FactReasoner: A Probabilistic Approach to Long-Form Factuality Assessment for Large Language Models

Anonymous ACL submission

Abstract

Large langauge models (LLMs) have demonstrated vast capabilities on generative tasks in recent years, yet they struggle with guaranteeing the factual correctness of the generated content. This makes these models unreliable in realistic situations where factually accurate responses are expected. In this paper, we propose FactReasoner, a new factuality assessor that relies on probabilistic reasoning to assess the factuality of a long-form generated response. Specifically, FactReasoner decomposes the response into atomic units, retrieves relevant contexts for them from an external knowledge source, and constructs a joint probability distribution over the atoms and contexts using probabilistic encodings of the logical relationships (entailment, contradiction) between the textual utterances corresponding to the atoms and contexts. FactReasoner then computes the posterior probability of whether atomic units in the response are supported by the retrieved contexts. Our experiments on labeled and unlabeled benchmark datasets demonstrate clearly that FactReasoner improves considerably over state-of-the-art prompt-based approaches in terms of both factual precision and recall.

1 Introduction

Large language models (LLMs) have achieved impressive improvements and demonstrated vast capabilities in recent years (Brown et al., 2020; Chowdhery et al., 2023), however they still struggle to guarantee the factual accuracy of the generated content. Specifically, LLMs often *hallucinate*, namely they produce factual errors in which a claim contradicts well-established ground-truth knowledge (Zhang et al., 2023; Sahoo et al., 2024; Huang et al., 2025). This makes the models unreliable in realistic situations that require factually accurate LLM-generated responses (Tonmoy et al., 2024).

Most modern approaches for assessing the factuality of LLM-generated long-form responses such as FactScore (Min et al., 2023), VeriScore (Song et al., 2024) and others (Wei et al., 2024; Bayat et al., 2025) are prompt-based approaches and consist of three main stages: 1) the response is decomposed into a set of atomic units (facts or claims) which are subsequently revised or decontextualized to make them self-contained; 2) relevant evidence (or context) is retrieved for each atomic unit from an external knowledge source such as Wikipedia, and 3) each atomic unit is evaluated against the retrieved context to determine whether it is supported (factually correct) or not and a factuality score is calculated for the response. These approaches sometimes struggle due to conflicting information between the model's internal knowledge and conflicting information within the retrieved contexts themselves. Therefore, they typically assume that the pieces of information retrieved do not conflict or overlap with each other (Min et al., 2023).

Contributions: In this paper, we provide a new perspective on long-form factuality assessment that departs from the prompt-based approach, especially in the evaluation stage of the assessment. Specifically, we propose a novel factuality assessor called FactReasoner that also decomposes the response into atomic units and retrieves the relevant contexts for them from an external knowledge source. However, instead of prompting another LLM to evaluate the atoms against the retrieved evidence, FactReasoner computes the probability of each atom being supported by reasoning over a graphical model that represents a joint probability distribution over the atoms and the retrieved contexts. The graphical model is constructed using probabilistic encodings of the entailment and contradiction relationships between the natural language utterances corresponding to the atoms and contexts. Furthermore, FactReasoner makes no assumptions regarding the existence of any conflicting information within the retrieved contexts.

We conduct an extensive empirical evaluation on labeled and unlabeled benchmark datasets for long-form factuality and compare against several state-of-the-art prompt-based approaches using open-source LLMs. Our results demonstrate clearly that FactReasoner improves significantly over its competitors in terms of factual precision and recall. We show that exploiting the logical relationships between atoms and all retrieved contexts, as well as between the contexts themselves, allows FactReasoner to identify correctly considerably more supported atoms than the competing approaches.

The Appendix contains additional examples, experimental results and implementation details.

2 Background

In this section, we provide preliminaries on probabilistic graphical models and long-form factuality assessment for large language models.

Graphical Models. A *graphical model* is a tuple $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F} \rangle$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a set of variables, $\mathbf{D} = \{D_1, \ldots, D_n\}$ is the set of their finite domains of values and $\mathbf{F} = \{f_1, \ldots, f_m\}$ is a set of discrete positive real-valued functions. Each function f_i (also called *factor*) is defined on a subset of variables $\mathbf{S}_i \subseteq \mathbf{X}$ called its *scope* and denoted by $vars(f_i)$. The model $\mathcal{M}$ defines a factorized probability distribution on $\mathbf{X}$: $P(\mathbf{x}) = \frac{1}{Z} \prod_{j=1}^{m} f_j(\mathbf{x})$ where $Z = \sum_{\mathbf{x} \in \Omega(\mathbf{X})} \prod_{j=1}^{m} f_j(\mathbf{x})$ is the normalization constant Z is known as the *partition function* and $\Omega(\mathbf{X})$ denotes the Cartesian product of the variables domains (Koller and Friedman, 2009).

A common inference task over graphical models is to compute the posterior marginal distributions over all variables. Namely, for each variable $X_i \in \mathbf{X}$ and domain value $x_i \in D_i$, compute: $P(x_i) = \sum_{\mathbf{x} \in \Omega(\mathbf{X})} \delta_{x_i}(\mathbf{x}) \cdot P(\mathbf{x})$, where $\delta_{x_i}(\mathbf{x})$ is 1 if X_i is assigned x_i in $\mathbf{x}$ and 0 otherwise.

Long-Form Factuality. Let y be the long-form response generated by an LLM to a query x. Following prior work (Min et al., 2023; Song et al., 2024; Wei et al., 2024), we assume that y can be decomposed into a set of n *atomic units* (or *atoms*) that can be either true or false, denoted by $\mathcal{A}_y = \{a_1, a_2, \ldots a_n\}$. An atomic unit $a_i \in \mathcal{A}_y$ is defined as a short sentence conveying one piece of information. Furthermore, given an external knowledge source $\mathcal{C}$[1], we say that an atomic unit $a_i \in \mathcal{A}_y$ is *supported* by $\mathcal{C}$ if there exists at least one piece of information in $\mathcal{C}$ (e.g., a passage) called a *context* that undebatably supports a_i. Otherwise, we say that the atomic unit is *not supported*. The *factual precision* $Pr(y)$ of the response y with respect to a knowledge source $\mathcal{C}$ is defined as: $Pr(y) = \frac{S(y)}{|\mathcal{A}_y|}$, where $S(y) = \sum_{i=1}^{n} \mathbb{I}[a_i \text{ is supported by } \mathcal{C}]$ is the number of supported atomic units. Furthermore, the notion of *factual recall* up to the K-th supported atomic unit denoted by $R_K(y)$ is given by: $R_K(y) = \min(\frac{S(y)}{K}, 1)$. Finally, an F_1 measure for long-form factuality denoted by $F1@K$ can be defined as: $F_1@K(y) = \frac{2 \cdot Pr(y) \cdot R_K(y)}{Pr(y) + R_K(y)}$ if $S(y) > 0$, and 0 otherwise (Wei et al., 2024).

3 The FactReasoner Assessor

In this section, we present FactReasoner, a novel long-form factuality assessor that uses probabilistic reasoning to assess the factuality of the generated response with respect to an external knowledge source $\mathcal{C}$. Specifically, FactReasoner builds a graphical model that represents a joint probability distribution over the atoms of the response and their relevant contexts in $\mathcal{C}$, and subsequently computes for each atom a_i the posterior marginal probability distribution $P(a_i)$ representing the probability of a_i being true (or supported) with respect to the information available in $\mathcal{C}$.

3.1 A Graphical Models Based Approach

Let y be the long-form response generated by an LLM for the input query x, and let $\mathcal{A}_y = \{a_1, \ldots, a_n\}$ be the set of n atomic units corresponding to y. For simplicity, but without loss of generality, we restrict ourselves to atomic units that are either *facts* or *claims* (Song et al., 2024). In addition, let $\mathcal{C}_y = \{c_1, \ldots, c_m\}$ be a set of m contexts relevant to y's atoms that were retrieved from an external knowledge source $\mathcal{C}$. We make no assumptions about these contexts, namely they may be overlapping and/or contradicting each other, which is often the case in realistic situations.

We next define the graphical model $\langle \mathbf{X}, \mathbf{D}, \mathbf{F} \rangle$ that represents a joint probability distribution over the atoms and their corresponding contexts.

Variables. We associate each atom $a_i \in \mathcal{A}_y$ and context $c_j \in \mathcal{C}_y$ with a bi-valued variable de-

[1] For example, $\mathcal{C}$ could be Wikipedia, Google Search, or a collection of documents embedded into a vector database.

noted by either A_i (for atoms) or C_j (for contexts). Therefore, we have that $\mathbf{X} = \mathbf{X}_a \cup \mathbf{X}_c$ where $\mathbf{X}_a = \{A_1, \dots, A_n\}$ and $\mathbf{X}_c = \{C_1, \dots, C_m\}$, respectively. The domains of the variables contain the values *true* and *false* indicating whether the corresponding atom or context is true or false. For simplicity, we use a_i and $\neg a_i$ (resp. c_j and $\neg c_j$) to denote the value assignments $A_i = true$ and $A_i = false$ (resp. $C_j = true$ and $C_j = false$).

Priors. For each variable $A_i \in \mathbf{X}_a$ (resp. $C_j \in \mathbf{X}_c$) we consider a unary factor denoted by $f(A_i)$ (resp. $f(C_j)$) representing the prior belief about the truthfulness of the corresponding atom (resp. context). Since we make no assumptions about the response, we set $f(a_i) = 0.5$ and $f(\neg a_i) = 0.5$, respectively. In contrast, the external knowledge source $\mathcal{C}$ is assumed to be reliable and therefore the retrieved contexts have high probability of being true (e.g., $f(c_j) = 0.99$). Note that if a context is retrieved from a less reliable source then its prior probability can be set to a smaller value.

Relationships. In addition, we also consider binary factors denoted by $f(A_i, C_j)$ and $f(C_j, C_k)$, defined on atom-context variable pairs as well as pairs of context variables. These factors are probabilistic representations of the logical relationships between the natural language utterances corresponding to the context and atom variables. For our purpose, we use a *relation model* $p_\theta(\cdot|t, t')$ to predict the most likely logical relationship between an ordered pair of natural language utterances from the choices {none, entail, contradict, equivalence}[2]. The relation model can be any pre-trained BERT or LLM (Liu et al., 2019; Touvron et al., 2023).

Specifically, let X and Y be two variables in $\mathbf{X}$ and let t_X and t_Y be their corresponding textual utterances. Let also $r^* = \text{argmax}_r\, p_\theta(r|t_X, t_Y)$ be the predicted relationship between the ordered pair (t_X, t_Y) and let p^* be its probability. Table 1 shows the binary factor $f(X, Y)$ corresponding to $r^* \in \{\text{entailment, contradiction, equivalence}\}$.

For instance, if r^* corresponds to entailment and (X, Y) is a context-atom pair then the context supports the atom. Alternatively, if r^* is a contradiction for the same (X, Y) pair then the context contradicts the atom. Finally, for BERT-based relation models, the probability p^* is given together with the predicted relationship r^*, whereas for instructed LLM-based relation models we can obtain p^* by applying any uncertainty quantification method (Lin et al., 2024; Gao et al., 2024). We use a simple white-box method that calculates p^* as the probability of the “entailment" or “contradiction" tokens produced by the model.

		entailment	contradiction	equivalence
X	Y	$f(X,Y)$	$f(X,Y)$	$f(X,Y)$
x	y	p^*	$1-p^*$	p^*
x	$\neg y$	$1-p^*$	p^*	$1-p^*$
$\neg x$	y	p^*	p^*	$1-p^*$
$\neg x$	$\neg y$	p^*	p^*	p^*

Table 1: Factors corresponding to logical relationships.

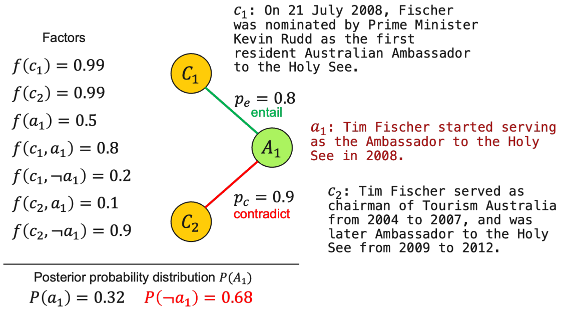

c_1: On 21 July 2008, Fischer was nominated by Prime Minister Kevin Rudd as the first resident Australian Ambassador to the Holy See.
Factors
$f(c_1) = 0.99$
$f(c_2) = 0.99$
$f(a_1) = 0.5$
$f(c_1, a_1) = 0.8$
$f(c_1, \neg a_1) = 0.2$
$f(c_2, a_1) = 0.1$
$f(c_2, \neg a_1) = 0.9$
C_1
$p_e = 0.8$
entail
A_1
a_1: Tim Fischer started serving as the Ambassador to the Holy See in 2008.
$p_c = 0.9$
contradict
C_2
c_2: Tim Fischer served as chairman of Tourism Australia from 2004 to 2007, and was later Ambassador to the Holy See from 2009 to 2012.
Posterior probability distribution $P(A_1)$
$P(a_1) = 0.32$
$P(\neg a_1) = 0.68$

Figure 1: FactReasoner: the graphical model corresponding to one atom A_1 and two contexts C_1 and C_2 such that C_1 entails A_1 and C_2 contradicts A_1.

Therefore, the set of factors $\mathbf{F}$ is:

$$
\begin{aligned}
\mathbf{F} = &\{f(C_j, A_i) \mid A_i \in \mathbf{X}_a, C_j \in \mathbf{X}_c\} \\
&\cup \{f(C_j, C_k) \mid C_j \in \mathbf{X}_c, C_k \in \mathbf{X}_c\} \\
&\cup \{f(A_i \mid \forall A_i \in \mathbf{X}_a)\} \\
&\cup \{f(C_j \mid \forall C_j \in \mathbf{X}_c)\}
\end{aligned}
$$

where we consider $r^* \in$ {entail, contradict} for the context-atom pairs, and $r^* \in$ {entail, contradict, equivalence} for the context pairs, respectively.

Example 1. *Figure 1 shows a simple example with one atomic unit a_1 and two contexts c_1 and c_2 retrieved from Wikipedia together with their corresponding natural language utterances. In this case, context c_1 entails the atom with probability $p_e = 0.8$ while context c_2 contradicts it with probability $p_c = 0.9$. The corresponding graphical model has 3 variables $\{A_1, C_1, C_2\}$, 3 unary factors $\{f(A_1),\ f(C_1),\ f(C_2)\}$ as well as 2 binary factors $\{f(C_1, A_1),\ f(C_2, A_2)\}$ encoding the two entailment and contradiction relationships.*

[2] The “equivalence" relationship is formed if entailment is predicted for both orderings of the utterances. The “none" relationship corresponds to neutrality meaning that the two utterances are not related to each other.

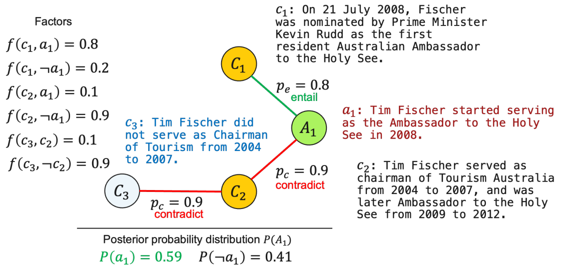

Factors
$f(c_1, a_1) = 0.8$
$f(c_1, \neg a_1) = 0.2$
$f(c_2, a_1) = 0.1$
$f(c_2, \neg a_1) = 0.9$
$f(c_3, c_2) = 0.1$
$f(c_3, \neg c_2) = 0.9$
c_1: On 21 July 2008, Fischer was nominated by Prime Minister Kevin Rudd as the first resident Australian Ambassador to the Holy See.
C_1
$p_e = 0.8$
entail
A_1
a_1: Tim Fischer started serving as the Ambassador to the Holy See in 2008.
c_3: Tim Fischer did not serve as Chairman of Tourism from 2004 to 2007.
C_3
$p_c = 0.9$
contradict
C_2
$p_c = 0.9$
contradict
c_2: Tim Fischer served as chairman of Tourism Australia from 2004 to 2007, and was later Ambassador to the Holy See from 2009 to 2012.
Posterior probability distribution $P(A_1)$
$P(a_1) = 0.59$ $P(\neg a_1) = 0.41$

Figure 2: FactReasoner: the graphical model corresponding to one atom A_1 and three contexts C_1, C_2 and C_3 such that C_3 contradicts C_2.

3.2 Inference and Factuality Assessment

The graphical model $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F} \rangle$ we just defined in the previous section represents a joint probability distribution over the set of atoms and relevant externally retrieved contexts. Therefore, we can use any probabilistic inference algorithm to compute the posterior marginal distribution $P(A_i)$ for each atom $A_i \in \mathcal{A}_y$ (Pearl, 1988; Koller and Friedman, 2009). Specifically, in our experiments, we use an approximate variational inference algorithm called Weighted Mini-Buckets (Liu and Ihler, 2011) to compute the marginals.

The number of supported atomic units $S(y)$ in a response y can be computed in this case as: $S(y) = \sum_{i=1}^{n} \mathbb{I}[P(a_i) > P(\neg a_i)]$, namely it is the number of atoms for which the probability of being true is larger than the probability of being false.

Example 2. *Looking again at Figure 1, we can see that in this case the posterior probability of the atom is $P(a_1) = 0.32$ and $P(\neg a_1) = 0.68$, which means that the atom is most likely false. Figure 2 continues the example and shows a third context c_3, possibly retrieved from another external knowledge source, that contradicts context c_2 and is neutral to atom a_1. As expected, the contradiction between c_2 and a_1 is much weaker now and therefore the posterior marginal probabilities are $P(a_1) = 0.59$ and $P(\neg a_1) = 0.41$, meaning that in light of the newly retrieved information, atom a_1 in more likely to be true than false. This example illustrates the kinds of conflicts that may exist between atoms and contexts and how they affect the factuality assessment.*

In addition to the factual precision $Pr(y)$ and $F_1@K$ measures, we define a new *entropy*-based factuality measure called $\mathcal{E}(y)$ that leverages the posterior probabilities of response y's atoms:

$$\mathcal{E} = \frac{1}{n} \sum_{i=1}^{n} -P(a_i) \cdot \log P(a_i) \tag{1}$$

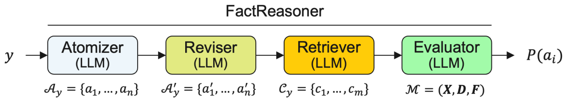

FactReasoner
y
Atomizer (LLM)
Reviser (LLM)
Retriever (LLM)
Evaluator (LLM)
$P(a_i)$
$\mathcal{A}_y = \{a_1, \ldots, a_n\}$
$\mathcal{A}'_y = \{a'_1, \ldots, a'_n\}$
$\mathcal{C}_y = \{c_1, \ldots, c_m\}$
$\mathcal{M} = (\boldsymbol{X}, \boldsymbol{D}, \boldsymbol{F})$

Figure 3: The FactReasoner pipeline.

where n is the number of atomic units in y.

Clearly, if all atoms in $\mathcal{A}_y$ have posterior probability $P(a_i) = 0.5$, there is virtually no external information to support or contradict the atoms (we refer to these atoms as *undecided atoms*) then $\mathcal{E}(y) = 0.150515$. On the other hand, if all atoms are true with absolute certainty ($P(a_i) = 1$), then $\mathcal{E}(y) = 0$ and if all atoms are false with absolute certainty then $\mathcal{E}(y) = \infty$. Therefore, when $\mathcal{E}(y)$ is closer to 0 the response is more truthful.

3.3 The FactReasoner Pipeline and Variants

The proposed FactReasoner pipeline for long-form factuality assessment is shown in Figure 3 and consists of four main stages called Atomizer, Reviser, Retriever and Evaluator, respectively. It takes as input a response y and outputs the marginal posterior probabilities $P(a_i)$ of y's atomic units together with the factuality measures described earlier, such as $Pr(y)$, $F_1@K(y)$ and $\mathcal{E}(y)$, respectively.

The **Atomizer** prompts an LLM to decompose the response y into a set of n atomic units $\mathcal{A}_y$ by applying any of the decomposition strategies proposed recently (Min et al., 2023; Bayat et al., 2025). Subsequently, the **Reviser** also uses an LLM to revise the atoms such that the pronouns, unknown entities, or incomplete names are replaced with their corresponding named entities in the response (Wei et al., 2024). Next, the **Retriever** is responsible for querying an external knowledge source to retrieve the contexts relevant to the response's atoms. At this stage, we can simply use the atoms' utterances as queries or prompt an LLM to generate them (Song et al., 2024). Finally, the **Evaluator** constructs the probabilistic graphical model representing the logical relationships between the atoms and contexts, and assess y's factuality via probabilistic reasoning, as described previously.

Depending on what relationships between atoms and contexts are considered, we define three versions of the FactReasoner pipeline, as follows:

FactReasoner 1 (FR1). In this case, for each atom variable A_i up to k most relevant contexts $\{C_1^i, \ldots, C_k^i\}$ are retrieved and only the relationships between atom A_i and its corresponding contexts are considered, namely only the factors

$f(A_i, C_j^i)$ are created (where $j = 1..k$).

FactReasoner 2 (FR2). This version also retrieves up to k contexts for each atom A_i, but it subsequently removes any duplicated contexts, thus resulting in m unique contexts denoted by $\{C_1, ..., C_m\}$. It then considers the relationships between atom A_i and all m contexts, creating the factors $f(A_i, C_j)$, where $j = 1..m$.

FactReasoner 3 (FR3). In this version, we consider the same contexts $\{C_1, ..., C_m\}$ as in FR2, but in addition to the atom-context relationships we also consider the context-context relationships. Thus, we create the factors $f(A_i, C_j)$ as well as the factors $f(C_j, C_k)$, where $j = 1..m$ and $k = 1..m$ and $j \neq k$, respectively.

4 Experiments

In this section, we empirically evaluate our proposed FactReasoner assessor for long-form factuality and compare it against state-of-the-art approaches on labeled and unlabeled datasets. Although the FactReasoner pipeline stages can be instantiated with different LLMs, in our implementation we use the same LLM throughout the entire pipeline and focus our empirical evaluation on the Evaluator stage (i.e., factuality assessment).

Baseline Assessors. For our purpose, we consider the following state-of-the-art prompt-based long-form factuality assessors: FactScore (FS) (Min et al., 2023), FactVerify (FV) (Bayat et al., 2025) and VeriScore (VS) (Song et al., 2024). FactScore is one of the first assessor that prompts an LLM to assess whether an atomic unit of the response is supported or not by a set of contexts relevant to the atom which are retrieved from an external knowledge source such as Wikipedia. FactVerify and VeriScore are more recent refinements of FactScore's original prompt that can accommodate other external knowledge sources such as Google Search results and enable the LLM's reasoning capabilities to evaluate the relationships between an atom and its relevant contexts. Unlike FactScore, the latter can label the atoms as supported, contradicted and undecided, respectively. In our experiments, we instantiated the competing assessors including the FactReasoner variants with open-source LLMs belonging to the IBM Granite[3], Meta LLama[4] and MistralAI Mixtral[5] families, namely: granite-3.0-8b-instruct, llama-3.1-70b-instruct, and mixtral-8x22b-instruct, respectively. All our LLMs are hosted remotely on compute nodes with A100 80GB GPUs and accessed via `litellm` APIs.

Datasets. We experimented with the following benchmark datasets: Biographies (Bio) (Min et al., 2023), AskHistorians (AskH) (Fangyuan Xu and Choi, 2023), ELI5 (Fangyuan Xu and Choi, 2023), FreshBooks (Books) (Song et al., 2024) and LongFact-Objects (LFObj) (Wei et al., 2024).

The Biographies is a *labeled* dataset that contains 157 biographies generated by ChatGPT for various person entities that have a Wikipedia page. Each biographic passage is also associated with a set of human generated atomic units (facts) that are labeled as *supported* (S) or *not-supported* (NS) by human annotators. We assume that this annotation is the ground truth.

The other four datasets are unlabeled and consist of collections of prompts (or questions). Specifically, the AskH and ELI5 datasets contain 200 questions each that were scraped from the reddit/AskHistorians and reddit/explainlikeimfive online forums, while the Books dataset consists of 10 paragraphs sampled from 20 non-fictional books that were published between 2023 and 2024, for a total of 200 paragraphs. Our version of the LFObj dataset is a subset of the original dataset (Wei et al., 2024) and contains 10 prompts sampled randomly from the original ones about objects spreading 38 different topics. For each prompt in these datasets, we generated a long-form response spanning up to two paragraphs using the llama-3.3-70b-instruct model (Touvron et al., 2023).

Measures of Performance. For each dataset $\mathcal{D}$ and each competing assessor, we report the factual precision Pr and $F_1@K$ measure, averaged over the number of prompts in $\mathcal{D}$. If $\mathcal{D}$ contains annotated atomic units (i.e., ground truth) then we also report the standard F_1 measure and the mean absolute error (MAE) given by:

$$\text{MAE} = \frac{1}{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} |Pr_j - Pr_j^*| \qquad (2)$$

where Pr_j and Pr_j^* are the precision and respectively the true factual precision of the j-th instance.

[3] `https://huggingface.co/ibm-granite`

[4] `https://huggingface.co/meta-llama`

[5] `https://huggingface.co/mistralai`

Dataset	# prompts	# atoms	# S*	Pr*	K
Biographies	157	31	20	0.62	32
AskH	200	22			22
Books	200	23			23
ELI5	200	22			21
LFObj	380	26			25

Table 2: Properties of the datasets used for evaluation.

Since the FactReasoner assessors calculate the posterior marginal distributions of the atoms, we also compute the $\mathcal{E}$-measure. Finally, we also report the mean number of supported (#S), contradicted (#C) and undecided atoms (#U), respectively.

External Knowledge Sources. We consider two external knowledge sources: Wikipedia and Google Search results. For a given atom, the top k results are retrieved as contexts either from wikipedia.org using the Wikipedia retriever available from LangChain[6], or from google.com using the Serper API[7]. In both cases, a context is a tuple (t, l, s, d), where t is the title of the wiki/web-page, l is the link, s is a short text snippet or summary and d is the content retrieved from l (but capped at max 4000 characters). We used $k = 3$ for the Wikipedia retriever and $k = 5$ for the Google Search results (Min et al., 2023; Wei et al., 2024).

In order to ensure a consistent evaluation, we decompose each response in the datasets into the corresponding atomic units (and subsequently revise them) using the same llama-3.3-70b-instruct model. Furthermore, we also retrieve and cache the relevant contexts for atoms from the two knowledge sources. This way, all competing assessors could be evaluated on the same sets of atoms and contexts. Table 2 summarizes the properties of the datasets, showing the number of prompts, the mean number of atoms and the median number of atoms (K). The latter is used for calculating the $F_1@K$ measure. In addition, for the labeled dataset, we also indicate the true number of supported atoms (S^*) and the true precision (Pr*).

4.1 Evaluating the Relation Model

We first evaluate the relation model used by the Evaluator stage of the FactReasoner assessor to extract the atom-context and context-context relationships required to construct the graphical model. Specifically, we consider two relation models based on a standard BERT-based model such as `vitc` (Schuster et al., 2021) and on a larger LLM such as `llama-3.1-70b-instruct` (Touvron et al., 2023) with a suitable few-shots prompt.

Assessor	# S	# C	# U	Pr↑	F_1 ↑	$F_1@K$ ↑	MAE↓	$\mathcal{E}$ ↓
BERT-based relation model: `albert-xlarge-vitaminc-mnli`								
FR1	12	5	12	0.40	0.66	0.39	0.25	0.11
FR2	10	16	4	0.32	0.53	0.31	0.34	0.09
FR3	10	16	4	0.32	0.53	0.31	0.33	0.09
LLM-based relation model: `llama-3.1-70b-instruct`								
FR1	13	1	16	0.41	0.70	0.41	0.23	0.10
FR2	**19**	2	9	**0.60**	**0.83**	**0.59**	**0.11**	**0.06**
FR3	**19**	2	9	**0.60**	**0.83**	**0.59**	**0.11**	**0.06**

Table 3: Results for the `vitc`- and `llama`-based relation models used by FactReasoner's Evaluator stage.

Assessor	# S	# C	# U	Pr ↑	F_1 ↑	$F_1@K$ ↑	MAE ↓	$\mathcal{E}$ ↓
`granite-3.0-8b-instruct`								
FS	18	12		0.59	0.70	0.57	0.17	
FV	14	2	14	0.45	0.67	0.44	0.21	
VS	15	8	6	0.49	0.64	0.48	0.21	
FR1 (ours)	14	2	14	0.43	0.70	0.43	0.22	0.12
FR2 (ours)	20	4	6	**0.62**	**0.78**	**0.61**	**0.12**	**0.06**
FR3 (ours)	19	4	6	0.60	0.78	0.59	0.13	**0.06**
`llama-3.1-70b-instruct`								
FS	19	12		0.59	0.73	0.58	0.16	
FV	15	1	14	0.47	0.73	0.47	0.19	
VS	12	0	18	0.38	0.64	0.38	0.27	
FR1 (ours)	13	1	16	0.42	0.71	0.42	0.23	0.10
FR2 (ours)	19	2	9	**0.60**	**0.83**	**0.59**	**0.11**	**0.06**
FR3 (ours)	19	2	9	**0.60**	**0.83**	**0.59**	**0.11**	**0.06**
`mixtral-8x22b-instruct`								
FS	19	12		0.59	0.74	0.58	0.16	
FV	15	1	13	0.49	0.72	0.48	0.19	
VS	13	1	15	0.42	0.65	0.42	0.25	
FR1 (ours)	14	0	15	0.44	0.72	0.44	0.21	0.10
FR2 (ours)	20	1	8	**0.63**	**0.83**	**0.62**	**0.11**	**0.07**
FR3 (ours)	20	1	9	**0.64**	**0.83**	**0.62**	**0.11**	**0.07**

Table 4: Results on the labeled Biographies dataset using Wikipedia contexts (mean number of supported (#S), contradicted (#C) and undecided (#U) atoms).

Table 3 shows the results obtained for the FR1, FR2 and FR3 assessors employing the two types of relation models on the Biographies dataset using Wikipedia retrieved contexts. We observe that using the LLM-based relation model which predicts entailments much more accurately than the BERT-based one leads to significant improvements in performance, especially for the FR2 and FR3 variants. For example, the `llama`-based FR2 achieves an F_1 score nearly twice as high compared with the `vitc`-based one (i.e., 0.83 versus 0.53). For this reason, we only employ LLM-based relation models for now on (see also the Appendix for more details).

4.2 Results on Labeled Datasets

Table 4 shows the results obtained on the labeled Biographies dataset using Wikipedia retrieved contexts (the best performance is highlighted). We see that in terms or mean absolute error (MAE),

[6] `https://python.langchain.com`

[7] `https://serper.dev`

precision and F_1 scores, the FR2 and FR3 assessors powered by stronger LLMs like llama-3.1-70b-instruct and mixtral-8x22b-instruct achieve the best performance compared to the other assessors. This is because both FR2 and FR3 can exploit the relationships between the atoms and all the retrieved contexts (as well as between the contexts themselves for FR3), not just the ones between an atom and its corresponding top k contexts. Therefore, it is often the case that a context retrieved for atom A_i may support or contradict another atom A_j for which it wasn't retrieved. This leads to a higher number of true positives and consequently larger F_1 scores. We also observe that the numbers of undecided atoms is also smaller for FR2/FR3 compared with the other assessors. The performance of FR3 is similar to that of FR2 because most of the context-context relationships are equivalence.

When looking at the prompt-based assessors, especially FV and VS, we see that they are more conservative in terms of number of supported atoms found. This can be explained by the relatively strict instructions specified in their prompts for identifying supported/contradicted atoms. Hence the number of undecided atoms is much larger than that of FR2/FR3. The simple prompt used by FS leads to finding a relatively large number supported atoms, across all the backend LLMs considered. However, many of these supported atoms are actually false positives which in fact is explained by the relatively smaller F_1 score compared with the best performing assessors FR2 and FR3, respectively.

We notice that the lightweight FR1 assessor performs on par with FV and VS in terms of precision, error and F_1 score. This shows that using only the top k contexts to determine whether an atom is supported or not is fairly limited. Furthermore, in situations when an atom is supported by several contexts but is contradicted by another context (which might as well be a spurious contradiction), the FR2/FR3 assessors are able to correctly label the atom as supported based on the strengths of the respective relationships (i.e., probabilities of entailment and contradiction) whereas the other assessors struggle and often label the atom as contradicted or undecided. This demonstrates clearly the power of the probabilistic approach to factuality as employed by the proposed assessors.

Assessor	# S	# C	# U	Pr ↑	$F_1@K$ ↑	$\mathcal{E}$ ↓
granite-3.0-8b-instruct						
FS	17	5		0.76	0.74	
FB	8	0	13	0.35	0.36	
FV	12	4	5	0.55	0.55	
FR1 (ours)	4	1	16	0.19	0.19	0.14
FR2 (ours)	10	9	2	0.46	0.47	0.09
FR3 (ours)	11	8	2	0.47	0.48	0.10
llama-3.1-70b-instruct						
FS	15	7		0.69	0.68	
FB	8	0	13	0.37	0.38	
FV	5	0	16	0.25	0.25	
FR1 (ours)	5	0	17	0.21	0.22	0.13
FR2 (ours)	10	1	10	0.45	0.46	0.09
FR3 (ours)	10	1	10	0.44	0.45	0.09
mixtral-8x22b-instruct						
FS	16	6		0.71	0.70	
FB	9	0	12	0.43	0.43	
FV	7	0	14	0.34	0.34	
FR1 (ours)	5	0	17	0.22	0.23	0.12
FR2 (ours)	11	0	11	0.46	0.47	0.09
FR3 (ours)	11	0	11	0.46	0.47	0.09
DeepSeek-v3	9	1	12	0.43	0.43	

Table 5: Results on the unlabeled AskH dataset using Wikipedia contexts (mean number of supported (#S), contradicted (#C) and undecided (#U) atoms).

Assessor	# S	# C	# U	Pr ↑	$F_1@K$ ↑	$\mathcal{E}$ ↓
granite-3.0-8b-instruct						
FS	18	3		0.82	0.81	
FV	14	1	7	0.62	0.62	
VS	14	3	3	0.65	0.65	
FR1 (ours)	13	4	4	0.60	0.60	0.08
FR2 (ours)	14	7	0	0.63	0.62	0.04
FR3 (ours)	15	7	0	0.67	0.66	0.06
llama-3.1-70b-instruct						
FS	18	3		0.82	0.80	
FV	16	1	5	0.71	0.70	
VS	15	0	7	0.66	0.65	
FR1 (ours)	12	1	8	0.53	0.54	0.08
FR2 (ours)	17	1	3	0.76	0.74	0.04
FR3 (ours)	17	2	3	0.75	0.74	0.04
mixtral-8x22b-instruct						
FS	18	3		0.82	0.80	
FV	15	0	6	0.67	0.67	
VS	15	0	6	0.68	0.67	
FR1 (ours)	14	0	8	0.60	0.60	0.07
FR2 (ours)	18	0	3	0.80	0.79	0.04
FR3 (ours)	18	0	3	0.80	0.79	0.04
DeepSeek-v3	15	2	5	0.69	0.69	

Table 6: Results on the unlabeled AskH dataset using Google Search contexts (mean number of supported (#S), contradicted (#C) and undecided (#U) atoms).

4.3 Results on Unlabeled Datasets

Tables 5 and 6 show the results obtained on the unlabeled AskH dataset using Wikipedia and Google Search retrieved contexts, respectively (we include in the the Appendix the experiments with the remaining datasets: Books, ELI5 and LFObj). Since there is no ground truth for this dataset, we only report the precision, $F_1@K$ (for $K = 22$) and the $\mathcal{E}$ measure. However, for reference, we also experimented with DeepSeek-v3, perhaps one of the strongest open models at the moment, using a suitable prompt (DeepSeek-AI, 2024).

We note that the AskH dataset covers a wider range of topics compared with the Biographies dataset and, therefore, the Wikipedia based contexts have a much smaller coverage in this case compared with the Google search results. This is reflected by the relatively smaller number of atoms supported by Wikipedia only contexts (Table 5) compared with those supported by Google Search results (Table 6), across all competing assessors. A similar pattern can be observed for the precision, $F_1@K$ and $\mathcal{E}$ measures reported in Table 5, as they are typically inferior to those shown in Table 6.

The prompt-based assessors FV and VS are fairly conservative in this case as well and find relatively fewer supported atoms compared with the FR2 and FR3 assessors. The latter two benefit from considering the relationships between an atom and all of the retrieved contexts and, therefore, find more supported atoms. The corresponding precision and $F_1@K$ values are also higher for FR2/FR3. We also observe that the $\mathcal{E}$-measure specific to the FR assessors correlates well with the number of supported atoms, namely as the number of supported atoms increases $\mathcal{E}$ gets closer to 0 .

When looking at the FS assessor, we notice again that it tends to find more supported atoms than the other assessors. However, we hypothesise that some of these atoms are false positive as before, but acknowledge that without any ground truth information it is difficult to verify this hypothesis.

Comparing the results obtained with DeepSeek-v3, we see that FV and VS come very close, especially for Google Search contexts. This is likely because the prompts used are fairly similar. In contrast, FR2/FR3 find slightly more supported atoms, although the results are very close for Wikipedia only contexts. We believe that this is caused by spurious context-atom entailment relationships which indicates that a better relation model is required.

In summary, our proposed FactReasoner assessor achieved the best performance on the labeled dataset, nearly matching the ground truth. However, on the unlabeled datasets, its performance was comparable with that of its competitors including DeepSeek-v3, a very powerful open model.

5 Related Work

The assessment of LLMs' adherence to factual knowledge has gained significant attention in recent years due to their widespread adoption. Several well-established benchmarks, including TruthfulQA (Lin et al., 2022), FreshQA (Vu et al., 2023), HaluEval (Li et al., 2023), HalluQA (Cheng et al., 2023), and FELM (Chen et al., 2023), focus on short-form response evaluation, where an LLM's knowledge is tested through individual factoids classified as either true or false. More recent studies (Min et al., 2023; Wei et al., 2024; Bayat et al., 2025; Song et al., 2024) have extended this approach to long-form generations by decomposing responses into distinct factual elements, which are then evaluated separately against relevant evidence retrieved from an external source of knowledge. These previous works typically assume that the retrieved pieces of information do not overlap or conflict with each other.

Conflicting information is prevalent in external knowledge sources (Xu et al., 2024) and it typically impacts modern retrieval augmented-generation systems that aim to reduce hallucinations in LLMs (Lewis et al., 2021). Other works have developed new benchmarks for capturing knowledge conflicts in realistic situations (Hou et al., 2024; Marjanović et al., 2024; Su et al., 2024; Pham et al., 2024).

Our work is closely related with recent studies on self-consistency that aim at improving the logical consistency of the LLM's response with respect to the input query by leveraging various methods including formal reasoning (Wang et al., 2023; Dohan et al., 2022; Mitchell et al., 2022).

6 Conclusion

The paper provides a new perspective on long-form factuality assessment and proposes FactReasoner, a new factuality assessor that employs probabilistic reasoning to assess the factuality of an LLM-generated long-form response. FactReasoner proceeds in a manner similar to existing prompt-based assessors by decomposing the response into atomic units and retrieving contexts relevant to them from an external knowledge source. However, unlike those methods, FactReasoner evaluates the factuality of the atoms by probabilistic reasoning over a graphical model that represents the logical relationships between the textual utterances corresponding to the atoms and contexts. We experiment with labeled and unlabeled benchmark datasets and demonstrate conclusively that FactReasoner improves significantly over the state-of-the-art prompt based approaches for long-form factuality evaluation. For future work, we plan to leverage the new FactReasoner assessor as part of a self-reflection loop to facilitate correction of the response.

Limitations

We acknowledge further limitations of the proposed FactReasoner approach.

First, the Atomizer stage is sensitive to the quality of the prompt and few shot examples used as well as the LLM employed to perform the atomic unit decomposition of the response. In our work we only consider open-source models from the LLaMA family (i.e., `llama-3.3-70b-instruct`). Furthermore, the decomposition of the response can be done at different granularities such as sentence level, paragraph level and the entire response level. Our implementation is limited to decomposing the entire response in one shot.

Second, the Reviser stage is also sensitive to how well the prompt is crafted as well as the quality of the few shot examples included in the prompt. Again, at this stage we only used the `llama-3.3-70b-instruct` model.

Third, the quality of the contexts retrieved for each atomic unit depends on the implementation of the retriever used as well as the structure of the query string that it receives. Our implementation is limited to off-the-shelf retrievers such as the one available from LangChain and we used the atomic unit's utterance as query. It is possible to prompt an LLM to generate better quality queries as suggested in previous work (Song et al., 2024). Therefore, employing a more advanced retriever will lead to better quality retrieved contexts and consequently will improve the overall performance of the proposed FactReasoner assessors.

Fourth, extracting the logical relationships between atoms and contexts as well as between the contexts themselves also depends on the quality of the prompt and the LLM. As before, for our relation model we only used open-source models such as granite-3.0-8b-instruct, llama-3.1-70b-instruct, and mixtral-8x22b-instruct with a fairly straightforward prompt. It is possible to craft better prompts that could lead to a better extraction of the relationships. Fine-tuning is another option to obtain a stronger relation model.

Finally, from a computational overhead perspective, the FR3 version requires $O(n \cdot m + m^2)$ prompts to extract the relationships between atoms and context, the FR2 version requires $O(n \cdot m)$ prompts while FR1 requires $O(k \cdot n)$ prompts, where n is the number of atomic units, m is the total number of non-duplicated contexts retrieved for the atoms, and k is maximum number of contexts retrieved per atom. In contrast, the prompt-based factuality assessor only require $O(n)$ prompts.

Ethical Statement

We recognize the positive and negative societal impacts of LLMs in general, including potential misuse of our work around uncertainty quantification for LLM generated output. We note that the datasets considered are public and peer reviewed, there are no human subjects involved, and as far as we know, there are no obvious harmful consequences from our work. All creators and original owners of assets have been properly credited and licenses and terms of use have been respected. We have not conducted crowd-sourcing experiments or research with human subjects.

References

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2025. Factbench: A dynamic benchmark for in-the-wild language model factuality evaluation. *Preprint*, arXiv:2410.22257.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *Preprint*, arXiv:2310.03368.

Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 4(1):1–113.

R. Dechter. 2003. *Constraint Processing*. Morgan Kaufmann Publishers.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *https://arxiv.org/html/2412.19437v1*.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. 2022. Language model cascades. *CoRR*, abs/2207.10342.

Mohit Iyyer Fangyuan Xu, Yixiao Song and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 2336–2346.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia. *Preprint*, arXiv:2406.13805.

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 Trans. Inf. Syst.*, 43(2).

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Q. Liu and A. Ihler. 2011. Bounding the partition function using Holder's inequality. In *International Conference on Machine Learning (ICML)*, pages 849–856.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, arXiv:1907.11692.

Sara Vera Marjanović, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. Dynamicqa: Tracing internal knowledge conflicts in language models. *arXiv preprint arXiv:2407.17023*.

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

Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1768, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Quang Hieu Pham, Hoang Ngo, Anh Tuan Luu, and Dat Quoc Nguyen. 2024. Who's who: Large language models meet knowledge conflicts in practice. *arXiv preprint arXiv:2410.15737*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *Preprint*, arXiv:2405.09589.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation. *Preprint*, arXiv:2310.03214.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

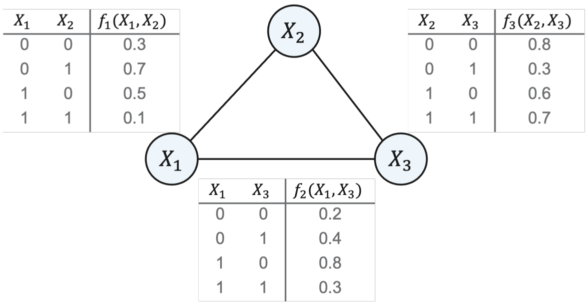
Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

A Details on Graphical Models

Graphical models such as Bayesian or Markov networks provide a powerful framework for reasoning about conditional dependency structures over many variables (Pearl, 1988; Koller and Friedman, 2009).



X_1	X_2	$f_1(X_1,X_2)$
0	0	0.3
0	1	0.7
1	0	0.5
1	1	0.1

X_2	X_3	$f_3(X_2,X_3)$
0	0	0.8
0	1	0.3
1	0	0.6
1	1	0.7

X_1	X_3	$f_2(X_1,X_3)$
0	0	0.2
0	1	0.4
1	0	0.8
1	1	0.3

Figure 4: A graphical model with three bi-valued variables X_1, X_2 and X_3, and three binary functions.

A *graphical model* is a tuple $\mathcal{M} = \langle \mathbf{X}, \mathbf{D}, \mathbf{F} \rangle$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a set of variables, $\mathbf{D} = \{D_1, \ldots, D_n\}$ is the set of their finite domains of values and $\mathbf{F} = \{f_1, \ldots, f_m\}$ is a set of discrete positive real-valued functions. Each function f_i (also called *factor*) is defined on a subset of variables $\mathbf{S}_i \subseteq \mathbf{X}$ called its *scope* and denoted by $vars(f_i)$. The model $\mathcal{M}$ defines a factorized probability distribution on $\mathbf{X}$:

$$P(\mathbf{x}) = \frac{1}{Z}\prod_{j=1}^{m} f_j(\mathbf{x}) \text{ s.t. } Z = \sum_{\mathbf{x}\in\Omega(\mathbf{X})}\prod_{j=1}^{m} f_j(\mathbf{x}) \tag{3}$$

where the normalization constant Z is known as the *partition function* and $\Omega(\mathbf{X})$ denotes the Cartesian product of the variables domains.

The function scopes of a model $\mathcal{M}$ define a *primal graph* whose vertices are the variables and its edges connect any two variables that appear in the scope of the same function.

A common inference task over graphical models is to compute the posterior marginal distributions over all variables. Namely, for each variable $X_i \in \mathbf{X}$ and domain value $x_i \in D_i$, compute:

$$P(x_i) = \sum_{\mathbf{x}\in\Omega(\mathbf{X})} \delta_{x_i}(\mathbf{x}) \cdot P(\mathbf{x}) \tag{4}$$

where $\delta_{x_i}(\mathbf{x})$ is 1 if X_i is assigned x_i in $\mathbf{x}$ and 0 otherwise (Koller and Friedman, 2009).

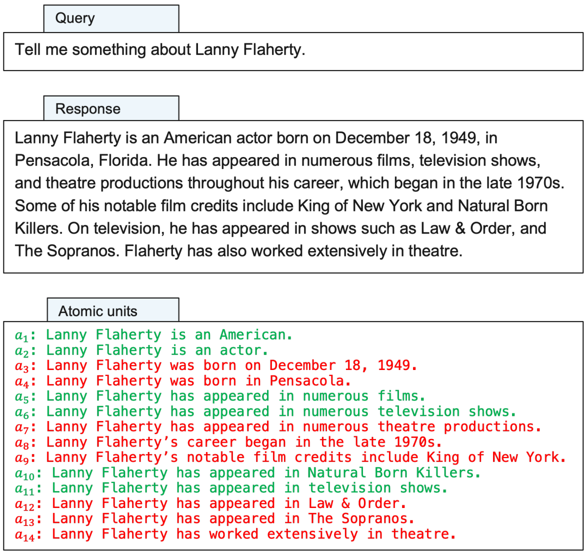
Example 3. *Figure 4 shows a graphical model with 3 bi-valued variables X_1, X_2 and X_3 and 3 binary functions $f_1(X_1, X_2)$, $f_2(X_1, X_3)$ and $f_3(X_2, X_3)$. The joint probability distribution is given by $P(X_1, X_2, X_3) = \frac{1}{Z} \cdot f_1(X_1, X_2) \cdot f_2(X_1, X_3) \cdot f_3(X_2, X_3)$. In this case, the posterior marginal distribution of X_1 is: $P(X_1 = 0) = 0.46$ and $P(X_1 = 1) = 0.54$, respectively.*

Solving Equation 4 can be done using any probabilistic inference algorithm for graphical models such as variable elimination (Dechter, 2003), belief propagation (Pearl, 1988) or variational inference (Liu and Ihler, 2011). In our implementation we used the Weighted Mini-Buckets (WMB) algorithm (Liu and Ihler, 2011). Specifically, WMB is parameterized by a so called i-bound which controls the complexity of inference. In our experiments we selected an i-bound of 6 which allowed us to solve all inference problems relatively efficiently.

B Details on Long-Form Factuality Assessment

Assessing the factuality of long form text generations is a challenging problem because these kinds of generations may contain a large number of informative statements and validating each piece of information against one or more reliable sources may be time-consuming, costly and often prone to errors (Min et al., 2023; Wei et al., 2024).

Formally, let y be the long form text generated by a large language model $\mathcal{L}$ in response to a query x. Following prior work (Min et al., 2023; Song et al., 2024), we assume that y consists of n *atomic units* (or *atoms*) that can be either true or false, denoted by $\mathcal{A}_y = \{a_1, a_2, \dots a_n\}$. An atomic unit $a_i \in \mathcal{A}_y$ is defined as a short sentence conveying one piece of information. Furthermore, given an external knowledge source $\mathcal{C}$[8], we say that an atomic unit $a_i \in \mathcal{A}_y$ is *supported* by $\mathcal{C}$ if there exists at least one piece of information in $\mathcal{C}$ (e.g., a passage) called a *context* that undebatably supports a_i. Otherwise, we say that the atomic unit is *not supported* (Min et al., 2023; Song et al., 2024).


Query

Tell me something about Lanny Flaherty.

Response

Lanny Flaherty is an American actor born on December 18, 1949, in Pensacola, Florida. He has appeared in numerous films, television shows, and theatre productions throughout his career, which began in the late 1970s. Some of his notable film credits include King of New York and Natural Born Killers. On television, he has appeared in shows such as Law & Order, and The Sopranos. Flaherty has also worked extensively in theatre.

Atomic units

a_1: Lanny Flaherty is an American.
a_2: Lanny Flaherty is an actor.
a_3: Lanny Flaherty was born on December 18, 1949.
a_4: Lanny Flaherty was born in Pensacola.
a_5: Lanny Flaherty has appeared in numerous films.
a_6: Lanny Flaherty has appeared in numerous television shows.
a_7: Lanny Flaherty has appeared in numerous theatre productions.
a_8: Lanny Flaherty’s career began in the late 1970s.
a_9: Lanny Flaherty’s notable film credits include King of New York.
a_{10}: Lanny Flaherty has appeared in Natural Born Killers.
a_{11}: Lanny Flaherty has appeared in television shows.
a_{12}: Lanny Flaherty has appeared in Law & Order.
a_{13}: Lanny Flaherty has appeared in The Sopranos.
a_{14}: Lanny Flaherty has worked extensively in theatre.

Figure 5: An example user prompt and the corresponding long form response together with its supported (green) and not supported (red) atomic units.

Therefore, the *factual precision* $Pr(y)$ of the response y with respect to a knowledge source $\mathcal{C}$ is defined as:

$$Pr(y) = \frac{S(y)}{|\mathcal{A}_y|} \tag{5}$$

where $S(y) = \sum_{i=1}^{n} \mathbb{I}[a_i \text{ is supported by } \mathcal{C}]$ is the number of supported atomic units. Similarly, the notion of *factual recall*[9] *up to the* K*-th supported atomic unit* denoted by $R_K(y)$ can be defined as follows:

$$R_K(y) = \min(\frac{S(y)}{K}, 1) \tag{6}$$

Combining Equations 5 and 6 yields an F_1 measure for factuality denoted $F1@K$ as follows:

$$F_1@K(y) = \begin{cases} \dfrac{2 \cdot Pr(y) \cdot R_K(y)}{Pr(y) + R_K(y)}, & S(y) > 0 \\ 0, & S(y) = 0 \end{cases} \tag{7}$$

Intuitively, $F_1@K(y)$ measures the long-form factuality of a model response y given the numbers of supported and not-supported atomic units in y. The parameter K indicates the number of supported atomic units required for a response to achieve full recall (Wei et al., 2024).

The precision and recall definitions however assume that the pieces of information in $\mathcal{C}$ do not conflict or overlap with each other (Min et al., 2023).

Example 4. *In Figure 5 we show an example of a long form generated text for a user prompt/query. In this case, the response y contains 14 atomic units $\mathcal{A}_y = \{a_1, a_2, \dots, a_{14}\}$. Furthermore, considering Wikipedia as our reliable knowledge source, we depict in green the supported atomic units, while the ones in red are not supported. The factual precision and $F_1@K$ of the response are $Pr(y) = 0.43$ and $F_1@K(y) = 0.57$ for $K = 7$, respectively.*

C Additional Experiments

Figure 6 plots the ROC curves for predicting contradiction and entailment relationships on the Expert

[8] For example, $\mathcal{C}$ could be Wikipedia, the Web, or a collection of documents embedded into a vector database.

[9] Measuring recall is quite challenging because it is almost impossible to come up with a definite set of atomic units that should be included in a long form response (Wei et al., 2024)

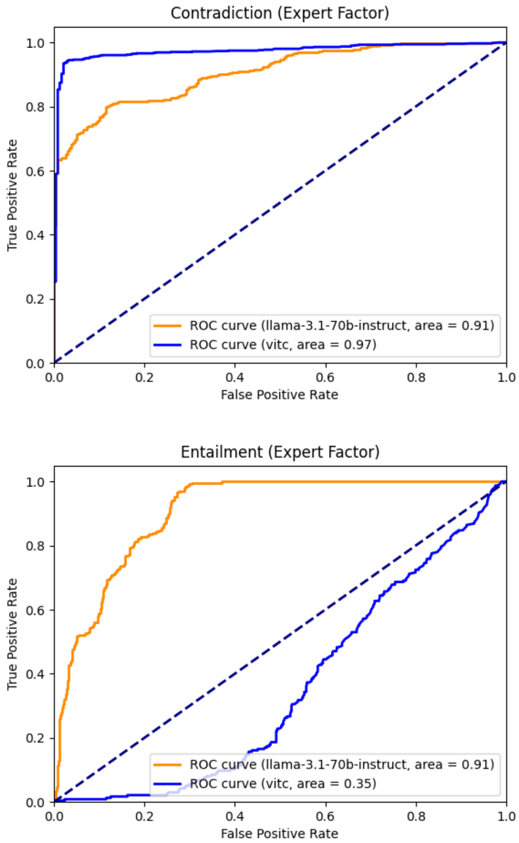


Figure 6: ROC curves for the `vitc`- and `llama`-based relation models predicting contradiction and entailment.

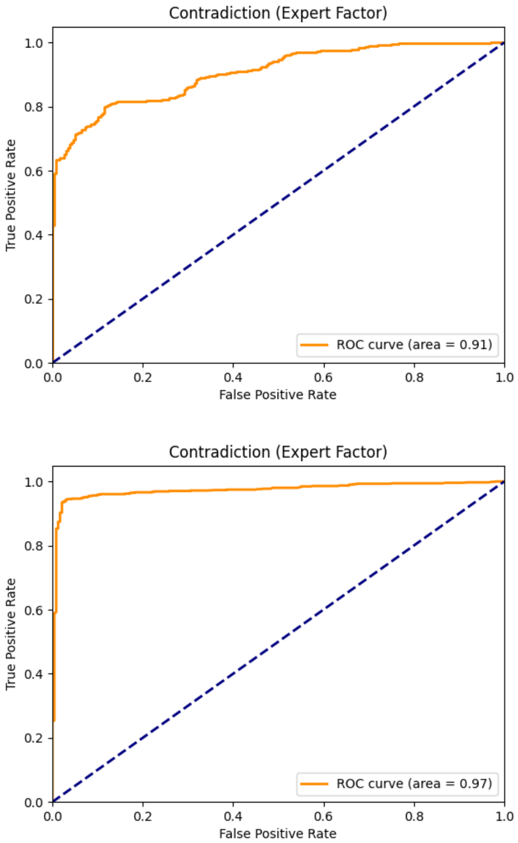


Figure 7: ROC curves for the `llama`- (top) `vitc`-based (bottom) relation models predicting contradiction on the Expert FACTOR dataset.

FACTOR dataset (Muhlgay et al., 2024). We see that the `vitc`-based model predicts contradictions fairly accurately compared with the `llama`-based one, but performs rather poorly on predicting the entailment relations.

Figures 7 and 8 plot the ROC curves for predicting the contradiction and entailment relationships by the `llama`- and `vitc`-based relation models on the Expert FACTOR dataset (Muhlgay et al., 2024). Figures 9 and 10 plot the ROC curves for predicting the contradiction and entailment relationships by the same relation models on the News FACTOR dataset (Muhlgay et al., 2024)

In Table 7 we show the results obtained on the same Biographies dataset but using Google Search results as contexts. We observe a similar pattern of the results compared with the previous case, namely FV and VS being more conservative than the FR assessors. However, we notice that in this case there are many more atoms labeled as supported (#S) and consequently more false positives which is reflected in the slightly higher MAE values for all competing assessors. We believe that this is most likely caused by the slightly noisier contexts compared with the Wikipedia only based ones which eventually leads to more spurious entailment relationships than in the previous case. As before, we note that the relatively simple prompt employed by FS leads to large numbers of atoms labeled as supported.

Tables 8 and 9 contain the detailed results obtained on the labeled Biographies dataset including the standard deviations for each of the reported performance measures.

Tables 10, 12, 14 and 16 report the detailed results obtained on the unlabeled datasets AskH, Books, ELI5 and LFObj using Wikipedia retrieved contexts. Tables 11, 13, 15 and 17 show the detailed results obtained on the unlabeled datasets Books, ELI5 and LFObj using Google Search results based contexts. All these additional results show a similar pattern to those reported for the AskH dataset in the main paper.

D Prompts

Tables 18, 19 and 20 show the prompt templates we used for the Atomizer, Reviser and Evaluator

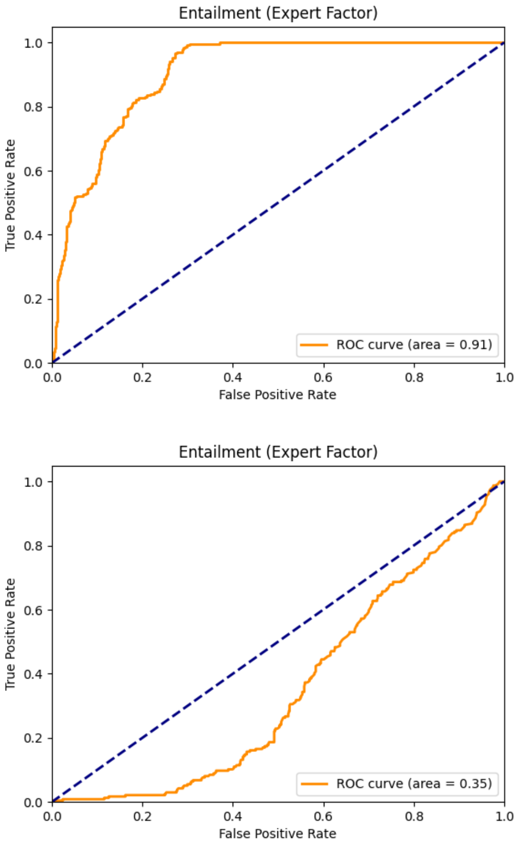


Figure 8: ROC curves for the `llama`- (top) `vitc`-based (bottom) relation models predicting entailment on the Expert FACTOR dataset.

Method	# S	# C	# U	Pr	F_1	$F_1@K$	MAE	$\mathcal{E}$
`granite-3.0-8b-instruct`								
FS	24	6		0.76	0.80	0.73	0.15	
FV	20	2	8	0.64	0.74	0.62	0.14	
VS	21	1	8	0.67	0.74	0.65	0.14	
FR1	23	3	4	0.73	0.79	0.70	0.14	0.08
FR2	24	5	0	0.78	0.80	0.75	0.19	0.04
FR3	24	5	0	0.78	0.79	0.74	0.18	0.04
`llama-3.1-70b-instruct`								
FS	23	7		0.73	0.82	0.71	0.14	
FV	23	3	4	0.72	0.82	0.70	0.13	
VS	23	1	6	0.72	0.81	0.70	0.13	
FR1	21	2	7	0.66	0.81	0.64	0.11	0.06
FR2	24	2	3	0.77	0.83	0.74	0.16	0.03
FR3	24	2	3	0.77	0.83	0.74	0.16	0.03
`mixtral-8x22b-instruct`								
FS	24	6		0.75	0.83	0.72	0.15	
FV	22	2	5	0.71	0.82	0.69	0.12	
VS	23	1	5	0.73	0.81	0.71	0.13	
FR1	22	1	6	0.71	0.81	0.69	0.13	0.05
FR2	25	1	3	0.81	0.82	0.77	0.19	0.03
FR3	25	2	3	0.80	0.82	0.77	0.19	0.03

Table 7: Results obtained on the labeled Biographies dataset using Google Search retrieved contexts.

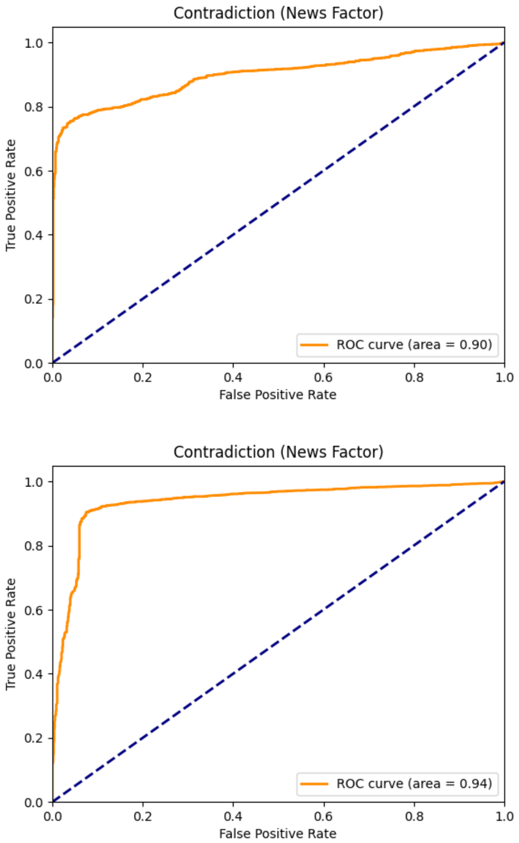


Figure 9: ROC curves for the `llama`- (top) `vitc`-based (bottom) relation models predicting contradiction on the News FACTOR dataset.

Assessor	# S	# C	# U	Pr	F_1	$F_1@K$	MAE	$\mathcal{E}$
`granite-3.0-8b-instruct`								
FS	18±8	12±5		0.59±0.17	0.70±0.17	0.57±0.20	0.17±0.14	
FV	14±7	2±1	14±6	0.45±0.19	0.67±0.15	0.44±0.21	0.21±0.14	
VS	15±8	8±4	6±3	0.49±0.20	0.64±0.19	0.48±0.22	0.21±0.14	
FR1	14±6	2±2	14±6	0.43±0.20	0.70±0.15	0.43±0.21	0.22±0.13	0.12±0.01
FR2	20±6	4±3	6±3	**0.62±0.21**	**0.78±0.15**	**0.61±0.23**	**0.12±0.13**	0.06±0.01
FR3	19±6	4±3	6±3	0.60±0.19	0.78±0.14	0.59±0.22	0.13±0.13	0.06±0.01
`llama-3.1-70b-instruct`								
FS	19±8	12±5		0.59±0.20	0.73±0.16	0.58±0.20	0.16±0.14	
FV	15±8	1±1	14±6	0.47±0.20	0.73±0.15	0.47±0.22	0.19±0.12	
VS	12±8	0	18±7	0.38±0.21	0.64±0.18	0.38±0.23	0.27±0.15	
FR1	13±8	1±2	16±6	0.42±0.20	0.71±0.15	0.42±0.21	0.23±0.13	0.10±0.02
FR2	19±9	2±2	9±5	**0.60±0.20**	**0.83±0.13**	**0.59±0.24**	**0.11±0.11**	**0.06±0.02**
FR3	19±9	2±2	9±5	**0.60±0.20**	**0.83±0.14**	**0.59±0.24**	**0.11±0.12**	**0.06±0.02**
`mixtral-8x22b-instruct`								
FS	19±8	12±5		0.59±0.18	0.74±0.16	0.58±0.20	0.16±0.13	
FV	15±7	1±1	13±5	0.49±0.18	0.72±0.14	0.48±0.21	0.19±0.12	
VS	13±7	1±1	15±6	0.42±0.18	0.65±0.16	0.42±0.20	0.25±0.14	
FR1	14±8	0±1	15±6	0.44±0.20	0.72±0.15	0.44±0.22	0.21±0.13	0.10±0.02
FR2	20±9	1±1	8±5	**0.63±0.20**	**0.83±0.14**	**0.62±0.24**	**0.11±0.11**	**0.07±0.01**
FR3	20±9	1±1	9±5	**0.64±0.21**	**0.83±0.14**	**0.62±0.24**	**0.11±0.12**	**0.07±0.01**

Table 8: Results obtained on the labeled Biographies dataset using Wikipedia retrieved contexts.

stages of the FactReasoner pipeline. Tables 21, 22 and 23 show the prompts used by the prompt-based assessors: FactScore (FS), FactVerify (FV) and VeriScore (VS), respectively.

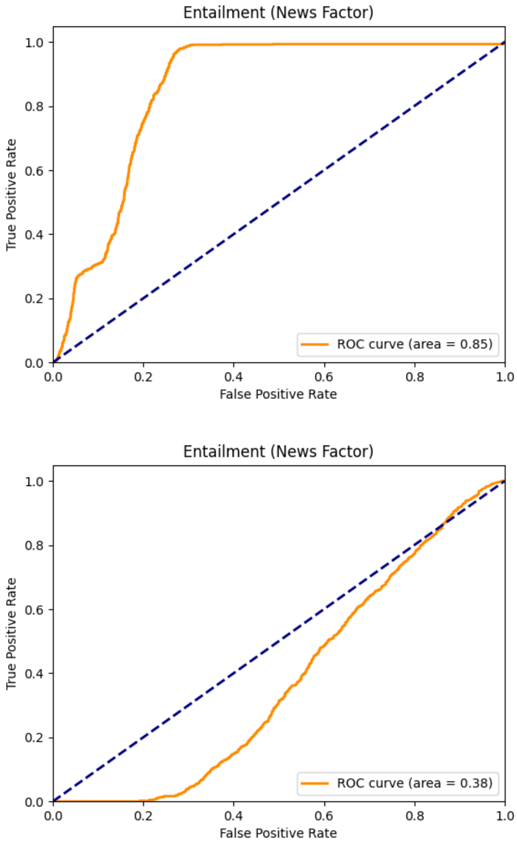


Figure 10: ROC curves for the `llama`- (top) `vitc`-based (bottom) relation models predicting entailment on the News FACTOR dataset.

Method	# S	# C	# U	Pr	F_1	$F_1@K$	MAE	$\mathcal{E}$
`granite-3.0-8b-instruct`								
FS	24±10	6±5		0.76±0.20	0.80±0.17	0.73±0.23	0.15±0.14	
FV	20±8	2±2	8±4	0.64±0.18	0.74±0.16	0.62±0.21	0.14±0.12	
VS	21±9	1±1	8±4	0.67±0.18	0.74±0.17	0.65±0.21	0.14±0.12	
FR1	23±9	3±2	4±3	0.73±0.19	0.79±0.15	0.70±0.22	0.14±0.14	0.08±0.01
FR2	24±10	5±5	0±1	0.78±0.20	0.80±0.18	0.75±0.23	0.19±0.16	0.04±0.01
FR3	24±10	5±6	0±1	0.78±0.21	0.79±0.18	0.74±0.24	0.18±0.16	0.04±0.01
`llama-3.1-70b-instruct`								
FS	23±10	7±5		0.73±0.20	0.82±0.15	0.71±0.23	0.14±0.13	
FV	23±10	3±2	4±3	0.72±0.20	0.82±0.16	0.70±0.23	0.13±0.12	
VS	23±10	1±1	6±5	0.72±0.21	0.81±0.15	0.70±0.24	0.13±0.12	
FR1	21±9	2±1	7±4	0.66±0.22	0.81±0.15	0.64±0.22	0.11±0.12	0.06±0.01
FR2	24±10	2±2	3±3	0.77±0.20	0.83±0.17	0.74±0.23	0.16±0.14	0.03±0.01
FR3	24±10	2±2	3±3	0.77±0.20	0.83±0.17	0.74±0.23	0.16±0.14	0.03±0.01
`mixtral-8x22b-instruct`								
FS	24±10	6±5		0.75±0.20	0.83±0.16	0.72±0.23	0.15±0.14	
FV	22±9	2±2	5±4	0.71±0.20	0.82±0.15	0.69±0.23	0.12±0.12	
VS	23±10	1±1	5±4	0.73±0.21	0.81±0.16	0.71±0.24	0.13±0.13	
FR1	22±9	1±1	6±4	0.71±0.20	0.81±0.15	0.69±0.23	0.13±0.13	0.05±0.01
FR2	25±10	1±2	3±3	0.81±0.18	0.82±0.17	0.77±0.22	0.19±0.16	0.03±0.01
FR3	25±10	2±4	3±3	0.80±0.19	0.82±0.17	0.77±0.22	0.19±0.17	0.03±0.01

Table 9: Results obtained on the labeled Biographies dataset using Google Search retrieved contexts.

Assessor	# S	# C	# U	Pr ↑	$F_1@K$ ↑	$\mathcal{E}$ ↓
`granite-3.0-8b-instruct`						
FS	17±6	5±3		0.76±0.15	0.74±0.17	
FB	8±4	0±1	13±4	0.35±0.15	0.36±0.17	
FV	12±5	4±2	5±2	0.55±0.16	0.55±0.18	
FR1 (ours)	4±3	1±1	16±4	0.19±0.14	0.19±0.13	0.14±0.01
FR2 (ours)	10±6	9±5	2±2	0.46±0.22	0.47±0.24	0.09±0.03
FR3 (ours)	11±6	8±4	2±2	0.47±0.22	0.48±0.24	0.10±0.03
`llama-3.1-70b-instruct`						
FS	15±5	7±3		0.69±0.15	0.68±0.16	
FB	8±5	0	13±4	0.37±0.19	0.38±0.21	
FV	5±4	0	16±5	0.25±0.16	0.25±0.18	
FR1 (ours)	5±4	0	17±4	0.21±0.16	0.22±0.18	0.13±0.02
FR2 (ours)	10±7	1±1	10±5	0.45±0.26	0.46±0.28	0.09±0.04
FR3 (ours)	10±7	1±1	10±5	0.44±0.25	0.45±0.27	0.09±0.04
`mixtral-8x22b-instruct`						
FS	16±5	6±3		0.71±0.15	0.70±0.16	
FB	9±5	0	12±4	0.43±0.17	0.43±0.19	
FV	7±4	0	14±5	0.34±0.17	0.34±0.19	
FR1 (ours)	5±4	0	17±4	0.22±0.18	0.23±0.20	0.12±0.02
FR2 (ours)	11±5	0	11±5	0.46±0.28	0.47±0.30	0.09±0.04
FR3 (ours)	11±8	0	11±5	0.46±0.30	0.47±0.30	0.09±0.04

Table 10: Results obtained on the unlabeled AskH dataset using Wikipedia retrieved contexts.

Assessor	# S	# C	# U	Pr ↑	$F_1@K$ ↑	$\mathcal{E}$ ↓
`granite-3.0-8b-instruct`						
FS	18±6	3±2		0.82±0.13	0.81±0.15	
FV	14±5	1±1	7±3	0.62±0.16	0.62±0.19	
VS	14±5	3±2	3±2	0.65±0.15	0.65±0.15	
FR1 (ours)	13±5	4±2	4±2	0.60±0.17	0.60±0.20	0.08±0.02
FR2 (ours)	14±8	7±5	0	0.63±0.27	0.62±0.28	0.04±0.03
FR3 (ours)	15±7	7±5	0	0.67±0.24	0.66±0.25	0.06±0.03
`llama-3.1-70b-instruct`						
FS	18±5	3±2		0.82±0.12	0.80±0.14	
FV	16±6	1±1	5±3	0.71±0.18	0.70±0.20	
VS	15±6	0	7±3	0.66±0.18	0.65±0.20	
FR1 (ours)	12±6	1±1	8±4	0.53±0.19	0.54±0.22	0.08±0.03
FR2 (ours)	17±6	1±1	3±3	0.76±0.18	0.74±0.20	0.04±0.03
FR3 (ours)	17±6	2±1	3±3	0.75±0.18	0.74±0.20	0.04±0.03
`mixtral-8x22b-instruct`						
FS	18±6	3±2		0.82±0.13	0.80±0.15	
FV	15±6	0±1	6±3	0.67±0.18	0.67±0.21	
VS	15±6	0±1	6±3	0.68±0.18	0.67±0.20	
FR1 (ours)	14±6	0	8±4	0.60±0.20	0.60±0.22	0.07±0.03
FR2 (ours)	18±7	0	3±3	0.80±0.17	0.79±0.19	0.04±0.03
FR3 (ours)	18±7	0	3±3	0.80±0.17	0.79±0.19	0.04±0.03

Table 11: Results obtained on the unlabeled AskH dataset using Google Search retrieved contexts.

Method	# S	# C	# U	Pr	$F_1@K$	$\mathcal{E}$
granite-3.0-8b-instruct						
FS	17±6	6±4		0.72±0.19	0.71±20	
FV	9±5	0	13±5	0.38±0.18	0.38±0.18	
VS	15±6	3±2	4±2	0.63±0.16	0.63±0.18	
FR1	8±6	0±0	14±6	0.34±0.23	0.34±0.24	0.11±0.03
FR2	15±9	0±1	7±6	0.64±0.29	0.63±0.29	0.06±0.04
FR3	13±8	7±6	2±3	0.55±0.27	0.54±0.28	0.09±0.03
llama-3.1-70b-instruct						
FS	16±6	7±4		0.69±0.18	0.68±0.19	
FV	10±6	0	12±5	0.43±0.22	0.43±0.23	
VS	5±4	0	17±4	0.24±0.18	0.24±0.18	
FR1	5±5	0±0	17±6	0.24±0.18	0.24±0.19	0.12±0.02
FR2	11±8	1±1	10±6	0.49±0.29	0.49±0.29	0.09±0.04
FR3	12±8	2±2	9±6	0.49±0.28	0.50±0.29	0.09±0.04
mixtral-8x22b-instruct						
FS	17±6	6±4		0.72±0.19	0.71±0.20	
FV	11±6	0	11±5	0.50±0.21	0.50±0.21	
VS	10±6	0	12±6	0.43±0.22	0.43±0.22	
FR1	6±5	0±0	17±6	0.25±0.20	0.25±0.21	0.12±0.03
FR2	12±8	0±0	10±6	0.51±0.29	0.51±0.30	0.08±0.04
FR3	12±8	0±0	10±6	0.51±0.30	0.51±0.30	0.08±0.04

Table 12: Results obtained on the unlabeled Books dataset using Wikipedia retrieved contexts.

Method	# S	# C	# U	Pr	$F_1@K$	$\mathcal{E}$
granite-3.0-8b-instruct						
FS	20±7	2±2		0.87±0.13	0.84±0.15	
FV	16±6	0±1	6±3	0.71±0.17	0.70±0.18	
VS	18±7	2±2	3±2	0.76±0.17	0.75±0.19	
FR1	18±8	0±1	3±3	0.79±0.19	0.77±0.20	0.04±0.03
FR2	21±7	1±1	0±1	0.90±0.14	0.86±0.15	0.02±0.02
FR3	17±8	5±5	0	0.74±0.27	0.72±0.27	0.04±0.03
llama-3.1-70b-instruct						
FS	20±7	3±3		0.84±0.15	0.82±0.17	
FV	18±8	0±1	4±4	0.78±0.20	0.76±0.21	
VS	17±8	0	5±5	0.72±0.23	0.71±0.23	
FR1	14±7	1±1	7±5	0.62±0.23	0.62±0.24	0.07±0.03
FR2	19±8	1±1	3±3	0.80±0.21	0.78±0.21	0.04±0.03
FR3	18±7	2±6	2±2	0.80±0.20	0.78±0.21	0.04±0.03
mixtral-8x22b-instruct						
FS	20±7	3±3		0.84±0.16	0.82±0.18	
FV	18±7	0	4±4	0.76±0.20	0.74±0.21	
VS	18±8	0	4±4	0.79±0.20	0.77±0.21	
FR1	16±7	0±0	6±4	0.69±0.22	0.68±0.23	0.06±0.03
FR2	20±7	0±0	2±3	0.86±0.17	0.83±0.18	0.03±0.03
FR3	20±7	0±0	2±3	0.86±0.17	0.83±0.18	0.03±0.03

Table 13: Results obtained on the unlabeled Books dataset using Google Search retrieved contexts.

Method	# S	# C	# U	Pr	$F_1@K$	$\mathcal{E}$
granite-3.0-8b-instruct						
FS	17±5	4±3		0.77±0.15	0.77±0.17	
FV	8±3	0	12±4	0.39±0.15	0.40±0.16	
VS	13±5	4±2	4±2	0.59±0.17	0.60±0.18	
FR1	5±4	1±1	15±4	0.23±0.15	0.24±0.16	0.13±0.01
FR2	14±6	4±3	3±2	0.63±0.22	0.63±0.24	0.08±0.03
FR3	14±6	4±3	3±2	0.64±0.21	0.64±0.23	0.08±0.03
llama-3.1-70b-instruct						
FS	16±5	5±3		0.74±0.14	0.74±0.16	
FV	10±5	0	10±4	0.47±0.21	0.47±0.22	
VS	6±4	0	15±5	0.29±0.18	0.30±0.19	
FR1	5±4	0	15±4	0.25±0.19	0.26±0.20	0.12±0.02
FR2	12±7	1±1	8±5	0.54±0.28	0.55±0.28	0.08±0.04
FR3	12±7	1±1	8±5	0.54±0.27	0.55±0.28	0.08±0.04
mixtral-8x22b-instruct						
FS	17±5	4±3		0.78±0.14	0.77±0.15	
FV	12±5	0	9±4	0.55±0.18	0.55±0.19	
VS	9±5	0	11±4	0.44±0.19	0.44±0.21	
FR1	6±5	0	15±4	0.27±0.20	0.28±0.21	0.12±0.03
FR2	12±7	0	8±6	0.55±0.29	0.56±0.30	0.08±0.04
FR3	12±7	0	9±6	0.55±0.31	0.56±0.31	0.08±0.04

Table 14: Results obtained on the unlabeled ELI5 dataset using Wikipedia retrieved contexts.

Method	# S	# C	# U	Pr	$F_1@K$	$\mathcal{E}$
granite-3.0-8b-instruct						
FS	18±5	3±2		0.85±0.11	0.84±0.13	
FV	15±5	0±1	5±2	0.69±0.14	0.70±0.16	
VS	16±5	2±2	3±1	0.71±0.14	0.72±0.17	
FR1	14±5	3±2	3±2	0.66±0.17	0.67±0.19	0.08±0.02
FR2	18±6	3±4	0	0.82±0.21	0.80±0.21	0.03±0.02
FR3	16±6	3±3	0	0.83±0.18	0.82±0.18	0.03±0.03
llama-3.1-70b-instruct						
FS	19±5	3±2		0.86±0.12	0.84±0.13	
FV	18±5	1±1	3±2	0.81±0.16	0.80±0.17	
VS	17±5	0	4±3	0.78±0.16	0.77±0.17	
FR1	14±6	1±1	6±4	0.65±0.20	0.66±0.21	0.07±0.03
FR2	19±5	1±1	1±1	0.86±0.14	0.85±0.16	0.03±0.03
FR3	19±6	1±1	1±1	0.86±0.15	0.84±0.16	0.03±0.03
mixtral-8x22b-instruct						
FS	19±5	2±2		0.87±0.11	0.86±0.13	
FV	17±5	0	3±2	0.79±0.15	0.79±0.17	
VS	17±5	0±1	3±2	0.79±0.15	0.78±0.17	
FR1	16±5	0	5±3	0.74±0.18	0.74±0.19	0.05±0.03
FR2	20±5	0	1±2	0.90±0.12	0.88±0.13	0.02±0.02
FR3	20±5	0	1±2	0.90±0.12	0.88±0.13	0.02±0.02

Table 15: Results obtained on the unlabeled ELI5 dataset using Google Search retrieved contexts.

Method	# S	# C	# U	Pr	$F_1@K$	$\mathcal{E}$
`granite-3.0-8b-instruct`						
FS	22±9	4±2		0.83±0.10	0.82±0.12	
FV	13±6	0±1	12±5	0.50±0.15	0.50±0.16	
VS	18±8	4±3	4±2	0.69±0.14	0.69±0.16	
FR1	12±8	0±0	13±7	0.46±0.23	0.47±0.24	0.09±0.03
FR2	20±11	0±1	4±6	0.79±0.24	0.78±0.25	0.04±0.04
FR3	15±11	8±6	2±6	0.58±0.28	0.58±0.28	0.08±0.04
`llama-3.1-70b-instruct`						
FS	18±9	7±4		0.71±0.16	0.71±0.17	
FV	14±9	0	11±6	0.53±0.21	0.54±0.21	
VS	10±7	0	15±7	0.41±0.21	0.41±0.22	
FR1	10±8	0±1	15±6	0.39±0.21	0.39±0.22	0.11±0.02
FR2	18±10	1±1	5±6	0.70±0.26	0.70±0.26	0.06±0.04
FR3	18±10	1±1	5±6	0.71±0.26	0.70±0.26	0.06±0.04
`mixtral-8x22b-instruct`						
FS	20±9	6±4		0.76±0.16	0.75±0.17	
FV	15±8	0	10±5	0.59±0.18	0.59±0.18	
VS	15±8	0	10±5	0.57±0.19	0.57±0.20	
FR1	11±8	0±0	14±7	0.41±0.22	0.42±0.23	0.10±0.03
FR2	19±10	0±0	6±6	0.74±0.26	0.74±0.26	0.05±0.04
FR3	19±10	0±0	6±7	0.74±0.26	0.74±0.26	0.05±0.04

Table 16: Results obtained on the unlabeled LFObj dataset using Wikipedia retrieved contexts.

Method	# S	# C	# U	Pr	$F_1@K$	$\mathcal{E}$
`granite-3.0-8b-instruct`						
FS	24±9	1±1		0.93±0.07	0.91±0.09	
FV	20±8	0	4±2	0.79±0.10	0.79±0.12	
VS	18±8	3±2	4±2	0.68±0.14	0.69±0.15	
FR1	24±9	0±0	1±2	0.93±0.09	0.91±0.10	0.02±0.02
FR2	25±5	1±9	0±0	0.97±0.07	0.94±0.09	0.00±0.01
FR3	23±7	4±16	0	0.89±0.22	0.86±0.22	0.02±0.02
`llama-3.1-70b-instruct`						
FS	23±9	2±2		0.91±0.09	0.89±0.10	
FV	23±9	0±1	1±2	0.91±0.10	0.89±0.12	
VS	10±7	0	15±7	0.40±0.21	0.40±0.22	
FR1	22±9	0±1	2±3	0.85±0.13	0.84±0.14	0.03±0.02
FR2	24±5	1±9	0±1	0.94±0.10	0.92±0.11	0.01±0.01
FR3	24±5	1±10	0	0.93±0.13	0.91±0.14	0.01±0.01
`mixtral-8x22b-instruct`						
FS	24±9	1±2		0.93±0.08	0.91±0.10	
FV	23±9	0±1	2±2	0.90±0.10	0.88±0.11	
VS	23±9	0	2±4	0.88±0.16	0.86±0.16	
FR1	23±9	0±0	2±2	0.90±0.10	0.88±0.12	0.03±0.02
FR2	24±5	0±9	0±1	0.96±0.09	0.93±0.10	0.01±0.01
FR3	24±5	0±9	0±1	0.96±0.09	0.94±0.10	0.01±0.01

Table 17: Results obtained on the LFObj dataset using Google Search retrieved contexts.

Table 18: Prompt template for few-shot atomic unit decomposition - Atomizer stage

,

Atomic unit decomposition (Few-Shot)
Instructions:

1. You are given a paragraph. Your task is to break the sentence down into a list of atomic statements without adding any new information.
2. An atomic statement is a sentence containing a singular piece of information directly extracted from the provided paragraph.
3. Atomic statements may contradict one another.
4. The paragraph may contain information that is factually incorrect. Even in such cases, you are not to alter any information contained in the paragraph and must produce atomic statements that are completely faithful to the information in the paragraph.
5. Each atomic statement in the outputted list should check a different piece of information found explicitly in the paragraph.
6. Each atomic statement is standalone in that any actual nouns or proper nouns should be used in place of pronouns or anaphoras.
7. Each atomic statement must not include any information beyond what is explicitly stated in the provided paragraph.
8. Where possible, avoid paraphrasing and instead try to only use language used in the paragraph without introducing new words.
9. Use the previous examples to learn how to do this.
10. You should only output the atomic statement as a list, with each item starting with "- ". Do not include other formatting.
11. Your task is to do this for the last paragraph that is given.

Few-Shot Examples:

Please breakdown the following paragraph into independent statements: Glenn Allen Anzalone (born June 23, 1955), better known by his stage name Glenn Danzig, is an American singer, songwriter, musician, and record producer. He is the founder of the rock bands Misfits, Samhain, and Danzig. He owns the Evilive record label as well as Verotik, an adult-oriented comic book publishing company.

- Glenn Allen Anzalone was born on June 23, 1955.
- Glenn Allen Anzalone is better known by his stage name Glenn Danzig.
- Glenn Danzig is an American singer, songwriter, musician, and record producer.
- Glenn Danzig is the founder of several rock bands, including Misfits, Samhain, and Danzig.
- Glenn Danzig owns the Evilive record label.
- Glenn Danzig owns Verotik, which is an adult-oriented comic book publishing company.

Please breakdown the following paragraph into independent statements: Luiz Inácio Lula da Silva (born 27 October 1945), also known as Lula da Silva or simply Lula, is a Brazilian politician who is the 39th and current president of Brazil since 2023. A member of the Workers' Party, Lula was also the 35th president from 2003 to 2010. He also holds the presidency of the G20 since 2023. Lula quit school after second grade to work, and did not learn to read until he was ten years old. As a teenager, he worked as a metalworker and became a trade unionist.

- Luiz Inácio Lula da Silva was born on October 27, 1945.
- Luiz Inácio Lula da Silva is also known as Lula da Silva or simply Lula.
- Lula is a Brazilian politician.
- Lula is the 39th and current president of Brazil since 2023.
- Lula is a member of the Workers' Party.
- Lula served as the 35th president of Brazil from 2003 to 2010.
- Lula holds the presidency of the G20 since 2023.
- Lula quit school after the second grade to work.
- Lula did not learn to read until he was ten years old.
- As a teenager, Lula worked as a metalworker.
- Lula became a trade unionist.

Please breakdown the following paragraph into independent statements: {}

Table 19: Prompt template for few-shot decontextualization - Reviser stage

Decontextualization (Few-Shot)

Instructions:
1. You are given a statement and a context that the statement belongs to. Your task is to modify the statement so that any pronouns or anaphora (words like "it," "they," "this") are replaced with the noun or proper noun that they refer to, such that the sentence remains clear without referring to the original context.
2. Return only the revised, standalone version of the statement without adding any information that is not already contained within the original statement. 3. If the statement requires no changes, return the original statement as-is without any explanation.
4. The statement that you return must start with #### and finish with #### as follows: ####<statement>####
5. Do not include any explanation or any additional formatting including any lead-in or sign-off text.
6. Learn from the provided examples below and use that knowledge to amend the last example yourself.

Few-Shot Examples:

Example 1: Context: John went to the store.
Statement: He bought some apples.
Standalone: ####John bought some apples.####

Example 2: Context: The presentation covered various aspects of climate change, including sea level rise.
Statement: This was a key part of the discussion.
Standalone: ####Sea level rise was a key part of the discussion.####

Example 3: Context: Maria Sanchez is a renowned marine biologist known for her groundbreaking research on coral reef ecosystems. Her work has contributed to the preservation of many endangered coral species, and she is often invited to speak at international conferences on environmental conservation.
Statement: She presented her findings at the conference last year.
Standalone: ####Maria Sanchez presented her findings at the conference last year.####

Example 4: Context: Nathan Carter is a best-selling science fiction author famous for his dystopian novels that explore the intersection of technology and society. His latest book, The Edge of Something, received widespread critical acclaim for its imaginative world-building and its poignant commentary on artificial cacti.
Statement: It was praised for its thought-provoking themes.
Standalone: ####The Edge of Tomorrow was praised for its thought-provoking themes.####

Now perform the task for the following example: Context: {}
Statement: {}
Standalone:

Table 20: Prompt template for few-shot NLI relation extraction.

NLI relation prompting (Few-Shot)

Instructions:
1. You are given a premise and a hypothesis and a context. Your task is to identify the relationship between them: does the premise entail, contradict, or remain neutral toward the hypothesis?
2. Your only output must be one of: (entailment | contradiction | neutral) without any lead-in, sign-off, new lines or any other formatting.
3. Do not provide any explanation or rationale to your output.
4. Use the following examples to learn how to do this, and provide your output for the last example given.

Few-Shot Examples:

Premise: Contrary to popular belief, the Great Wall is not visible from space without aid.
Hypothesis: Astronauts have managed to see the wall from Space unaided.
Context: The Great Wall of China is one of the most famous landmarks in the world. It stretches over 13,000 miles and was primarily built during the Ming Dynasty. Contrary to popular belief, the Great Wall is not visible from space without aid. The primary purpose of the Great Wall was to protect against invasions from nomadic tribes. The wall is a UNESCO World Heritage site and attracts millions of tourists each year. Astronauts have managed to see the wall from Space unaided.
Output: Contradiction

Premise: It is estimated that around 20 percent of the world's oxygen is produced by the Amazon.
Hypothesis: However, the Amazon Rainforest produces no significant amount of oxygen as the plants consume almost all of it through respiration.
Context: The Amazon Rainforest is often referred to as the lungs of the Earth due to its vast capacity to produce oxygen. This immense rainforest spans nine countries in South America. It is estimated that around 20 percent of the world's oxygen is produced by the Amazon. However, the Amazon Rainforest produces no significant amount of oxygen as the plants consume almost all of it through respiration. The biodiversity of the Amazon is unparalleled, hosting millions of species of plants and animals.
Output: Contradiction

Premise: It is estimated that around 20 percent of the world's oxygen is produced by the Amazon.
Hypothesis: This immense rainforest spans nine countries in South America.
Context: The Amazon Rainforest is often referred to as the lungs of the Earth due to its vast capacity to produce oxygen. This immense rainforest spans nine countries in South America. It is estimated that around 20 percent of the world's oxygen is produced by the Amazon. However, the Amazon Rainforest produces no significant amount of oxygen as the plants consume almost all of it through respiration. The biodiversity of the Amazon is unparalleled, hosting millions of species of plants and animals.
Output: Neutral

Premise: It is estimated that around 20 percent of the world's oxygen is produced by the Amazon.
Hypothesis: The Amazon Rainforest is often referred to as the lungs of the Earth due to its vast capacity to produce oxygen.
Context: The Amazon Rainforest is often referred to as the lungs of the Earth due to its vast capacity to produce oxygen. This immense rainforest spans nine countries in South America. It is estimated that around 20 percent of the world's oxygen is produced by the Amazon. However, the Amazon Rainforest produces no significant amount of oxygen as the plants consume almost all of it through respiration. The biodiversity of the Amazon is unparalleled, hosting millions of species of plants and animals.
Output: Entailment

Premise: {}
Hypothesis: {}
Context: {}
Output:

Table 21: Prompt template used by the FactScore (FS) assessor.

Answer the input question based on the given context.
{CONTEXTS}
Input: {ATOM} True or False?
Output:

Table 22: Prompt template used by the FactVerify (FV) assessor.

Instructions:

You are provided with a STATEMENT and several KNOWLEDGE points.

Your task is to evaluate the relationship between the STATEMENT and the
KNOWLEDGE, following the steps outlined below:

1. Summarize KNOWLEDGE Points: Carefully analyze the KNOWLEDGE points one by one and assess their relevance to the STATEMENT.

Summarize the main points of the KNOWLEDGE.

2. Evaluate Evidence: Based on your reasoning:

- If the KNOWLEDGE strongly implies or directly supports the STATEMENT, explain the supporting evidence.
- If the KNOWLEDGE contradicts the STATEMENT, identify and explain the conflicting evidence.
- If the KNOWLEDGE is insufficient to confirm or deny the STATEMENT, explain why the evidence is inconclusive.

3. Restate the STATEMENT: After considering the evidence, restate the STATEMENT to maintain clarity.

4. Final Answer: Based on your reasoning and the STATEMENT, determine your final answer.

Your final answer must be one of the following, wrapped in square brackets:

- [Supported] if the STATEMENT is supported by the KNOWLEDGE.
- [Contradicted] if the STATEMENT is contradicted by the KNOWLEDGE.
- [Undecided] if the KNOWLEDGE is insufficient to verify the STATEMENT.

Your task:

KNOWLEDGE:

{}

STATEMENT:

{}

Table 23: Prompt template used by the VeriScore (VS) assessor.

Instructions
You need to judge whether a claim is supported or contradicted by Google search results, or whether there is no enough information to make the judgement. When doing the task, take into consideration whether the link of the search result is of a trustworthy source. Mark your answer with ### signs.
Below are the definitions of the three categories:
Supported: A claim is supported by the search results if everything in the claim is supported and nothing is contradicted by the search results. There can be some search results that are not fully related to the claim.
Contradicted: A claim is contradicted by the search results if something in the claim is contradicted by some search results. There should be no search result that supports the same part.
Undecided: A claim is inconclusive based on the search results if:
- a part of a claim cannot be verified by the search results,
- a part of a claim is supported and contradicted by different pieces of evidence,
- the entity/person mentioned in the claim has no clear referent (e.g., "the approach", "Emily", "a book").

Here are some examples:
Claim: Characters Lenny and Carl on The Simpsons are hearing but are depicted as close friends of the Simpsons family.
Search result 1
Title: Character Spotlight: Lenny Leonard and Carl Carlson
Content: Their friendship is a pretty singular aspect on the show – save Bart and Milhouse (or to some degree, Mr. Burns and Smithers) – they always ...
Link: https://nohomers.net/forums/index.php?threads/character-spotlight-lenny-leonard-and-carl-carlson-barflies.23798/
Search result 2
Title: The Simpsons: Lenny and Carl's History, Explained - CBR
Content: Introduced in the show's first season, the pair were portrayed as background characters at Homer's work, usually appearing together in minor ...
Link: https://www.cbr.com/the-simpsons-lenny-carl-history-explained/
Search result 3
Title: Are Lennie and Carl Homer Simpson's real or fake friends? - Quora
Content: Lenni is a pal, Carl doesn't consider any of them to be 'friends' they're just shallow guys he hangs out with. Lenny and Carl have a special ...
Link: https://www.quora.com/Are-Lennie-and-Carl-Homer-Simpson-s-real-or-fake-friends
Your decision: ###Undecided###
Claim: The championship match of the FIFA World Cup 2026 will be hosted by the United States.
Search result 1
Title: World Cup 2026 | New York New Jersey to host final - FIFA
Content: New York New Jersey Stadium has been confirmed as the location for the FIFA World Cup 26 final on Sunday, 19 July 2026. The full match schedule for the ...
Link:https://www.fifa.com/fifaplus/en/tournaments/mens/worldcup/canadamexicousa2026/articles/new-york-new-jersey-stadium-host-world-cup-2026-final
Search result 2
Title: 2026 FIFA World Cup - Wikipedia
Content: The tournament will take place from June 11 to July 19, 2026. It will be jointly hosted by 16 cities in three North American countries: Canada, Mexico, and the ...
Link: https://en.wikipedia.org/wiki/2026_FIFA_World_Cup
Search result 3
Title: World Cup 2026 | Dallas to host nine matches - FIFA
Content: Dallas Stadium will host nine matches from the FIFA World Cup 26, including four knockout games in the latter stages of the tournament.
Link:https://www.fifa.com/fifaplus/en/tournaments/mens/worldcup/canadamexicousa2026/articles/dallas-stadium-host-nine-world-cup-matches
Your decision: ###Supported###
Claim: Vikings used their longships to transport livestock.
Search result 1
Title: How did the Vikings transport animals on their ships? - Quora
Content: The Vikings transported horses overseas in boats very similar to Viking longships, but with flat flooring built within the hulls, which allowed ...
Link: https://www.quora.com/How-did-the-Vikings-transport-animals-on-their-ships
Search result 2
Title: The Truth Behind Vikings Ships
Content: They could land on any beach, permitting lightning-quick embarking and attacks. Great loads could be carried, including horses and livestock.
Link: https://www.vikings.com/news/the-truth-behind-vikings-ships-18274806
Search result 3
Title: Viking ships | Royal Museums Greenwich
Content: Cargo vessels were used to carry trade goods and possessions. They were wider than the longships and travelled more slowly.
Link: https://www.rmg.co.uk/stories/topics/viking-ships
Your decision: ###Contradicted###
Your task:
Claim: {}
{}
Your decision:

Table 24: Prompt template used by DeepSeek-v3.

Instructions:
You are provided with a STATEMENT and several external EVIDENCE points.
Your task is to use your internal knowledge as well as the provided EVIDENCE to
reason about the relationship between the STATEMENT and the EVIDENCE.

1. Carefully analyze the EVIDENCE points one by one and assess their relevance to the STATEMENT.
2. Use your reasoning and your internal knowledge, evaluate the EVIDENCE as follows:
- If the EVIDENCE strongly implies or directly supports the STATEMENT, explain the supporting evidence.
- If the EVIDENCE contradicts the STATEMENT, identify and explain the conflicting evidence.
- If the EVIDENCE is insufficient to confirm or deny the STATEMENT, explain why the evidence is inconclusive.
3. Based on your reasoning and your explanations, determine your final answer.
Your final answer must be one of the following, wrapped in square brackets:
- [Supported] if the EVIDENCE supports the STATEMENT.
- [Contradicted] if the EVIDENCE contradicts the STATEMENT.
- [Undecided] if the EVIDENCE is insufficient to assess the STATEMENT.

Your task:
EVIDENCE: {}
STATEMENT: