompt in Appendix A.5).

Second, we compute semantic consistency between reasoning steps using cosine similarity of vectors from a sentence transformer (`all-MiniLM-L6-v2`). This includes both the similarity of each step to the original question and between adjacent steps to measure coherence. The average of these measures captures how well the reasoning chain maintains semantic alignment and progression. The final SAQ is obtained by aggregating the LLM scores and semantic similarity scores.

Finally, we propose a holistic FQ-Score, a unified factuality score designed to capture model performance across correctness, attribution, and step-level grounding. The FQ-Score is simply defined as:

$$\text{FQ-Score} = \alpha \cdot \text{AC} + \beta \cdot \text{QAQ} + \gamma \cdot \text{SAQ}$$

where $\alpha$, $\beta$, and $\gamma$ are tunable parameters, initially set to equal weights ($\alpha = \beta = \gamma = 1/3$) for balanced evaluation. FQ-Score can be computed per example and represents a single interpretable number that reflects the factual reliability of model-generated answers.

## 5.2 Clinical Reader Studies

To validate FQ-Score as a proxy for expert judgment in the medical domain, we conducted two preliminary clinical reader studies on MedQA (Jin et al., 2020), a benchmark based on United States Medical Licensing Exam (USMLE) questions. We first used 593 questions from USMLE Steps 2 and 3 in the official test set and evaluated model performance with ToQ using GPT-4o and DeepSeek-Chat (Table 5). While all baseline prompting methods already achieved strong performance, ToQ matched or very slightly exceeded them in accuracy and F1.

| **Method** | Accuracy ↑ | F1 ↑ |
|---|---|---|
| | *GPT-4o* | |
| Zero-Shot | 0.855 | 0.853 |
| CoT | **0.886** | 0.885 |
| CoVe | 0.867 | 0.865 |
| CoQ | 0.862 | 0.861 |
| ToQ (Ours) | 0.886 (0.871–0.913) | **0.886** (0.870–0.911) |
| | *DeepSeek-Chat* | |
| Zero-Shot | 0.816 | 0.814 |
| CoT | 0.766 | 0.771 |
| CoVe | 0.782 | 0.778 |
| CoQ | 0.825 | 0.812 |
| ToQ (Ours) | **0.841** (0.830–0.860) | **0.839** (0.828–0.859) |

Table 5: Accuracy and F1 on MedQA across prompting methods.

### 5.2.1 Reader Study Setup

Each study involved asking clinically trained evaluators of various backgrounds and experiences (from a board-certified radiation oncologist to a resident doctor in acute general medicine) to annotate 20 samples from the MedQA test set. The same samples were used between these clinicians to allow inter-rater agreement measurement. Evaluation guidelines were adapted from prior clinical natural language processing works (Van Veen et al., 2024; Xu et al., 2024).

In the first study, 5 clinicians reviewed only ToQ outputs. We heuristically selected a diverse set of examples — including correct/incorrect answers and strong/weak reasoning chains — to cover the full quality spectrum. For each reasoning step, clinicians were asked: (1) whether the subanswer was factually correct (`yes|no`), and (2) how relevant and useful the step was to the original question and final answer (0–10). They were further asked how logically consistent the entire reasoning chain was (0–10), and were instructed to penalize logical leaps or circular reasoning.

In the second study, 2 clinicians evaluated a blinded comparison of CoT, CoQ, and ToQ outputs selected based on a similar heuristic as in the first

study. Outputs were manually post-processed into a unified format to minimize method-identifying cues. Each sample was scored on a 0–10 scale for: completeness (whether it covered all clinically relevant considerations), correctness (factual accuracy of reasoning and answers), and logical flow (coherence and structure of the reasoning path).

#### 5.2.2 Correlations between Clinicians and with FQ-Score

In the first reader study, we observed low inter-rater reliability for the overall score, with an intraclass correlation coefficient (ICC) of 0.406 (95% CI: 0.21 to 0.65). This suggests considerable variability in clinician judgments regarding sample quality. Despite this, we adopt the mean opinion score (MOS) from clinicians as our reference standard, which is consistent with common practices in subjective evaluation domains.

Our model demonstrates a strong Pearson correlation with the clinician MOS ($r = 0.866$, $p < 0.001$), indicating close alignment with average expert judgment. However, this correlation should be interpreted with caution given the limited sample size and only moderate inter-rater agreement (Mukaka, 2012). To enhance the robustness of our conclusions, future work will include a broader panel of raters with a greater number of samples. The second reader study exhibited notably higher inter-rater agreement, particularly for the fine-grained completeness score. The mean score agreement reached an ICC of 0.539 (95% CI: 0.16 to 0.78), indicating moderate to substantial reliability among raters.

We provide a full breakdown of clinician inter-rater correlation values in Table 6, with visual comparisons plotted in Figure 2 for both reader studies. Additionally, Figure 3 shows the correlation between FQ-scores and clinician ratings from the first reader study.

Finally, Table 7 summarizes results across multiple methods evaluated in the second reader study. As not all methods included quotes, we computed FQ-Scores without the QAQ component, and evenly weighted AC and SAQ components. We observed that ToQ outputs are scored the highest by both FQ-Score and clinicians with strong correlations, providing support for FQ-Score to be a practical proxy for expert evaluations.

| **Metric** | ICC(2,1) ↑ | Pearson r ↑ | Spearman $\rho$ ↑ |
|---|---|---|---|
| *Reader Study 1* | | | |
| Logical Flow | 0.467 (95% CI, 0.27 0.69) | 0.484 | 0.428 |
| Overall Score[1] | 0.406 (95% CI, 0.21 0.65) | 0.407 | 0.382 |
| *Reader Study 2* | | | |
| Completeness | 0.617 (95% CI, 0.26 0.83) | 0.622 | 0.577 |
| Correctness | 0.457 (95% CI, 0.05 0.74) | 0.522 | 0.617 |
| Logical Flow | 0.435 (95% CI, 0.03 0.73) | 0.516 | 0.478 |
| Mean Score | 0.539 (95% CI, 0.16 0.78) | 0.592 | 0.628 |

Table 6: Inter-rater agreement metrics across both reader studies. [1]Overall Score represents the mean of the step relevance/usefulness scores for each step (halving those with an incorrect subanswer in the corresponding step), averaged with the logical flow score.

### 5.3 Human Attribution Evaluation

To accurately measure quote attribution directly, we also conducted a separate human evaluation study using the same 40 samples from our two clinical reader studies. Annotators manually checked whether each quote used in the model responses appeared in the online sources. In the first reader study with 20 ToQ samples, 37/45 quotes appeared in Wikipedia/PubMed, 7/45 appeared in another online source, and only 1 did not have any match. In the second reader study comparing methods, 14/17 quotes appeared in Wikipedia/PubMed for ToQ, while 3 appeared in another online source. None of the 3 quotes from the other methods appeared. This explicitly validates ToQ's improved quote attribution quality.

## 6 Discussion

Our results highlight that ToQ prompting meaningfully improves factual attribution and answer quality, especially on complex multi-hop questions. ToQ's quote-grounded step generation helps ensure that each reasoning step is not just a hallucinated bridge between premise and conclusion but is anchored by a verifiable quote. This structure encourages models to generate reasoning chains that are both faithful to their sources and logically coherent. Furthermore, the iterative retry-and-score mechanism ensures that the model does not progress based on low-confidence or unverifiable intermediate steps, effectively reducing compounding errors.

ToQ is also easily extensible: its quoting mechanism can be redirected to domain-specific sources in a retrieval-augmented manner, and external documents (*e.g.*, those provided with many multi-hop datasets or patient-specific context from the health record) can be embedded into the prompt as fixed

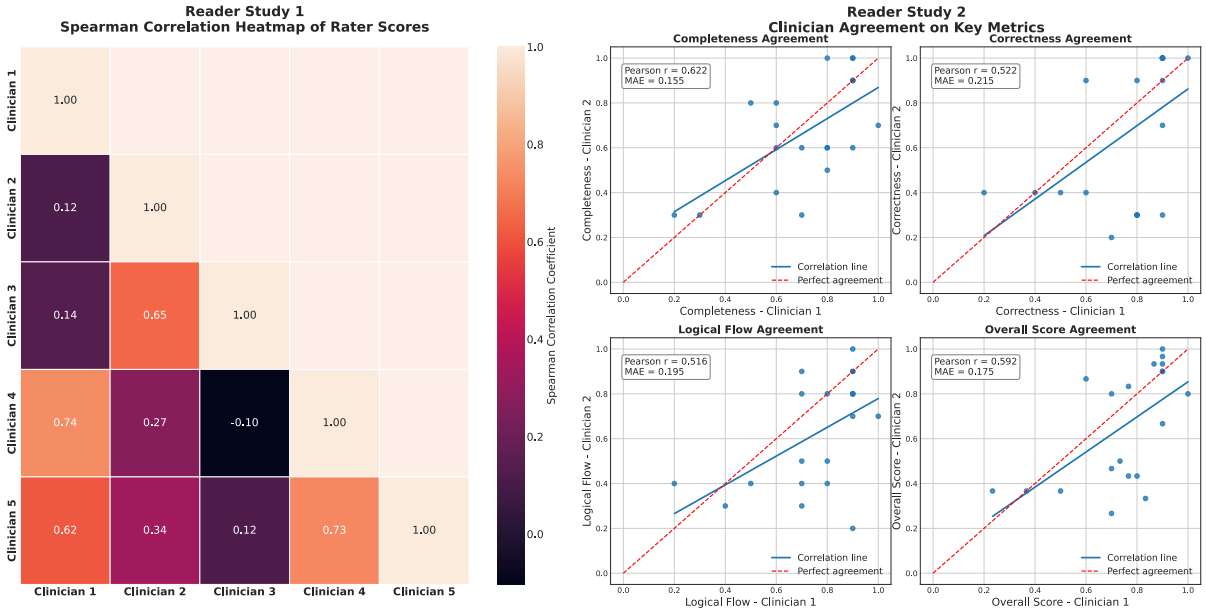


Figure 2: Inter-rater agreement and correlations between clinician raters for both reader studies.

| **Method** | Mean FQ-Score ↑ | Mean Clinician Score ↑ | Pearson r (p) ↑ | Spearman $\rho$ (p) ↑ | MAE ↓ |
|---|---|---|---|---|---|
| CoT | 0.591 | 0.790 | 0.758 (p=0.138) | 0.667 (p=0.219) | 0.199 |
| CoQ | 0.635 | 0.750 | **0.918 (p=0.028)** | 0.800 (p=0.104) | 0.144 |
| ToQ (Ours) | **0.770** | **0.817** | 0.889 (p=0.044) | **0.990 (p=0.001)** | **0.083** |

Table 7: Comparison of FQ-Scores, Clinician Scores, and their correlations between prompting methods in Reader Study 2.

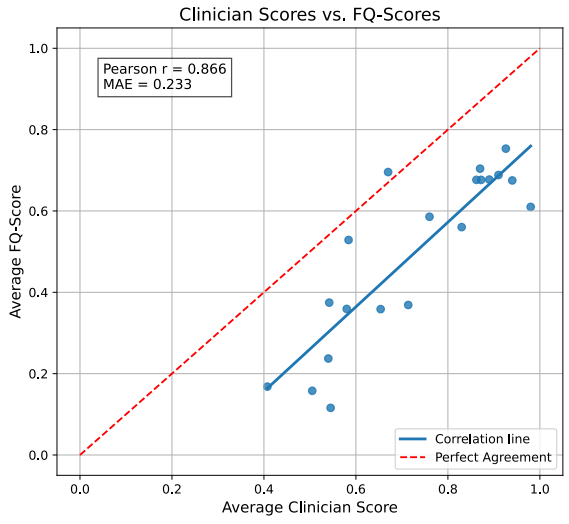


Figure 3: Correlation between FQ-Scores and Clinician Scores in Reader Study 1.

context. This modularity enables ToQ to be applied across both open-domain QA and closed-context reasoning settings, supporting greater interpretability and domain alignment.

We note that the benchmark performances can be sensitive to the specific combination of model and prompt design. Small changes — such as allowing the model to quote from Wikipedia vs. only on explicitly provided context — can noticeably affect outcomes. In some cases, quoting improves factual grounding, but this depends on the task setup. For example, if models are expected to rely solely on provided context, encouraging external referencing can actually hurt performance on the benchmark (*i.e.,* MuSiQue). Interestingly, although 2Wiki's task formulation includes accompanying context, referencing Wikipedia proves to be beneficial.

Finally, while ToQ's stepwise format may produce longer responses than baseline methods, which may be perceived as of higher quality, we note that in clinical contexts, verbosity is not inherently preferred — concise, focused reasoning is often the norm. ToQ is intended to support interpretability and teaching, not replicate chart-style brevity. However, we acknowledge that LLM

judges may favor verbosity, which we did not explicitly penalize for. Future iterations could incorporate length-normalized metrics or human-calibrated baselines to reduce this potential bias.

### 6.1 Dataset Examples

To better understand the value of quote-aware reasoning and adequate evaluation functions, we examined specific questions from the multi-hop datasets and call out two examples below.

Example 1

**Question:** What team does the 2018 PFA player of the year play for?
**Reference:** Egypt national football team
**Candidate:** Liverpool

While GPT-4o-ToQ's candidate "Liverpool" would be marked incorrect by most metrics (*i.e.*, EM, MM, SM, F1, etc.), it is arguably the more precise and contextually relevant answer. The PFA is an English football award, and in that context, the relevant affiliation is the player's English club team. At the time, Mohamed Salah — the PFA 2018's Player of the Year — played for both the Egypt national team and Liverpool, but only Liverpool is an English football club and the one that was actually mentioned in the dataset-provided context. Hence, Liverpool would be a more-than-reasonable answer. This highlights a fundamental limitation in current evaluation setups, where questions and reference answers may be incomplete or underspecified, penalizing valid predictions.

Example 2

**Question:** When did the maker of the Acura Legend, the manufacturer of Toyopet Master, and Nissan open US assembly plants?
**Reference:** 1981
**Candidate:** Honda, Toyota, and Nissan all opened their US assembly plants in the early 1980s

The model's response to this example is a broader but factually valid answer compared to the reference. In this case, both the specific year (1981) and the broader decade (early 1980s) were mentioned in the dataset-provided context. Since the question did not explicitly require an exact year (*e.g.*, by asking "`In which year...`"), the decade-level response is a fair interpretation.

These issues are especially pertinent in the high-stakes domain of medicine, where hallucinations impact patient care. In such settings, having a model that not only generates answers, but also quotes the source of each intermediate claim can dramatically improve trustworthiness and downstream usability. By design, ToQ is well-suited for these domains, enabling users to audit and interpret the reasoning chain in a fine-grained manner.

### 6.2 Small Language Models

To evaluate ToQ in lower-resource settings, we briefly applied it to Google's `gemma-3-12B-it` model (Team et al., 2025). While zero-shot ToQ was too complex for reliable outputs, performance significantly improved in a few-shot setup — particularly when using clinician-generated examples for answering MedQA.

## 7 Conclusion

In this work, we introduced Tree-of-Quote (ToQ) prompting, a novel prompting framework that tightly integrates quote attribution with stepwise reasoning. Unlike prior approaches that treat quoting and reasoning as separate stages, ToQ embeds quoting directly into the reasoning process and validates quotes through explicit scoring and retry loops. This results in more factually grounded and verifiable reasoning paths.

While ToQ makes attribution more explicit and controllable, failure modes remain. Quote retries can spiral if the model persistently fails to retrieve high-quality evidence, and subquestion generation may drift, producing shallow or redundant branches. Improving quote search robustness and adding dynamic pruning mechanisms could help prevent bloated or uninformative reasoning paths. Additionally, current evaluations rely heavily on reference-based metrics, which often fail to reward semantically correct yet differently phrased answers. A key direction is to develop more reliable, model-agnostic attribution checks — such as corpus-level quote tracing (*e.g.*, OLMOtrace (Liu et al., 2025)) — and to explore LLM-free alternatives to the scoring pipeline. Finally, ToQ's modularity makes it a candidate for integration with retrieval-augmented systems, domain-specific corpora, and real-time user-in-the-loop verification, all of which open paths toward more interactive and trustworthy AI assistants.

## Limitations

While ToQ improves factuality and attribution, several limitations remain. First, our evaluation scripts and metrics are still constrained by reference-based correctness, which, as shown in our case studies, can be brittle or incomplete. For instance, in the question: "Which missionary helped spread the religion widely practiced in region having the second largest rainforest in the world?" with the simple reference answer: "Sufi missionaries", GPT-4o-ToQ answered: "Sufi missionaries helped spread Islam, the widely practiced religion in Southeast Asia, which has the second largest rainforest in the world." This is a high-quality, well-supported answer, yet it fails exact match, scores poorly on F1, and may even be marked incorrect by semantic match due to wording differences, despite being a faithful and factually accurate response.

In addition, failure modes emerge in complex reasoning chains. If the quote extraction fails or yields a low-quality quote repeatedly, the pipeline may get "stuck" in a retry loop. Similarly, the subquestion generation logic can sometimes drift, producing redundant or trivial subquestions, leading to bloated reasoning chains that don't contribute new information. This mirrors limitations seen in other chain-of-reasoning methods, and highlights the need for better subquestion pruning and more adaptive control over when to terminate reasoning.

Lastly, compared to standard prompting, ToQ can also be more computationally expensive. For instance, ToQ averaged 2.26 LLM calls per question with GPT-4o, and 3.62 LLM calls per question with DeepSeek-Chat on multi-hop reasoning datasets. Overall, ToQ costs about 2x more than CoT in terms of wall clock time and tokens, but less than CoVe (Appendix B, Table 8). We hope to explore this cost tradeoff further in future work.

## Acknowledgments

JX gratefully acknowledges joint support from Canadian Institutes of Health Research (CIHR) Project ID 202410BCB-535721-77482 (Bioinformatics and Computational Biology), Nuffield Department of Medicine (NDM), and Oxford University Press (OUP). ZZ acknowledges support from Nuffield Department of Population Health (NDPH) and OUP. AYHL is a recipient of the University of Oxford Croucher Scholarship. MJ is supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project ID 389792660-TRR 248 (Foundations of Perspicuous Software Systems). DWE is supported by a Robertson Fellowship from the Big Data Institute (BDI). We also acknowledge contributions to the original project idea by Jack Jingyu Zhang of Johns Hopkins University (JHU).

## References

Sumera Anjum, Hanzhi Zhang, Wenjun Zhou, Eun Jin Paek, Xiaopeng Zhao, and Yunhe Feng. 2024. HALO: Hallucination Analysis and Learning Optimization to Empower LLMs with Retrieval-Augmented Context for Guided Clinical Decision Making. *arXiv preprint*. ArXiv:2409.10011 [cs] version: 2.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint*. ArXiv:2310.11511 [cs].

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2023. Factuality Challenges in the Era of Large Language Models. *arXiv preprint*. ArXiv:2310.05189 [cs].

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Xing Cao and Yun Liu. 2021. Coarse-grained decomposition and fine-grained interaction for multi-hop question answering. *arXiv preprint*. ArXiv:2101.05988 [cs].

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, and 68 others. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint*. ArXiv:2401.02954 [cs].

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv preprint*. ArXiv:2309.11495 [cs].

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large

language models using semantic entropy. *Nature*, 630(8017):625–630. Publisher: Nature Publishing Group.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to Plan and Generate Text with Citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling Large Language Models to Generate Text with Citations. *arXiv preprint*. ArXiv:2305.14627 [cs].

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *arXiv preprint*. ArXiv:2101.02235 [cs].

Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. 2024. LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models. *arXiv preprint*. ArXiv:2404.05221 [cs].

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55. ArXiv:2311.05232 [cs].

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint*. ArXiv:2009.13081 [cs].

Yanyang Li, Shuo Liang, Michael R. Lyu, and Liwei Wang. 2024. Making Long-Context Language Models Better Multi-Hop Reasoners. *arXiv preprint*. ArXiv:2408.03246 [cs].

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, YenSung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, Cassidy Trier, Aaron Sarnat, Jenna James, Jon Borchardt, Bailey Kuehl, Evie Cheng, Karen Farley, Sruthi Sreeram, Taira Anderson, and 12 others. 2025. OLMoTrace: Tracing Language Model Outputs Back to Trillions of Training Tokens.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint*. ArXiv:2203.11147 [cs].

M. M. Mukaka. 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal: The Journal of Medical Association of Malawi*, 24(3):69–71.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint*. ArXiv:2112.09332 [cs].

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A Comprehensive Overview of Large Language Models. *arXiv preprint*. ArXiv:2307.06435 [cs].

Tin Nguyen, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. 2025. HoT: Highlighted Chain of Thought for Referencing Supporting Facts from Inputs. *arXiv preprint*. ArXiv:2503.02003 [cs].

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o System Card. *arXiv preprint*. ArXiv:2410.21276 [cs].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv preprint*. ArXiv:1606.05250 [cs].

Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Vasan Srinivasan. 2024. Enhancing Post-Hoc Attributions in Long Document Comprehension via Coarse Grained Answer Decomposition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17790–17806, Miami, Florida, USA. Association for Computational Linguistics.

Abhilasha Sancheti, Koustava Goswami, and Balaji Srinivasan. 2024. Post-Hoc Answer Attribution for Grounded and Trustworthy Long Document Comprehension: Task, Insights, and Challenges. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 49–57, Mexico City, Mexico. Association for Computational Linguistics.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. MoreHopQA: More Than Multi-hop Reasoning. *arXiv preprint*. ArXiv:2406.13397 [cs].

Jiajun Shen, Tong Zhou, Yubo Chen, and Kang Liu. 2024. Citekit: A Modular Toolkit for Large Language Model Citation Generation. *arXiv preprint*. ArXiv:2408.04662 [cs].

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180. Publisher: Nature Publishing Group.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950. Publisher: Nature Publishing Group.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute First, then Generate: Locally-attributable Grounded Text Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs] version: 1.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop Questions via Single-hop Question Composition. *arXiv preprint*. ArXiv:2108.00573 [cs].

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142. Publisher: Nature Publishing Group.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint*. ArXiv:2201.11903 [cs].

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. "According to ...": Prompting Language Models Improves Quoting from Pre-Training Data. *arXiv preprint*. ArXiv:2305.13252 [cs].

Jingtian Wu and Claire Cardie. 2025. Reasoning Court: Combining Reasoning, Action, and Judgment for Multi-Hop Reasoning. *arXiv preprint*. ArXiv:2504.09781 [cs] version: 1.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint*. ArXiv:2501.09686 [cs].

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the First Shared Task on Clinical Text Generation: RRG24 and "Discharge Me!".

In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint*. ArXiv:2305.10601 [cs].

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. *arXiv preprint*. ArXiv:2305.16739 [cs].

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2024. LongCite: Enabling LLMs to Generate Fine-grained Citations in Long-context QA. *arXiv preprint*. ArXiv:2409.02897 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint*. ArXiv:1904.09675 [cs].

# A Prompts

## A.1 Tree-of-Quote Initialization Prompt

The following prompt was used for the Initialization step in Tree-of-Quote prompting. Provided sources are optional if the datasets do not include them.

**Initialization**

```
You are an initialization assistant for a reasoning system that aims to answer questions with
factually correct answers.

Your task is to output the first subquestion that helps identify a fact needed for
step-by-step reasoning. Start by uncovering any prerequisite knowledge.

For reference, you are provided with a list of sources that may contain relevant information.

Respond only in the following format:

<response>
  <type>Subquestion</type>
  <subquestion>[Your first subquestion to begin step-by-step reasoning]</subquestion>
</response>

---

<question>{question}</question>
<provided_sources>{sources}</provided_sources>
```

## A.2 Tree-of-Quote Quoting Prompt

The following prompt was used for the Quoting step in Tree-of-Quote prompting. Provided sources are optional if the datasets do not include them.

**Quoting**

```
You are a quoting agent in a multi-step factual reasoning system.

Your task is to answer a given subquestion using quotes from the provided sources (some of which
may be irrelevant) or Wikipedia (which is in your training data). Always include one accurate,
word-for-word quote and begin the reasoning with: According to [source], [quote].

Next, think step-by-step based on the quoted material while incorporating any provided prior
context. Continue the sequence without repeating earlier steps.

Respond only in the following format:

<response>
  <reasoning>
    [Begin with: According to [source], [quote].]
    [Provide your reasoning here.]
  </reasoning>
</response>

---

<original_question>{question}</original_question>
<provided_sources>{sources}</provided_sources>
<context>
{context}
</context>
<subquestion_to_answer>{subquestion}</subquestion_to_answer>
```

### A.3 Tree-of-Quote Questioning Prompt

The following prompt was used for the Questioning step in Tree-of-Quote prompting.

Questioning

```
You are a questioning agent in a multi-step factual reasoning system.

Your task is to:
1. Assess whether the current context contains sufficient factual information to answer the
original question.
2. If so, think step-by-step and then provide a concise final answer (not a full sentence).
3. If not, generate the next most informative subquestion.

Respond only in one of the following formats:

If context is sufficient:
<response>
  <action>Answer original question</action>
  <content>
    <explanation>[Brief step-by-step reasoning using the context]</explanation>
    <answer>The answer is: [Your concise answer to the original question]</answer>
  </content>
</response>

If more information is needed:
<response>
  <action>Generate subquestion</action>
  <content>
    <subquestion>[The next best subquestion to ask based on current context]</subquestion>
  </content>
</response>

---

<original_question>{question}</original_question>
<context>
{context}
</context>
```

### A.4 Semantic Match Evaluation Prompt

The following prompt was used for the semantic match (SM) metric during evaluation.

**Semantic Match**

```
You are a QA evaluator. Compare the candidate answer to the reference answer given the
question.

Question: {question}
Reference: {reference}
Candidate: {candidate}

Does the candidate answer express the *same meaning* as the reference answer? Some answers
may also contain additional information. If the reference is a subset of the candidate, or
if the candidate is a subset of the reference, they are considered to express the same
meaning.

### Special cases to consider:
- If both reference and candidate mention the same fact but at different levels of detail
(e.g. adding surrounding context), they are considered equivalent unless the question
specifically asks for a certain level of detail.
- Minor numeric discrepancies (e.g. rounding) may still be considered equivalent unless
materially different or the question specifies a certain level of precision.
- Answers referring to broader or more specific geographic/political regions
(e.g. city vs. state vs. country) should be evaluated for semantic overlap.
- If the candidate includes multiple possible answers, it must still contain the reference
to be a match.
- Slight differences in temporal expressions (e.g. year vs. decade) may be allowed if the
meaning is preserved, unless the question specifies a certain time frame.
- Differences in terminology (e.g. specific names vs. their parent organizations or aliases)
should be judged based on contextual equivalence.
- Partial elaboration or addition of related facts is acceptable if the core meaning aligns.

Reply with 'Yes' or 'No'.
```

### A.5 FQ-Score Step-Wise Attribution Quality Evaluation Prompt

The following prompt was used for the LLM component of the step-wise attribution quality (SAQ) subscore in FQ-Score.

FQ-Score SAQ

```
You are an expert in logical reasoning and explainability. Given a reasoning trace generated
in response to a question, your task is to rate each step based on:

1. Relevance to the original question (0-10)
2. Usefulness in deriving the final answer (0-10)
3. Correctness and factuality of the subanswer (0-10)
4. Completeness of the reasoning within the subanswer (0-10)

Please also give:
- An overall judgment on whether the answer was derived using sound reasoning (True/False)
- How logically consistent and valid the whole reasoning chain is (0-10). Consider the logical
flow of the reasoning and how well each step connects to the next, as well as if there are
gaps in logical flow or circular reasoning.

Instructions:
- Read the reasoning chain carefully.
- Provide your evaluation in the XML format as shown below.
- Be as harsh and rigorous as possible in your evaluation. No reasoning step is perfect, and
you should be able to find something to improve for each step.

Question: {question}

Reasoning Trace:
{context}

Respond in the following XML format:
<evaluation>
  <steps>
    <step>
      <step_text>...</step_text>
      <relevance>0-10</relevance>
      <usefulness>0-10</usefulness>
      <correctness>0-10</correctness>
      <completeness>0-10</completeness>
    </step>
    ...
  </steps>
  <final_answer_valid>true/false</final_answer_valid>
  <logical_flow>0-10</logical_flow>
</evaluation>
```

## B Computational Costs of ToQ

| **Method** | Average wall clock time per question (secs) ↓ | Approximate number of tokens per question (#) ↓ |
|---|---|---|
| | *GPT-4o* | |
| Zero-Shot | 1.134 (1.00x) | 10.82 (1.00x) |
| CoT | 3.058 (2.70x) | 113.09 (10.45x) |
| CoVe | 8.515 (7.51x) | 531.56 (49.13x) |
| CoQ | 3.324 (2.93x) | 135.04 (12.48x) |
| ToQ (Ours) | 9.621 (8.48x) | 189.71 (17.53x) |
| | *DeepSeek-Chat* | |
| Zero-Shot | 5.392 (1.00x) | 9.25 (1.00x) |
| CoT | 10.275 (1.91x) | 125.45 (13.56x) |
| CoVe | 75.246 (13.96x) | 924.68 (99.97x) |
| CoQ | 11.911 (2.21x) | 158.74 (17.16x) |
| ToQ (Ours) | 47.724 (8.85x) | 236.92 (25.61x) |

Table 8: ToQ costs on 2WikiMultiHopQA across prompting methods.