InteGround: On the Evaluation of Verification and Retrieval Planning in Integrative Grounding

Cheng Jiayang♠ **Qianqian Zhuang**♠ **Haorao Li**♠
Chunkit Chan♠ **Xin Liu**♠ **Lin Qiu**♡ **Yangqiu Song**♠
♠The Hong Kong University of Science and Technology
♡Shanghai Jiaotong University
{jchengaj, yqsong}@cse.ust.hk

Abstract

Grounding large language models (LLMs) in external knowledge sources is a promising approach to ensuring faithful and accurate predictions. While existing grounding approaches work well for simple queries, many real-world information needs require synthesizing multiple pieces of evidence. We introduce "integrative grounding" — the challenge of retrieving and verifying multiple interdependent pieces of evidence to support a hypothesis query. To systematically study this problem, we repurpose data from four domains for evaluating integrative grounding capabilities. Our investigation reveals two critical findings: First, when verifying groundedness, while LLMs are robust to redundant evidence, they tend to rationalize using their internal knowledge when the provided grounding information is incomplete. Second, in examining retrieval planning strategies, we find that undirected planning can degrade performance through the introduction of noise, while premise abduction emerges as a promising approach due to its logical constraints. Additionally, we observe that LLMs' zero-shot self-reflection capabilities consistently enhance grounding quality. These insights provide valuable directions for developing more effective integrative grounding systems. [1]

1 Introduction

Large language models (LLMs) are notorious for their tendency to hallucinate – generating content that appears plausible but is factually incorrect or unsupported (Ji et al., 2023; Zhang et al., 2023). To alleviate this issue, grounding LLMs to external knowledge sources has emerged as a promising approach. By anchoring model outputs to verifiable information (Min et al., 2023; Rashkin et al., 2023; Asai et al., 2023), grounding has enabled LLMs with more faithful decision-making and responsible generation.

In typical grounding setups, systems retrieve relevant documents in response to a query. While this approach has shown success for simple queries where a single piece of evidence suffices, many real-world information needs are inherently complex and require *synthesizing multiple pieces of evidence* to form a complete answer (Figure 1). For instance, answering questions about comparative analysis, multi-step reasoning (Yang et al., 2018; Trivedi et al., 2022), or claims requiring evidence from different sources (Min et al., 2023; Kamoi et al., 2023; Dalvi et al., 2021) often necessitates the integration of multiple pieces of information. We term this problem "*integrative grounding*": given a hypothesis query, a grounding system needs to retrieve multiple interdependent pieces of evidence to support it.

Despite its significance, the integrative grounding problem lacks systematic evaluation in current research. Existing work on Retrieval-augmented Generation (RAG) primarily focuses on end-to-end reasoning performance (Yao et al., 2022; Shinn et al., 2024), reasoning with pre-retrieved evidence (Fang et al., 2024), or post-generation evaluation in specific domains (Trautmann et al., 2024; Song et al., 2024). While these approaches implicitly address aspects of integrative grounding, they offer no comprehensive analysis of the grounding problem itself. Similarly, research in automated theorem proving (Dalvi et al., 2021; Sprague et al., 2022) explores complex inference chains but operates in restricted domains without robust evaluation of the broader integrative grounding challenges. While related fields like Natural Language Inference (NLI) and automated proof generation have their own evaluation paradigms, they typically operate under the assumption that sufficient evidence is provided. A critical gap remains in systematically evaluating grounding under the sub-optimal evidence conditions that are common in open-world settings. Our work addresses this gap by intro-

[1] Our code is available at https://github.com/HKUST-KnowComp/InteGround.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13587–13602
November 4-9, 2025 ©2025 Association for Computational Linguistics

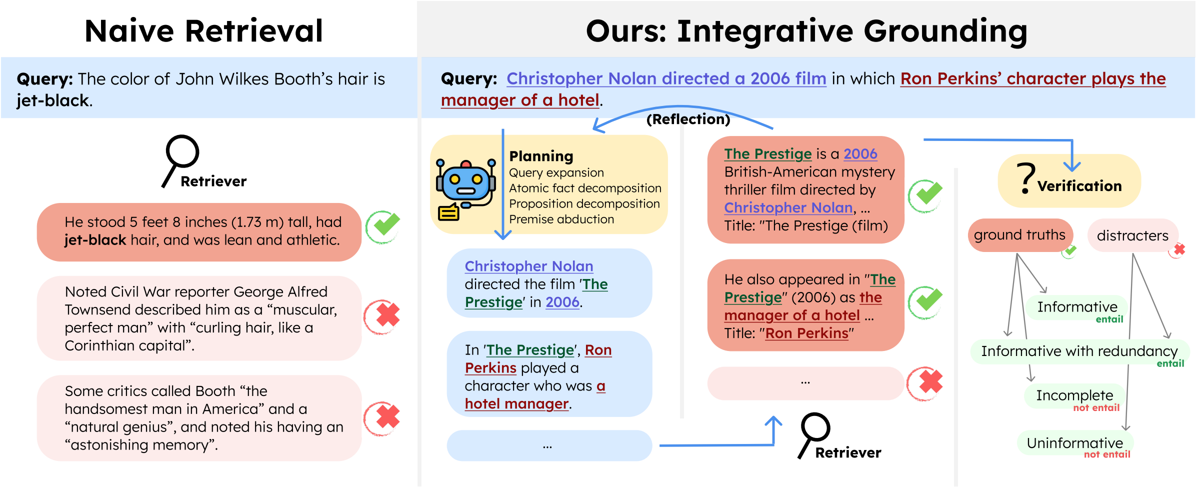
Naive Retrieval
Query: The color of John Wilkes Booth's hair is jet-black.
Retriever
He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair, and was lean and athletic.
Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".
Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory".
Ours: Integrative Grounding
Query: Christopher Nolan directed a 2006 film in which Ron Perkins' character plays the manager of a hotel.
(Reflection)
Planning
Query expansion
Atomic fact decomposition
Proposition decomposition
Premise abduction
Christopher Nolan directed the film 'The Prestige' in 2006.
In 'The Prestige', Ron Perkins played a character who was a hotel manager.
...
The Prestige is a 2006 British-American mystery thriller film directed by Christopher Nolan, ... Title: "The Prestige (film)
He also appeared in "The Prestige" (2006) as the manager of a hotel ... Title: "Ron Perkins"
...
Retriever
Verification
ground truths
distracters
Informative
entail
Informative with redundancy
entail
Incomplete
not entail
Uninformative
not entail

Figure 1: Overview of the integrative grounding problem.

ducing InteGround, a novel evaluation framework designed specifically for *integrative grounding*. Unlike standard benchmarks, InteGround systematically tests a model's behavior across four distinct evidence scenarios: complete, redundant, incomplete, and uninformative. This allows us to rigorously analyze critical failure modes like rationalization—where models invent facts to fill evidence gaps—that are often missed by traditional evaluations. We argue that robust integrative grounding is a crucial prerequisite for building truly reliable and faithful generative systems.

In this work, we focus on two critical aspects for evaluating integrative grounding. First, we investigate whether models can effectively *verify groundedness* by determining if multiple evidence pieces collectively support a query hypothesis (**RQ1**). Second, recognizing the challenges posed by reasoning dependencies among evidence, we examine *effective planning strategies* for LLMs to reformulate search queries and guide the retrieval process (**RQ2**).

To address these questions, we first construct an evaluation dataset from four domains to evaluate a system's ability to integrate multiple pieces of evidence. For RQ1, we conduct experiments on *groundedness verification*. Our experiments reveal that although LLMs are robust to redundant or distracting evidence, they exhibit a strong tendency to compensate for incomplete information by drawing on internal knowledge instead of strictly adhering to the retrieved content.

For RQ2, we present a systematic investigation of different *planning strategies for retrieval*, including decomposition-based, query expansion-based, and premise abduction-based approaches. Our findings reveal that planning does not universally improve retrieval performance. In fact, undirected planning can degrade performance by introducing noise, while decomposition-based planning shows limited improvement due to its conservative nature. Notably, we identify premise abduction as a particularly promising approach that shows consistent improvements and generalizes well to other datasets. This is likely due to its strong logical constraints, which encourage a directed expansion of the search space. Furthermore, we demonstrate that leveraging zero-shot self-reflection consistently enhances grounding quality across all planning strategies, highlighting the value of iterative refinement.

2 Preliminaries

Proposition. A proposition is a statement that has a truth value. For instance "The sky is green." is a proposition since it can be verified as true or false, while "Look up at the sky" is not a proposition. We use p to denote general propositions. For convenience, we use ϕ to denote a proposition that serves as hypothesis. In this paper, the queries are all propositions.

Knowledge Base (KB). A knowledge base $\mathcal{K} = \{p_i\}_{i=1,2,\ldots,K}$ is a set of consistent propositions.

Asking a KB. The $\texttt{Ask}_{\mathcal{K}}(p)$ operation queries KB $\mathcal{K}$ about proposition p, which returns three possible responses: `Entailment` ($\mathcal{K} \models p$), `Contradiction` ($\mathcal{K} \models \neg p$) , and `Contingent` (neither of the above). These two responses are considered as *informative* as they indicate that $\mathcal{K}$ contains related knowledge about p. Here, the operator $\models$ tells whether p follows logically from the premises in $\mathcal{K}$, where it is impossible for the premises to be true and p to be false. For example, ϕ="Socrates is mortal." *deductively* follows from $\mathcal{K}$={"All men are mortal.", "Socrates is a man."}, i.e. $\mathcal{K} \models \phi$, ($\texttt{Ask}_{\mathcal{K}}(p)$ = $\texttt{Entailment}$).

Grounding. Given a hypothesis proposition ϕ and a knowledge base $\mathcal{K}$, the task of *grounding* is to find a subset of consistent premises $\Sigma \subseteq \mathcal{K}$ through planning and retrieval, such that Σ is informative enough to ground the query hypothesis (i.e., $\Sigma \models p$ or $\Sigma \models \neg p$). In practice, Σ is obtained through the top-n retrieval results.

3 Constructing evaluation data

Our evaluation is based on data repurposed from two tasks where integrative grounding is required: multiple premise entailment (Lai et al., 2017; Dalvi et al., 2021; Kamoi et al., 2023) and multi-hop question answering (QA) (Yang et al., 2018; Trivedi et al., 2022).

3.1 Evaluation formulation

In the evaluation data (examples shown in Table 1), each hypothesis ϕ is accompanied by a set of ground-truth evidence $\Sigma^{gt} = \{p_1^{gt}, p_2^{gt}, \cdots\} \subseteq \mathcal{K}$ and a larger set of candidate evidence $\mathcal{K} = \Sigma^{gt} \cup \Sigma^{distr}$ that includes both the ground-truth and additional distracting facts.

For groundedness verification evaluation (Section 4), we test verification models' ability to accurately classify whether a target query ϕ is grounded by retrieval results Σ. For retrieval planning evaluation (Section 5), we assess how effectively different planning strategies retrieve relevant evidence. Given a hypothesis ϕ and a candidate evidence set $\mathcal{K}$, an integrative grounding system retrieves related evidence through query planning (as shown in Figure 1).

3.2 Data composition

We construct our evaluation dataset from four data sources, totaling 1,625 items. Dataset examples are shown in Table 1.

Multi-premise Entailment.

- EntailmentBank (Dalvi et al., 2021) contains hypotheses and corresponding multi-step entailment tree annotations from the science facts in WorldTree (Xie et al., 2020). We adapt the test set under task 2 setting[2] for our use, where the leaf nodes of entailment trees are kept as ground-truth evidence. Since task 2 already provides a set of (hard) distractors, we follow their setting and treat all candidate evidence for a given hypothesis as its corresponding KB, $\mathcal{K}$.

- WiCE (Kamoi et al., 2023) is a fine-grained textual entailment dataset linking natural claims and Wikipedia evidence. We consider both the claim and sub-claim level annotations, where for each claim several distinct groups of ground-truth evidence are annotated. We treat the evidence set for each claim as the KB $\mathcal{K}$, filtering out instances with more than 200 evidence items. Because WiCE often provides multiple valid evidence sets for a single hypothesis, we filter out instances where there is any single answer evidence (i.e., #GT == 1). Furthermore, to simplify evaluation, we select only the first group of ground-truth evidence ids as the ground truth. We only retain instances with the `supported` label.

Multi-hop QA. In the literature, a question-answer pair can be seen as a hypothesis (Dalvi et al., 2021). For multi-hop QA datasets, we prompt an LLM to transform each question and its corresponding answer into an equivalent hypothesis. As test sets are not always publicly available, we sample 500 instances from the validation sets of these datasets, specifically selecting those that require at least 3 pieces of evidence to answer.

- HotpotQA (Yang et al., 2018) is a dataset of question-answer pairs derived from Wikipedia, designed to evaluate complex reasoning and explanation generation. The questions in this dataset necessitate finding and reasoning over multiple supporting documents to formulate answers. To create evidence pieces, we append the document titles to the end of corresponding sentences. The evidences with titles that appear in the supporting facts are considered ground truths.

- MuSiQue (Trivedi et al., 2022) is created by composing questions from single-hop datasets. To create pieces of evidence, we concatenate the titles to the end of the corresponding paragraph texts to preserve context information. We treat

[2] `https://allenai.org/data/entailmentbank`. We use the split under `v3_May6_2022/entailment_trees_emnlp2021_data_v3/dataset/task_2/`.

Hypothesis ϕ	Candidate Evidence KB $\mathcal{K}$	Domain
Northern hemisphere will have the most sunlight in summer.	The northern hemisphere is a kind of hemisphere of earth. // A hemisphere of earth is a kind of place. // If a place is in summer, then it will have the most sunlight. If an object/something is in the sunlight then that object/that something will absorb solar energy. // Daylight is when the sun shines on a location. // The northern hemisphere is a kind of hemisphere of earth. // ...	Ent-Bank
Salih won the election with 219 votes to 22.	(meta data) TITLE: Iraq: Parliament elects Barham Salih as new president | News | Al Jazeera // Salih routed his main rival, Fuad Hussein, with 219 votes to 22. The Kurdish moderate politician has named veteran Shia politician Adel Abdul Mahdi as prime minister-designate. // Salih is a former deputy prime minister of the Iraqi federal government [Reuters] // ...	WiCE
Christopher Nolan directed a 2006 film in which Ron Perkins' character plays the manager of a hotel.	The Prestige is a 2006 British-American mystery thriller film directed by Christopher Nolan, from a screenplay ... from Christopher Priest's 1995 novel of the same name. Title: "The Prestige (film)" // Ron Perkins is an American actor who has been active since the early 1960s. Title: "Ron Perkins" // He also appeared in "The Prestige" (2006) as the manager of a hotel visited by Hugh Jackman's character in Colorado Springs, as well as ... Inland Empire is an internationally co-produced 2006 film written and directed by David Lynch. Title: "Inland Empire (film)" // The film is a co-production of France, Poland and the United States. Title: "Inland Empire (film)" // ...	HotpotQA
The maker of the Acura Legend, the manufacturer of the Scion xB, and Nissan opened US assembly plants in 1981.	The Acura Legend is a mid-size luxury/executive car manufactured by Honda. It was sold ... // The Scion xB is a compact car (subcompact car in its first generation) made by Toyota for the United States market and sold ... // ... A decade after the 1973 oil crisis, Honda, Toyota and Nissan, affected by the 1981 voluntary export restraints, opened US assembly plants and established their luxury divisions (Acura, Lexus and Infiniti, respectively) to ... The Nissan Rogue is a compact crossover SUV produced by the Japanese automaker Nissan. It made its debut in October 2007 for the 2008 model year... // The Acura EL is a subcompact executive car that was built at Hondaś Alliston, Ontario, plant... // ...	MuSiQue

Table 1: Example instances in the evaluation data. Related information in the Hypothesis and Candidate Evidence columns is color-coded for easier identification, using purple and blue. Distracting evidence are marked with gray. Due to large number of candidate instances, only part of distracting evidence are shown and the rest are left out. "Ent-Bank" is short for "EntailmentBank".

	Items	$\lvert\Sigma^{gt}\rvert$	$\lvert\mathcal{K}\rvert$
EntailmentBank	340	4.5±2.4	25.0±0.0
WiCE	285	2.8±0.9	85.2±43.4
HotpotQA	500	3.4±0.6	42.7±10.9
MuSiQue	500	3.4±0.5	20.0±0.1

Table 2: Statistics of INTEGROUND. $|\Sigma^{gt}|$ is the number of ground-truth snippets. $|\mathcal{K}|$ is the number of all candidate snippets.

the evidences corresponding to paragraph support indices as ground-truth evidences, and all others as distracting evidence.

3.3 Overview

Table 2 presents a comprehensive overview of all four data sources, detailing the number of items, and the means and standard deviations of ground-truth evidence ($|\Sigma^{gt}|$) and candidate evidence ($|\mathcal{K}|$) numbers.

As shown in Table 1 and 2, the evaluation data represent diverse domains and complexity levels. EntailmentBank, WiCE, HotpotQA, and MuSiQue each present distinct challenges—from logical deduction to information synthesis from news content to connecting facts across multiple sources. This diversity allows for comprehensive evaluation of integrative grounding capabilities across different contexts.

4 Groundedness verification

In this section, we aim to investigate whether models are capable of doing groundedness verification (RQ1).

4.1 Evaluation setup

Evaluation set creation Given candidate evidence set $\hat{\Sigma}_t$, we aim to test verification methods' classification performance on whether $\hat{\Sigma}_t$ provides informative enough clues for hypothesis ϕ.

We randomly sample from the retrieval datasets to construct such set. Specifically, we consider four cases: (1) informative ($\hat{\Sigma} = \Sigma^{gt}$); (2) informative with redundancy ($\hat{\Sigma} \supset \Sigma^{gt}$); (3) incomplete ($\hat{\Sigma} \subset \Sigma^{gt}$); (4) uninformative ($\hat{\Sigma} \not\subseteq \Sigma^{gt}$ and $\hat{\Sigma} \not\supseteq \Sigma^{gt}$). Case (1) and (2) have ground-truth label

	EntailmentBank				WiCE				HotpotQA				MuSiQue			
	Info.	Redun.	Inc.	Uninfo.	Info.	Redun.	Inc.	Uninfo.	Info.	Redun.	Inc.	Uninfo.	Info.	Redun.	Inc.	Uninfo.
NLI models																
NLI-xxlarge	76.8	75.9	83.5	93.2	45.6	42.1	97.5	99.6	82.2	76.4	93.2	96.0	43.6	26.6	98.0	98.6
NLI-xlarge	87.1	87.1	79.4	88.2	54.7	51.9	96.1	100.0	82.2	79.2	90.2	94.8	47.6	49.8	94.8	92.6
LLMs																
Llama3.1 8B Instr.	95.9	95.9	18.8	34.1	87.4	84.2	28.4	69.1	82.8	82.0	55.0	82.6	40.0	35.6	84.0	91.0
Llama3.1 70B Instr.	97.4	97.9	26.5	40.3	83.2	82.5	49.1	86.3	87.8	85.4	69.2	89.8	58.6	51.4	84.6	94.2
Claude3 Haiku	96.8	97.1	20.3	28.5	86.7	83.2	61.1	86.3	80.8	78.8	70.8	91.8	35.6	28.8	89.6	96.8
Claude3 Sonnet	99.7	99.4	16.8	43.5	83.2	84.6	67.0	91.6	86.6	84.0	65.4	88.0	47.6	42.4	86.6	94.6
Claude3.5 Sonnet	85.3	84.1	67.9	88.8	44.2	38.9	97.2	99.3	67.0	67.6	95.0	97.8	31.4	26.8	98.4	98.2
GPT-4o	97.4	97.1	34.4	54.4	68.4	63.5	85.3	96.8	82.2	85.4	74.8	92.2	48.0	51.6	84.2	95.2

Table 3: Accuracy (%) of verification methods on different types (Informative, Informative with Redundancy, Incomplete, Uninformative) of items across datasets. Since it is a two-way classification, chance accuracy is 50%.

	EntailmentBank			WiCE			HotpotQA			MuSiQue		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
NLI models												
NLI-xxlarge	86.8	76.3	81.2	96.9	43.9	60.4	93.6	79.3	85.9	95.4	35.1	51.3
NLI-xlarge	84.3	87.1	85.7	96.5	53.3	68.7	91.5	80.7	85.8	88.5	48.7	62.8
LLMs												
Llama3.1 8B Instr.	56.7	96.0	71.3	64.2	87.6	74.1	73.6	82.5	77.8	77.5	38.0	51.0
Llama3.1 70B Instr.	59.4	97.6	73.9	72.0	82.8	77.0	80.9	86.6	83.6	83.8	55.0	66.4
Claude3 Haiku	56.2	96.9	71.1	78.3	84.9	81.5	81.8	79.8	80.8	85.0	32.2	46.7
Claude3 Sonnet	58.8	99.6	73.9	80.2	83.9	82.0	78.5	85.3	81.8	82.7	45.0	58.3
Claude3.5 Sonnet	79.7	84.7	82.1	96.0	41.6	58.0	94.9	67.3	78.8	94.5	29.1	44.5
GPT-4o	63.6	97.2	76.9	88.1	66.0	75.4	83.5	83.8	83.7	82.9	49.8	62.2

Table 4: Classification performance (%) of verification methods.

informative (`Entailment`), while (3) and (4) are uninformative (`Not entailment`).

The evidence sets are created as follows. For (2), we randomly sample distractors ($\Sigma^{gt} \cup d$) and add into the ground-truth set. For (3), we randomly sample a strict subset from Σ^{gt}. For (4), we repeatedly sample a set from all candidate evidence until the set satisfy the condition ($\hat{\Sigma} \not\subseteq \Sigma^{gt}and\hat{\Sigma} \not\supseteq \Sigma^{gt}$).

Verification methods We evaluate verification methods that output two-way classification labels: {`Entailment, Not entailment`}. These two labels naturally exist in the multi-premise entailment benchmarks (Dalvi et al., 2021; Aghahadi and Talebpour, 2022; Kamoi et al., 2023).

We examine the characteristics of several verification methods, including (1) Natural language inference (NLI) models (He et al., 2021); (2) Large language models (LLMs), including GPT-4o (OpenAI, 2024), Claude-3, Claude-3.5 (Anthropic, 2024a), and Llama-3.1 Instruct (Meta, 2024a).

4.2 Results

Classification results are presented in Table 4. We also present per-type accuracies in Table 3. We have the following observations. We also investigate the impact of different prompt structures on verification performance; a detailed analysis in Appendix A.2.4 shows that while more complex prompts can help in specific cases, they do not consistently outperform a basic, direct prompt.

NLI models are precise verifiers. It is found that NLI models achieve high precision across datasets (Table 4). However, they suffer from low recall on WiCE$_{\text{claim}}$ and WiCE$_{\text{subclaim}}$. Further, NLI models are very conservative when predicting entailment (Table 3).

LLM verifiers tend to rationalize incomplete evidence with internal knowledge. From Table 3, it can be observed that LLMs are prone to classify incomplete evidence sets (Inc.) as "entailment", leading to much worse than random performance on EntailmentBank. This phenomenon is more pronounced in dataset with simpler languages (e.g., EntailmentBank). In contrast, supervised NLI classifiers are more conservative in terms of using internal knowledge. This behavior supports the notion that LLMs tend to fill gaps with their internal knowledge. This tendency to "fill in the gaps" highlights a significant reliability challenge. Future work could explore mitigation strategies, such as instruction-tuning models to explicitly forbid relying on internal knowledge, or incorporating more

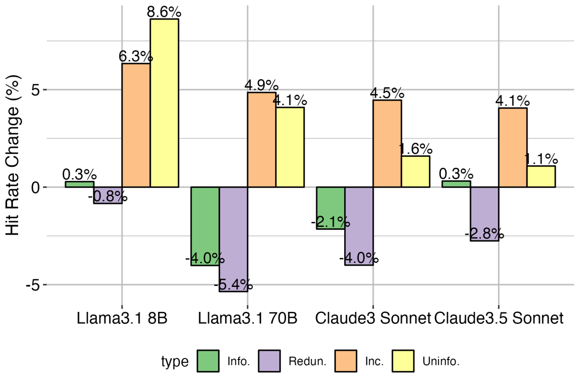

Figure 2: Hit Rate Change (%) of LLMs + NLI model predictions. The numbers are averaged across four sources in INTEGROUND.

conservative verification mechanisms, like the NLI models we evaluated, to act as a safeguard against such rationalization.

Redundant evidence has little impact on model predictions. Though sensitive to uninformative or incomplete information, models are relatively robust to redundant information.

Combining LLM and NLI predictions leads to more conservative judgments. We investigate the effect of incorporating NLI predictions into LLM prompts on verification performance. As shown in Figure 2, this ensemble approach improves LLMs' ability to identify incomplete and uninformative instances, though at the cost of reduced accuracy in detecting informative and redundant evidence. This suggests that NLI models' more stringent criteria for entailment can help counteract LLMs' tendency to rationalize, albeit with a trade-off in overall verification performance.

4.3 Qualitative analysis

A running example is presented in Table 5, showing how evaluation sets are constructed and how NLIs and LLMs perform. The NLI model, NLI-xlarge, shows consistent and precise verification capabilities across different evaluation sets. In contrast, LLMs like Llama3.1 8B and Claude models exhibit a tendency to rationalize incomplete evidence, as seen in the "Incomplete" set where they often incorrectly predict entailment. The "Redundancy" set demonstrates that additional, non-essential information has minimal impact on model predictions.

Hypothesis ϕ **and Evidences** $\mathcal{K}$	**Evaluation Sets** $\hat{\Sigma}_t$ **and Predictions**
Hypothesis: Northern hemisphere will have the most sunlight in summer.	***Informative:*** [k_1, k_2, k_3] – ENT NLI-xlarge: ENT Llama3.1 8B Instr.: ENT Claude3 Sonnet: ENT Claude3.5 Sonnet: ENT
Ground-truth Set: k_1: The northern hemisphere is a kind of hemisphere of earth. k_2: If a place is in summer, then it will have the most sunlight. k_3 A hemisphere of earth is a kind of place.	***Redundancy:*** [k_1, k_4, k_2, k_3] – ENT NLI-xlarge: ENT Llama3.1 8B Instr.: ENT Claude3 Sonnet: ENT Claude3.5 Sonnet: ENT
Sample distracter(s): k_4: Daylight hours means time during which there is daylight. k_5: Receiving sunlight is synonymous with absorbing sunlight. k_6 Period of daylight is synonymous with amount of daylight. k_7: Sunshine means sunlight.	***Incomplete:*** [k_1] – not ENT NLI-xlarge: not ENT Llama3.1 8B Instr.: ENT Claude3 Sonnet: ENT Claude3.5 Sonnet: not ENT ***Uninformative:*** [k_5, k_6, k_7] – not ENT NLI-xlarge: not ENT Llama3.1 8B Instr.: not ENT Claude3 Sonnet: ENT Claude3.5 Sonnet: not ENT

Table 5: A case study of verification results. Ground-truth evidences are marked with purple, and distracting evidences are marked with gray for easier identification. "ENT" and 'not ENT" are short for dataset labels "`Entailment`" and "`Not entailment`".

5 Retrieval planning for integrative grounding

In the sections above, we assume models are given a fixed set of queries (hypotheses) for retrieving evidence. Recent advancements of proof systems (Sprague et al., 2022; Tafjord et al., 2022; Weir and Van Durme, 2022) and LLM agents (Yao et al., 2022; Shinn et al., 2024) provide another perspective on this setting, where *planning* is integrated to proactively intervene retrieval processes. The objective of *planning* is to increase the success rate of grounding, i.e., biasing the search space so that it is more likely for an informative set Σ to be found.

The following part comprises experiments and discussions for three research questions: can planning and verification help retrieval and grounding? Moreover, how to optimize these components to maximize grounding performance?

5.1 Evaluation setup

Retrievers For sparse retrievers, we evaluate BM25 (Robertson et al., 2009). We also test dense retrievers, including MiniLM (miLM, Wang et al., 2020), Sentence T5 (ST5, Ni et al., 2021), and Microsoft E5-instruct (mE5, Wang et al., 2024). Given ϕ and $\mathcal{K}$, retrieval methods predict similarities and a ranking order of propositions in $\mathcal{K}$.

		No planning	Query Exp.	Fact Decomp.	Prop. Decomp.	Premise Abd.
EntailmentBank	BM25	64.4	53.6	**65.8**	64.9	56.3
	miLM	67.7	66.6	66.8	67.0	**68.3**
	mE5	66.8	64.9	66.7	66.7	**67.2**
	ST5	67.5	65.3	67.0	67.0	**67.6**
WiCE	BM25	**61.1**	48.5	56.9	58.5	52.4
	miLM	**58.1**	51.4	56.6	56.6	53.9
	mE5	**64.0**	58.8	61.5	62.6	61.7
	ST5	**63.3**	54.5	60.9	61.5	59.1
HotpotQA	BM25	67.5	65.8	68.7	68.7	**70.8**
	miLM	69.5	**73.9**	68.0	67.8	72.3
	mE5	**80.1**	79.3	75.2	75.8	79.7
	ST5	71.7	71.8	71.7	71.7	**74.0**
MuSiQue	BM25	60.9	64.1	57.8	58.4	**67.7**
	miLM	64.0	**70.9**	63.0	62.7	70.3
	mE5	65.2	**74.2**	64.6	65.3	73.7
	ST5	58.0	64.4	57.1	57.6	**66.6**

Table 6: Planning performance comparison based on Recall@5 (%). Best and second-best results are shown in **bold** and underlined, respectively. Results represent mean performance values across different LLMs.

Planning methods Given the retrieval history Σ_{t-1} and last step queries Φ_{t-1}, we define "planning" as reasoning to generate a new set of queries Φ_t to guide the next retrieval step.

$$\Phi_t \leftarrow \texttt{Plan}(\Phi_{t-1}, \Sigma_{t-1})$$

Specifically, the planners we use can be summarized as follows.

Planners that do not depend on retrieval history ($\texttt{Plan}(\Phi_{t-1}, \varnothing)$), including:

- **Query expansion** (Gao et al., 2023; Wang et al., 2023). This line of work expand writing based on the input query with an LLM. We adopt official prompts from HyDE (Gao et al., 2023).

- **Atomic fact decomposition** (Min et al., 2023; Kamoi et al., 2023). The hypothesis text is decomposed into multiple atomic factoids with a few-shot prompt. We reuse prompts from (Min et al., 2023).

- **Proposition decomposition** (Chen et al., 2022). Similar to atomic fact decomposition, proposition decomposition breaks down the input hypothesis text into multiple propositions. Since many have found that LLMs achieve reasonably good performance when prompted with few-shot examples (Min et al., 2023; Kamoi et al., 2023; Chen et al., 2024), we use the prompts provided in (Chen et al., 2024).

- **Premise abduction** (Tafjord et al., 2022). Given an input hypothesis text, premise abduction methods generate all premises required to entail the hypothesis through abduction. We curate few-shot prompts with examples from EntailmentBank (Tafjord et al., 2022).

In addition, we also evaluate planners that take both input hypothesis and planning history as inputs ($\texttt{Plan}(\Phi_{t-1}, \Sigma_t)$). This group of planners are closely related to the agentic behaviors of **self-reflection** (Yao et al., 2022; Khot et al., 2022; Shinn et al., 2024). Intuitively, integrative grounding may benefit from adjusting the queries to missing information and past queries. In our setup, this is implemented by prompting the LLM to analyze the retrieved evidence from the previous step, identify missing information, and generate a new, more targeted set of queries to guide the next retrieval iteration. The full prompt can be found in Appendix A.2.3 (Table 11). Intuitively, integrative grounding may benefit from adjusting the queries to missing information and past queries.

We prompt three state-of-the-art LLMs, GPT-4o, Claude-3.5 Sonnet and Llama-3.1 70B Instruct models as the base LLMs for planning evaluation. Each planning step produces multiple queries (in Φ_t). Following previous work in informative retrieval (Gao et al., 2023; Wang et al., 2023), we concatenate them as the query for the next step retrieval. The implementation details are in Appendix A.2.

5.2 Results

We compared the retrieval results of directly feeding hypotheses to retrievers ("No planning") against the performance of feeding both hypotheses and rewritten plans to retrievers. The main experimental results are presented in Table 6.

Adding planning modules to refine the queries does not always help. In some cases, it can even hurt performance: For instance, query expansion methods almost always led to decreased performance compared to no planning. This suggests that arbitrary query rewriting and expansion can introduce noise that hinders effective retrieval (Weller et al., 2023).

Limited impact of decomposition-based planning: Notably, planning based on atomic fact decomposition (Min et al., 2023) or proposition decomposition (Chen et al., 2022) showed little improvement in grounding performance. We hypothesize that this is because decomposition-based planning does not introduce new information to the queries, potentially resulting in retrieved results that overlap significantly with the no-planning baseline.

Abduction-based planning shows significant improvement: Among the four planning methods tested, Premise Abduction performed best. Although this method's intuition stems from strict textual entailment data (Dalvi et al., 2021), it appears to generalize well to broader datasets. This success could be attributed to the directed nature of such planning methods. Compared to "Query expansion" planning, premise abduction imposes an additional logical reasoning constraint (i.e., the possible premises of the hypothesis) to expand the search space effectively.

Self-reflection enhances integrative grounding: As illustrated in Figure 3, incorporating a zero-shot self-reflection step consistently improved the integrative grounding task. Most planning methods surpassed the No-planning baseline after the reflection step. Notably, query expansion and decomposition-based planning methods showed the greatest improvements. This may be because reflection helps mitigate the weaknesses of other planners: it provides a 'directed' bias that undirected query expansion lacks, and it introduces new information and context that conservative decomposition methods do not generate on their own.

5.3 Qualitative analysis

The example in Table 7 illustrates the workings of different planning methods. The proposition decomposition method breaks down the hypothesis into four key components, focusing on the relationship between the northern hemisphere, seasons,

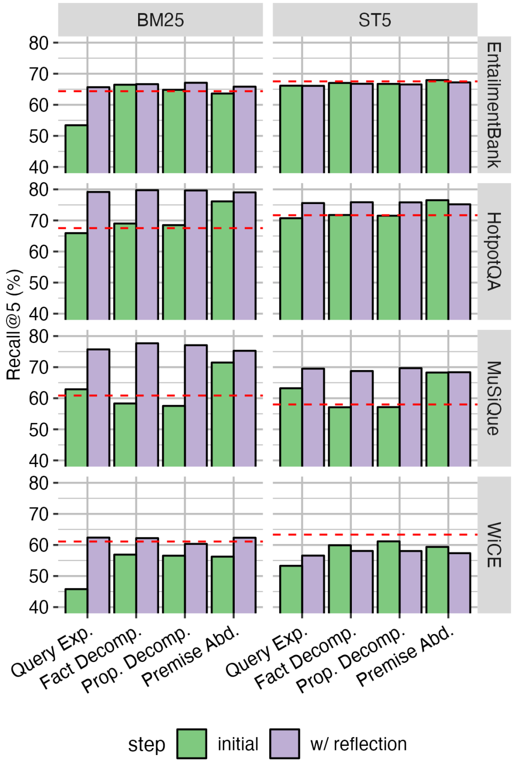

Figure 3: Performance comparison of planning with ("w/ reflection") and without ("initial") reflection step using Recall@5 (%). Dashed red lines indicate baseline retrieval performance without planning.

Planning	Retrieval results
Hypothesis: Northern hemisphere will have the most sunlight in summer.	
Proposition decomposition: •The northern hemisphere has more sunlight in summer. •The northern hemisphere experiences seasons. •Summer is a season in the northern hemisphere. •Sunlight varies by season in the northern hemisphere.	•The northern hemisphere is a kind of hemisphere of earth. •If a place is in summer, then it will have the most sunlight. •A hemisphere is a part of earth. •Being in the sun is synonymous with being in the sunlight. •A hemisphere of earth is a kind of place.
Atomic fact decomposition: •The northern hemisphere experiences seasons. •Summer is one of the seasons. •The northern hemisphere receives sunlight. •The northern hemisphere receives the most sunlight in summer.	•The northern hemisphere is a kind of hemisphere of earth. •If a place is in summer, then it will have the most sunlight. •Being in the sun is synonymous with being in the sunlight. •A hemisphere is a part of earth. •A hemisphere of earth is a kind of place.

Table 7: Illustration of the planning process. The "Planning" column shows representative queries generated by Claude-v3.5-sonnet. For brevity, only proposition decomposition and atomic fact decomposition are presented. The "Retrieval" column displays evidence pieces retrieved using BM25. Ground truth statements are highlighted in purple.

and sunlight variation. Similarly, the atomic fact decomposition method generates four fundamental statements about these elements. Notably, neither method introduces new information beyond the original hypothesis. In this case, both methods capture all three ground truth evidences.

6 Related work

Natural Language Proof Generation. Prior work has developed proof writing algorithms that generate proof trees based on models' internal knowledge (Tafjord et al., 2022; Sprague et al., 2022). While NELLIE (Weir and Van Durme, 2022) incorporates retrieved facts for hypothesis decomposition, these approaches typically operate in restricted domains. More recent work has also focused on enhancing this process using principles from informal logic to improve decompositional inference (Weir et al., 2024).

Fact Verification with LLMs. Recent work on fact-checking LLM-generated content (Min et al., 2023; Tang et al., 2024; Rashkin et al., 2023) primarily focuses on single-premise verification. Other studies have investigated how well LLMs ground their outputs in provided sources, confirming that even state-of-the-art models struggle with faithfully adhering to evidence (Lee et al., 2023), especially when it contains conflicting information (Jiayang et al., 2024). While related to our verification component, our work focuses on a comprehensive evaluation of integrative retrieval and planning.

RAG and Multi-hop QA Agents. Recent LLM agents for multi-hop question answering (Yao et al., 2022; Shinn et al., 2024) employ iterative retrieval strategies but focus primarily on reasoning rather than addressing integrative grounding challenges. Methods like TRACE (Fang et al., 2024) construct reasoning chains from already-retrieved evidence, whereas our approach emphasizes dynamically planning what to retrieve next based on evidence interdependencies. Similarly, while studies like (Trautmann et al., 2024) focus on post-generation evaluation and others (Song et al., 2024) enhance citation quality, we address verification during the retrieval process to evaluate whether multiple documents collectively support a hypothesis.

7 Conclusion

In this work, we introduce "integrative grounding" as a critical challenge for LLMs and provide a systematic evaluation framework, InteGround, to assess it. Our investigation yields several key insights with direct implications for building more reliable systems. We demonstrate that while LLMs are robust to redundant evidence, they exhibit a strong tendency to "rationalize" with internal knowledge when faced with incomplete information, posing a significant risk to faithfulness. In retrieval planning, we find that intuitive strategies like undirected query expansion can degrade performance by introducing noise, whereas logically constrained methods show significant promise. Notably, premise abduction prove effective by expanding the search space in a directed manner, and zero-shot self-reflection consistently improve performance across all planning methods by enabling iterative refinement. Our findings offer direct guidance for building more robust grounding systems, with a detailed discussion of practical applications for RAG pipelines provided in Appendix A.2.5.

Limitations

Limited ground-truth setting. This study assumes that a single, unique ground-truth evidence set. However, in many real-world scenarios, multiple valid ground-truth evidence sets may exist to support a hypothesis.

Grounding to structured data. Our evaluation is restricted to the textual domain. Grounding to structured or semi-structured data, such as tabular data or knowledge graphs, is also an important and promising direction for future research.

Limited language. We primarily focus on English corpora, meaning our findings may not generalize to other languages. Evaluation of grounding on multilingual corpora is left for future work.

Evaluation-centric approach. Our work prioritizes the comprehensive evaluation of integrative grounding rather than direct application development. While our findings have implications for RAG systems, translating these insights into optimized real-world applications requires additional engineering effort beyond the scope of this study.

Limited Scope of Planners. Our study focuses exclusively on LLM-driven planning strategies. A comparison with established non-LLM planning techniques, such as classical symbolic planners, was beyond our scope but remains an important direction for future comparative analysis.

Ethical statement

Our work primarily utilizes open-source retrieval models, datasets, and publicly available Large Language Models (LLMs). Given the nature of our research context, the outputs generated by these LLMs are unlikely to contain harmful or dangerous information. We have carefully considered the ethical implications of our study and foresee no significant concerns or potential risks associated with our methodology or findings.

Acknowledgements

The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong, SAR, China. We would like to express our sincere gratitude to all the reviewers for their invaluable contributions through their comments and suggestions.

References

Zeinab Aghahadi and Alireza Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. *Journal of Applied Non-Classical Logics*, pages 1–17.

AI Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

AI Anthropic. 2024b. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

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*.

Elias Bassani. 2022. ranx: A blazing-fast python library for ranking evaluation and comparison. In *ECIR (2)*, volume 13186 of *Lecture Notes in Computer Science*, pages 259–264. Springer.

JC de Borda. 1781. M'emoire sur les' elections au scrutin. *Histoire de l'Acad'emie Royale des Sciences*.

Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, Dan Roth, and Tal Schuster. 2022. Propsegment: A large-scale corpus for proposition-level segmentation and entailment recognition. *arXiv preprint arXiv:2212.10750*.

Sihao Chen, Hongming Zhang, Tong Chen, Ben Zhou, Wenhao Yu, Dian Yu, Baolin Peng, Hongwei Wang, Dan Roth, and Dong Yu. 2024. Sub-sentence encoder: Contrastive learning of propositional semantic representations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1596–1609.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. *arXiv preprint arXiv:2406.11460*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Cheng Jiayang, Chunkit Chan, Qianqian Zhuang, Lin Qiu, Tianhang Zhang, Tengxiao Liu, Yangqiu Song, Yue Zhang, Pengfei Liu, and Zheng Zhang. 2024. Econ: On the detection and resolution of evidence conflicts. *arXiv preprint arXiv:2410.04068*.

Cheng Jiayang, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, et al. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. *arXiv preprint arXiv:2310.12874*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Alice Lai, Yonatan Bisk, and Julia Hockenmaier. 2017. Natural language inference from multiple premises. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 100–109.

Hyunji Lee, Sejune Joo, Chaeeun Kim, Joel Jang, Doyoung Kim, Kyoung-Woon On, and Minjoon Seo. 2023. How well do large language models truly ground? *arXiv preprint arXiv:2311.09069*.

AI Meta. 2024a. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

AI Meta. 2024b. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

OpenAI. 2024. Hello gpt-4o. *OpenAI*.

David L Poole and Alan K Mackworth. 2010. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2024. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse. *arXiv preprint arXiv:2409.11242*.

Zayne Sprague, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction with incomplete information. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8230–8258.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2022. Entailer: Answering questions with faithful and truthful chains of reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Oguzhan Topsakal and Tahir Cetin Akinci. 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, volume 1, pages 1050–1056.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Alan Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. *arXiv preprint arXiv:2410.08764*.

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. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Nathaniel Weir, Kate Sanders, Orion Weller, Shreya Sharma, Dongwei Jiang, Zhengping Zhang, Bhavana Dalvi Mishra, Oyvind Tafjord, Peter Jansen, Peter Clark, et al. 2024. Enhancing systematic decompositional natural language inference using informal logic. *arXiv preprint arXiv:2402.14798*.

Nathaniel Weir and Benjamin Van Durme. 2022. Dynamic generation of grounded logical explanations in a neuro-symbolic expert system. *arXiv preprint arXiv:2209.07662*.

Orion Weller, Kyle Lo, David Wadden, Dawn Lawrie, Benjamin Van Durme, Arman Cohan, and Luca Soldaini. 2023. When do generative query and document expansions fail? a comprehensive study across

	ϕ	Σ^{gt}	$\mathcal{K}$
EntailmentBank	11.8±4.3	9.5±4.3	9.4±5.0
WiCE	25.9±11.7	21.7±16.2	12.5±13.8
HotpotQA	21.4±7.4	34.4±13.7	33.1±13.3
MuSiQue	25.9±6.3	110.2±63.9	105.2±59.0

Table 8: Mean and standard deviation for numbers of tokens in INTEGROUND.

methods, retrievers, and datasets. *arXiv preprint arXiv:2309.08541*.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. Worldtree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Tianshi Zheng, Jiayang Cheng, Chunyang Li, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Y Wong, and Simon See. 2025. Logidynamics: Unraveling the dynamics of logical inference in large language model reasoning. *arXiv preprint arXiv:2502.11176*.

A Appendix

A.1 Details in the construction of INTEGROUND

The number of tokens are presented in Table 8.

A.2 Experimental details

A.2.1 Baselines

Retrieval methods:

- BM25 (Robertson et al., 2009). We use the `rank-bm25` python package to implement the algorithm.
- Sentence-transformers. We use the LangChain(Topsakal and Akinci, 2023) implementation to embed corpus, and the cosine similarities between embeddings as the similarity for retrieval.

Verification and planning methods:

- NLI. We use the state-of-the-art NLI models (He et al., 2020), including DeBERTa (`xlarge`) and DeBERTa-v2 (`xxlarge`). Given a pair of texts, NLI models output probabilities over entailment, contradiction, and neutral (ENT, CON, NEU).
- LLMs. We use the state-of-the-art LLMs, including Llama 3.1 8B Instruct, Llama 3.1 70B Instruct (Meta, 2024b), Claude 3 Haiku, Claude 3 Sonnet, and Claude 3.5 Sonnet (Anthropic, 2024b). Llama and Claude models are accessed through Amazon Bedrock.

A.2.2 Experimental settings

Direct retrieval Ranking evaluation. We use the `ranx` (Bassani, 2022) [3] package for computing ranking metrics in the evaluation of retrievers. In the main evaluation, F1@5 and Acc@5 are reported.

Stepwise retrieval We add ground-truth evidence step by step to the hypothesis.

Combining planning and retrieval In this setting, we generate multiple rankings by retrieving with different sub-queries. To consolidate these rankings into a single ranking, we address it as a rank aggregation problem (Dwork et al., 2001). We implement Borda's rank aggregation strategy (Borda, 1781) to produce a unified rank.

A.2.3 LMs prompting details

The prompt templates for LLMs in this research are presented in Table 10 for hypothesis generation and verifications and Table 11 for plannings.

A.2.4 Additional experiment on prompt structures

As shown in Table 12, we test GPT-4o models with different prompts on the groundedness verification task.

Basic Prompt (in the paper): Simple instruction to assess if evidence supports a hypothesis query.

"You are a helpful logical reasoner. Please help classify a hypothesis with {labels} based solely on

[3] https://github.com/AmenRa/ranx

	EntailmentBank		WiCE		HotpotQA		MuSiQue	
	F1@5	Acc@5	F1@5	Acc@5	F1@5	Acc@5	F1@5	Acc@5
BM25	52.5	34.1	40.7	30.5	53.3	26.6	48.4	15.0
SimCSE$_{\text{RoBERTa}}$	50.2	29.7	38.8	25.6	50.4	21.8	40.6	8.6
MiniLM-L6	55.3	36.2	38.5	25.6	55.1	27.8	50.9	17.6
ST5$_{\text{large}}$	55.1	37.1	**42.0**	31.2	56.7	32.8	46.1	14.2
GTR$_{\text{T5-large}}$	**57.6**	**40.0**	41.3	**32.3**	56.6	32.8	50.9	18.8
mE5$_{\text{large-instruct}}$	54.5	37.1	41.2	30.2	**62.8**	**48.0**	**54.4**	**24.2**

Table 9: Direct retrieval results.

a set of evidence. Evidence set: {e1} Hypothesis: {e2}"

Structured Reasoning Prompt: Include explicit steps for verification (check completeness, redundancy, etc.)

"Evidence set: {e1} Hypothesis: {e2} Assess whether the evidence is sufficient to support the query by checking: 1. Relevance: Is all the evidence relevant to the query? 2. Completeness: Does the evidence contain all necessary information to address the query? 3. Redundancy: Is there unnecessary repetition in the evidence? Based on this assessment, is the provided evidence sufficient to support the hypothesis? Briefly explain your assessment, then choose your answer among {labels}."

Chain-of-Thought Prompt: Ask the model to think step-by-step before concluding

"Evidence set: {e1} Hypothesis: {e2} Think step by step to determine if the provided evidence is sufficient to support the hypothesis. The following steps are an example: - What key information does the query require? - What information does the evidence provide? - What information, if any, is missing? - What additional evidence would be needed to fully address the query? After thinking step by step, determine if the provided evidence is sufficient to support the hypothesis. Choose your answer among {labels}."

We have the following observations:

- Prompts with complex structures (Structured, CoT) enhance models' detection of incomplete information, but often reduce accuracy on "Informative" labels—suggesting a potential overthinking effect.
- Despite this trade-off, overall classification performance (F1) remains robust across prompting schemes, with complex prompts did not outperform the Basic prompting, and in some domains, performed significantly worse."

A.2.5 Applications to RAG Systems

Our research offers significant practical applications for enhancing RAG systems. First, our analysis of planning strategies provides actionable methods for improving evidence retrieval in production pipelines. Specifically, premise abduction addresses cases with incomplete evidence by generating plausible intermediate premises that guide subsequent retrievals. Similarly, fact decomposition simplifies complex queries, substantially improving retrieval accuracy in noisy information environments.

Second, our groundedness verification findings directly inform RAG system design. The observed tendency of LLMs to rationalize when evidence is incomplete underscores the need for dedicated verification mechanisms to detect and mitigate hallucinations. Our evaluation framework offers a approach for evaluating the effectiveness of such safeguards in practical applications.

Finally, the performance disparities across different verification strategies provide clear guidance for RAG system architecture decisions. By incorporating these insights, developers can create more reliable systems that not only retrieve relevant information but also accurately assess whether the retrieved evidence collectively supports the generated content.

A.3 Grounding systems

A grounding system serves the goal of finding a subset Σ from a KB $\mathcal{K}$, given a hypothesis ϕ. Although there may be various ways to achieve this goal, there are common stages among all the grounding systems: the *planning* stage which involves reasoning over the hypothesis, and the *linking* stage which retrieves candidates from $\mathcal{K}$ and verifies the

Function	Inputs	Prompt
Hypothesis Generation	q: question a: answer	Paraphrase the given question and answer pair to a proposition. Your response should be formatted as {{"Proposition": "PROPOSITION TEXT"}}. Question: When did the maker of the Acura Legend, the manufacturer of Toyopet Master, and Nissan open US assembly plants? Answer: 1981 {{"Proposition": "The maker of the Acura Legend, the manufacturer of Toyopet Master, and Nissan opened US assembly plants in 1981."}} Question: Signed with Maybach Music Group in 2011, which artist was featured as a guest in Fire of Zamani? Answer: Wale {{"Proposition": "Wale, who signed with Maybach Music Group in 2011, was a featured guest artist on Fire of Zamani."}} Question: $\{q\}$ Answer: $\{a\}$
Verification (LLMs)	e_1: evidence set e_2: hypothesis	You are a helpful logical reasoner. Please help classify a hypothesis with {labels} based solely on a set of evidence. Evidence set:$\{e_1\}$ Hypothesis: $\{e_2\}$ Result in JSON format (e.g. {{"label": "{labels}"}}):
Verification (NLIs+LLMs)	$label$: NLI's prediction e_1: evidence set e_2: hypothesis	You are a helpful logical reasoner. Please help classify a hypothesis with {labels} based solely on a set of evidence. For your reference, an external supervised Natural Language Inference model's prediction is: $\{label\}$. Evidence set: $\{e_1\}$ Hypothesis: $\{e_2\}$ Result in JSON format (e.g. {{"label": "{labels}"}}):

Table 10: Prompts for LLMs: Hypothesis Generation and Verifications

groundedness of such candidates set. In the literature of logical reasoning (Poole and Mackworth, 2010), forward chaining and backward chaining provides insights on the possible implementations of the stages.

Suppose each grounding has at most T ($T \geq 1$) steps. Let Σ_t denote the grounded set and Φ_t denote the hypotheses tree at time step t ($t \in \{0, 1, \cdots, T\}$). Initially, $\Sigma_0 = \{\}$ and Φ_0 contains only the root node ϕ.

Linking. The system first conduct linking to update the candidate set:

$$\hat{\Sigma}_t \leftarrow \texttt{Retrieve}(\Sigma_{t-1}, \Phi_{t-1}, \mathcal{K})$$

The candidate set is then judged for testing whether the hypotheses are consistent with, using the $\texttt{Ask}$ function, where the qualifies subset is retained

$$\Sigma_t \leftarrow \texttt{Verify}(\texttt{Ask}_{\hat{\Sigma}_t}(\Phi_{t-1}))$$

The linking process trigger exiting condition when $\texttt{Ask}_{\Sigma_t}(\Phi_{t-1})$ returns *informative* response for all the leaf nodes in Φ_{t-1}.

Essentially, forward chaining is applied to test whether the hypothesis follows Σ_t.

Planning. Given Σ_t and Φ_{t-1}, a grounding system do reasoning to update the hypotheses tree so as to guide the next linking step.

$$\Phi_t \leftarrow \texttt{Plan}(\Sigma_t, \Phi_{t-1})$$

Although how to implement the reasoning function here is up to each grounding system's design, this reasoning stage is essentially backward-chaining. Developing robust reasoning functions is a significant challenge, as LLMs often fall short on complex cognitive tasks like story-level analogy (Jiayang et al., 2023; Zheng et al., 2025). Backward chaining, or abductive reasoning based on tree Φ_{t-1} and premises set Σ_t, can provide additional coverage for searching over $\mathcal{K}$ (Sprague et al., 2022; Tafjord et al., 2022).

Function	Inputs	Prompt
Planning (Premise abduction)	k: hypothesis	Given the following hypothesis, try to generate a set of premises that can prove the hypothesis. Please format the premises as {{"Premises": ["PREMISE 1 TEXT", "PREMISE 2 TEXT", ...]}}. Hypothesis: The earth revolving around the sun causes leo to appear in different areas in the sky at different times of year. {{"Premises": ["Leo is a kind of constellation.", "A constellation contains stars.", "The earth revolving around the sun causes stars to appear in different areas in the sky at different times of year."]}} Hypothesis: The earth rotating on its axis causes stars to move relative to the horizon during the night. {{"Premises": ["Apparent motion is when an object appears to move relative to another object's position.", "The earth rotating on its axis causes stars to appear to move across the sky at night.", "Earth is a kind of celestial object.", "A star is a kind of celestial object / celestial body.", "Stars appear to move relative to the horizon during the night."]}} Hypothesis: {k}
Planning (Atomic fact decomposition)	s: sentence	Example 0: Please breakdown the following sentence into independent facts: He made his acting debut in the film The Moon is the Sun's Dream (1992), and continued to appear in small and supporting roles throughout the 1990s. {{"facts": ["He made his acting debut in the film.", "He made his acting debut in The Moon is the Sun's Dream.", "The Moon is the Sun's Dream is a film.", "The Moon is the Sun's Dream was released in 1992.", "After his acting debut, he appeared in small and supporting roles.", "After his acting debut, he appeared in small and supporting roles throughout the 1990s."]}} Example 1: Please breakdown the following sentence into independent facts: He is also a successful producer and engineer, having worked with a wide variety of artists, including Willie Nelson, Tim McGraw, and Taylor Swift. {{"facts": ["He is successful.", "He is a producer.", "He is a engineer.", "He has worked with a wide variety of artists.", "Willie Nelson is an artist.", "He has worked with Willie Nelson.", "Tim McGraw is an artist.", "He has worked with Tim McGraw.", "Taylor Swift is an artist.", "He has worked with Taylor Swift."]}} Example 2: Please breakdown the following sentence into independent facts: In 1963, Collins became one of the third group of astronauts selected by NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission. {{"facts": ["Collins became an astronaut.", "Collins became one of the third group of astronauts.", "Collins became one of the third group of astronauts selected.", "Collins became one of the third group of astronauts selected by NASA.", "Collins became one of the third group of astronauts selected by NASA in 1963.", "He served as the Command Module Pilot.", "He served as the back-up Command Module Pilot.", "He served as the Command Module Pilot for the Gemini 7 mission."]}} Example 3: Please breakdown the following sentence into independent facts: {s}
Planning (Proposition decomposition)	s: sentence	Given the following sentence, tell me what claims they are making. Please split the sentence as much as possible, but do not include information not in the sentence. Sentence: The Andy Warhol Museum in his hometown, Pittsburgh, Pennsylvania, contains an extensive permanent collection of art. {{"Claims": ["The Andy Warhol Museum is in Pittsburgh.", "Andy Warhol's hometown is in Pittsburgh.", "Pittsburgh is in Pennsylvania.", "The Andy Warhol Museum contains an extensive permanent collection of art."]}} Sentence: {s}
Planning (Query expansion)	k: claim	Please write a passage to support/refute the claim. Claim: k Passage (in the format "{{"passage": "PASSAGE TEXT"}}"):
Planning (with history)	k: hypothesis q: previous queries s: previous search results	You are an AI information retrieval specialist trained to optimize search queries for finding relevant evidence in factual sources. Task: Generate targeted search queries to find evidence that could either support or disprove the given hypothesis. Requirements: 1. Generate 3-5 refined search queries 2. Each query should be specific and focused 3. Consider both supporting and contradicting evidence 4. You may retain effective queries from the previous round Input Hypothesis: k Previous Information: - Previous queries: q - Previous search results: s Output Format: {{"queries": ["QUERY TEXT 1", "QUERY TEXT 2", ...]}}

Table 11: Prompts for LLMs: Planning

	EntailmentBank			WiCE			HotpotQA			MuSiQue		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Basic	63.6	97.2	76.9	88.1	66.0	75.4	83.5	83.8	83.7	82.9	49.8	62.2
Structured	77.2	86.0	81.4	96.4	47.4	63.5	90.0	79.6	84.5	88.9	44.1	59.0
CoT	73.1	88.5	80.1	90.6	52.5	66.4	85.1	79.9	82.4	86.0	50.2	63.4

Table 12: Comparison of three different prompting schemes in groundedness verification.

	EntailmentBank				**WiCE**				**HotpotQA**				**MuSiQue**			
	Info.	Redun.	Inc.	Uninfo.	Info.	Redun.	Inc.	Uninfo.	Info.	Redun.	Inc.	Uninfo.	Info.	Redun.	Inc.	Uninfo.
Basic	97.4	97.1	34.4	54.4	68.4	63.5	85.3	96.8	82.2	85.4	74.8	92.2	48	51.6	84.2	95.2
Structured	88.2	83.8	62.9	86.2	49.8	44.9	96.8	99.6	79.4	79.8	87.4	95	44.6	43.6	90.8	98.2
CoT	88.8	88.2	57.4	77.4	53	51.9	90.9	98.2	78.4	81.4	79.6	92.4	51.4	49	88	95.6

Table 13: Per-type performance in the comparison of three different prompting schemes in groundedness verification.