CONFORMAL FACTUALITY WITH LEARNED STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models (LLMs) are increasingly used to generate content, evaluating the correctness and factuality of their outputs has become essential. We introduce a new conformal inference framework for flagging factually incorrect claims in a generated text while controlling the false negative rate (FNR). In contrast to existing conformal inference-based methods that assess the correctness of each claim in isolation, our approach constructs a score of factuality by leveraging the natural language inference (NLI) relationships among claims within the text. We evaluate the effectiveness of our approach on three datasets, including FABLES, FELM, and a newly annotated dataset.[1] The results demonstrate that our method controls the FNR while significantly improving retention and accuracy over baseline procedures. Remarkably, our method, even when instantiated with free open-source models, matches the performance obtained when using state-of-the-art closed-source LLMs.

1 INTRODUCTION

As large language models (LLMs) and generative artificial intelligence (AI) begin to be adopted in all spheres of human activity, vast quantities of AI-generated text are entering scientific and industrial workflows. However, even as AI adoption has surged, efficient methods to verify the correctness of AI-generated text remain in short supply. In recent years, motivated by these considerations, there has been a slew of proposed algorithms to assess *factuality* and detect hallucinations in LLM-generated text (Min et al., 2023; Manakul et al., 2023; Wei et al., 2024; Song et al., 2024; Tang et al., 2024; Hu et al., 2024; Chern et al., 2025; Marinescu et al., 2025), building on the broader literature on automated fact-checking that predates modern generative AI (Hassan et al., 2017; Thorne et al., 2018; Augenstein et al., 2019; Guo et al., 2022).

One particularly promising line of work focuses on factuality algorithms that are accompanied by rigorous *statistical guarantees* on their outputs (Quach et al., 2024; Mohri & Hashimoto, 2024; Cherian et al., 2024). In particular, the seminal work Mohri & Hashimoto (2024) proposes the *conformal factuality* algorithm, which filters a given piece of text in such a way that the output is completely correct with a prespecified confidence. Specifically, the algorithm first partitions the input text into atomic *claims;* each claim is assigned a scalar *score,* and all claims with scores exceeding a learned threshold are retained. In order to estimate this threshold on the scores, Mohri & Hashimoto (2024) leverage tools from the rich literature on *conformal prediction* and distribution-free uncertainty quantification (Vovk et al., 2005). Follow-up work (Cherian et al., 2024) extends the results to control the expectation of arbitrary monotone losses via the *conformal risk control* framework of Angelopoulos et al. (2024).

However, a crucial limitation of prior work is the fact that each claim is scored *in isolation,* without explicitly taking into account the logical relationships that exist between the claims. For instance, suppose that claim A is factual, and that claim B is logically contradicted by claim A; then claim B should be nonfactual. Motivated by this observation, in this paper, we propose a novel conformal factuality algorithm that leverages the logical interrelations between pairs of claims in order to perform better filtering, while preserving distribution-free factuality guarantees on the output text.

[1] Code is available at https://anonymous.4open.science/r/Conformal-Factuality-with-Learned-Structure-4F84

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

Specifically, we leverage ideas from the field of *Natural Language Inference* (NLI) to categorize possible logical relations between pairs of claims. Figure 1 illustrates our running example, which highlights fundamental differences between our approach to scoring claims and prior work (Mohri & Hashimoto, 2024; Cherian et al., 2024).

Input text x: Lev Landau theoretically demonstrated the existence of transition radiation, supersymmetric particles and special properties of condensed matter. For the latter he received Nobel Prize in 1963.

Extracted claims $\mathcal{C}(x)$:
c_1: Lev Landau received a Nobel Prize for discovering special properties of condensed matter;
c_2: Lev Landau received Nobel Prize in 1963;
c_3: Lev Landau theoretically demonstrated the existence of transition radiation;
c_4: Lev Landau theoretically demonstrated special properties of condensed matter;
c_5: Lev Landau theoretically demonstrated the existence of supersymmetric particles.

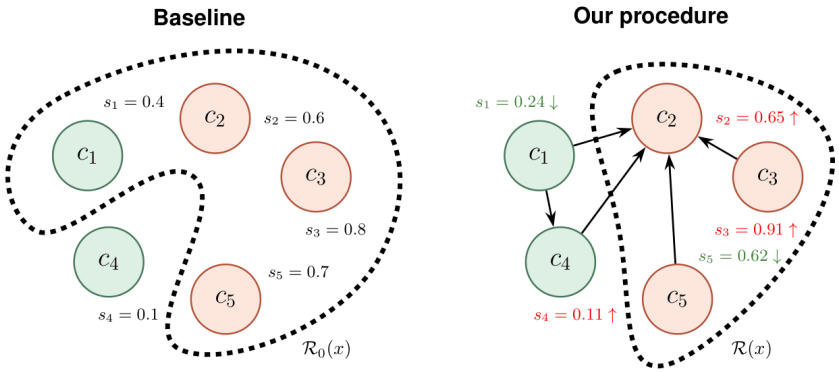


Figure 1: Illustration of our proposed algorithm. Given an input text x, we begin by extracting a set $\mathcal{C}(x)$ of constituent claims; here, the factual claims (c_1 and c_4) are highlighted in green, while the claims in red (c_2, c_3, and c_5) are non-factual. The conformal factuality algorithm (Mohri & Hashimoto, 2024), depicted at left, uses scores $s_1, \ldots, s_5$ computed on *individual* claims to reject a subset $\mathcal{R}_0(x)$ that it deems are non-factual. In contrast, our algorithm leverages the *logical interrelations* between pairs of claims to update the scores and compute a smaller rejection set $\mathcal{R}(x)$.

1.1 Contributions

We can summarize our contributions as follows:

Methodological contributions. Unlike previous works that update beliefs of claim correctness (Hou et al., 2025; Rubin-Toles et al., 2025; Marinescu et al., 2025) using NLI relations, we apply conformal risk control (Angelopoulos et al., 2024) after an update to control the FNR. Moreover, we propose several posterior update strategies to which we apply risk control: based on a combinatorial posterior update, variational inference (Wainwright & Jordan, 2008), and belief tree propagation (Hou et al., 2025).

Empirical validation. We validate our approach on four empirical datasets: FABLES (Kim et al., 2024), FELM (Zhao et al., 2023), WiCE (Kamoi et al., 2023), as well as a newly constructed dataset in our paper. We show that our procedure outperforms a strong baseline, and can in fact match the accuracy and retention of current state-of-the-art models. These experiments support our claim that our method can be a practically useful approach for conformal factuality control.

1.2 Related work

Here we provide a brief overview of the literature, focusing on algorithms that perform a posterior probability update, as well as those that apply conformal inference.

Claim-level factuality. Prior work on factuality may be split into two groups: (a) evidence-grounded methods, which verify atomic claims, using retrieved evidence (Min et al., 2023; Marinescu et al., 2025; Tang et al., 2024; Wanner et al., 2025), and (b) retrieval-free approaches, which estimate factuality leveraging a generative model (Liu et al., 2023; Hou et al., 2025). Although search-augmented methods such as FActScore and SAFE are particularly effective when reliable evidence can be retrieved, here we use context-independent factuality scores for illustration (though our methods are also applicable beyond this).

Natural Language Inference refinement. In order to determine relations among claims, we use Natural Language Inference (NLI) inference tools, similar to (Laban et al., 2021; Marinescu et al., 2025), and graph-based approaches for score updates similar to (Rubin-Toles et al., 2025; Hou et al., 2025). One refinement is inspired by probabilistic inference in graphical models, where local factors are combined to update marginal beliefs (Yedidia et al., 2005; Wainwright & Jordan, 2008; Ihler et al., 2005; Elidan et al., 2012). In contrast to previous approaches that either treat scores independently (Mohri & Hashimoto, 2024) or use conditional statistics (Cherian et al., 2024) to change the scores, we exploit pairwise NLI relations to obtain logically-consistent score updates.

Conformal factuality. Conformal prediction provides distribution-free coverage guarantees, corresponding to controlling a zero-one loss (Vovk et al., 2005) and has been extended to control general monotonic losses such as the FNR (Angelopoulos et al., 2024). As mentioned, recent works have applied these ideas to factuality (Mohri & Hashimoto, 2024; Cherian et al., 2024; Rubin-Toles et al., 2025; Wang et al., 2026; Shihab et al., 2026). Our work is different, in that it exploits nontrivial structure between claims linked by NLI relations.

2 PRELIMINARIES

2.1 PROBLEM FORMULATION

Let $\mathcal{X}$ denote the set of all possible texts under consideration, and fix an input text $x \in \mathcal{X}$. Here, x could represent the output of an LLM, a social media post, or an article. Let $\mathcal{S}$ denote the set of atomic claims. In what follows, we assume that any text $x \in \mathcal{X}$ can be decomposed into a fixed number K of claims $(c_1, \ldots, c_K)$. In order to formalize this, we introduce the notion of a *K-decomposer*.

Definition 1 (K-decomposer)**.** *Fix a deterministic rule that decomposes each piece of text x into an ordered tuple of atomic claims $(c_1, \ldots, c_k)$, where k may depend on x. Its associated K-decomposer is the map $\mathcal{X} \mapsto \mathcal{S}^K$ given by*

$$\phi_K(x) := \begin{cases} (c_1, \cdots, c_k, \varnothing, \cdots, \varnothing), & 0 \le k < K, \\ (c_1, \cdots, c_K), & k \ge K, \end{cases}$$

for all $x \in \mathcal{X}$, where $\varnothing$ denotes an empty claim used to pad a k-tuple to reach length K.

For example, if we set $K = 10$, then the text $x =$ "Henri Poincaré was a mathematics professor at Sorbonne University" can be decomposed as

$$\begin{aligned} \phi_K(x) = \big(&\text{Henri Poincaré was a professor,} \\ &\text{Henri Poincaré worked at Sorbonne University, Henri Poincaré taught mathematics,} \\ &\text{Sorbonne University had a professorship in mathematics, } \varnothing, \ldots, \varnothing\big). \end{aligned}$$

We assume that each claim is either *factual* (correct) or *non-factual* (incorrect).[2] Formally, we let $h : \mathcal{S} \to \{0, 1\}$ denote an *oracle verifier* that encodes the correctness of a claim: for each $c \in \mathcal{S}$,

$$h(c) = \begin{cases} 0 & \text{if } c \text{ is correct,} \\ 1 & \text{if } c \text{ is incorrect.} \end{cases}$$

As in prior work (Mohri & Hashimoto, 2024), factuality will always be defined with respect to this oracle verifier. Having introduced the verifier h, we define $\mathcal{H}_1(x) = \{i \in [K] : h(c_i) = 1\}$ to be

[2] When ambiguous claims exist, our results pertain to the set of unambiguous claims; thus, we would redefine the FNR to be computed over only the unambiguous subset.

the set of non-factual claims in x, where $[K] := \{1, \ldots, K\}$. Note that $\mathcal{H}_1(x)$ is precisely the set of claims that our algorithm seeks to filter out.

Calibration data. We assume that we have access to a *calibration set* of texts $\mathcal{D} = \{x_i\}_{i=1}^N$, where $x_1, \ldots, x_N \overset{\text{iid}}{\sim} P$ are sampled from an unknown distribution P of user texts. On this calibration set, we also have access to ground-truth correctness labels $\{\mathcal{H}_1(x_i)\}_{i=1}^N$.

Goal. Our goal is to construct a *decision process* $\mathcal{A} : \mathcal{S}^K \mapsto [K]$ that, given a new text x_{N+1}, returns a set $\mathcal{A}(x_{N+1})$ of claims flagged as incorrect. We evaluate $\mathcal{A}$ using the *false negative rate* (FNR), the average value of the *false negative proportion* (FNP):

$$\text{FNR}(\mathcal{A}, x_{N+1}) := \mathbb{E}\big[\text{FNP}(\mathcal{A}, x_{N+1})\big], \text{ where } \text{FNP}(\mathcal{A}, x_{N+1}) := \frac{|\mathcal{H}_1(x_{N+1}) \setminus \mathcal{A}(x_{N+1})|}{|\mathcal{H}_1(x_{N+1})| \vee 1},$$

and where the expectation is over the randomness of the calibration texts $\{x_i\}_{i=1}^N$. Here, $|\cdot|$ denotes the cardinality of a set, and for $a, b \in \mathbb{R}$ we define $a \vee b := \max\{a, b\}$. Given a target level $\alpha \in (0, 1)$, our goal is for $\mathcal{A}$ to ensure that $\text{FNR}(\mathcal{A}, x_{N+1}) \leq \alpha$.

We choose to control the FNR in order to prioritize the detection of factual errors. When flagged claims are referred for further verification, an unflagged correct claim may mislead users or compromise downstream reasoning. In contrast, unnecessarily flagged correct claims increase the verification burden, which may often be a lesser risk. By measuring the fraction of incorrect claims that escape detection, FNR captures the risk of missed errors; thus, it is a reasonable metric. Since flagging every claim yields zero FNR, our objective is to control the FNR while retaining as much correct content as possible.

2.2 Recap: Conformal Risk Control

Conformal inference (see e.g., Vovk et al., 1999; Saunders et al., 1999; Vovk et al., 2005; Lei et al., 2018, etc) is a flexible framework for predictive inference with finite-sample, distribution-free guarantees. In split conformal prediction (Papadopoulos et al., 2002), a fitted prediction model is calibrated using a separate labeled dataset. Angelopoulos et al. (2024) extend this framework to *conformal risk control* (CRC), which controls the expected value of a broad class of loss functions. Specifically, given $B \in \mathbb{R}$, consider an exchangeable sequence of non-increasing functions $L_i : \Lambda \mapsto (-\infty, B]$ for $i \in [N+1]$. Given a nominal risk level α, the goal of CRC is to use the first N losses to select $\hat{\lambda} \in \Lambda$ such that

$$\mathbb{E}\big[L_{N+1}(\hat{\lambda})\big] \leq \alpha, \tag{1}$$

where $\alpha \in (-\infty, B)$ is a pre-specified error rate.[3] CRC selects $\hat{\lambda}$ to be the first point at which a conservative version of the empirical risk crosses the level α; formally, we define

$$\hat{\lambda} = \inf\left\{\lambda \in \Lambda : \frac{1}{N+1}\sum_{i=1}^{N} L_i(\lambda) + \frac{B}{N+1} \leq \alpha\right\},$$

where we use the convention $\hat{\lambda} = \sup \Lambda$ if the argument of the infimum is empty. Under the stated assumptions, Angelopoulos et al. (2024) establishes the guarantee in (1) in Theorem 1.

In what follows, we apply CRC to control the FNR of our claim-flagging procedure. To see that the FNP is a valid loss for CRC, simply note that the FNP lies in $[0, 1]$ by definition, and is non-increasing due to the fact that the rejection set is increasing in λ. As we shall see, these key properties allow us to calibrate the decision process in order to achieve finite-sample FNR control.

3 Method

3.1 Conformal factuality with textual relations

We present *Conformal factuality with textual relations (ConFacTR)*, a claim-flagging procedure with finite-sample FNR control. Given a labeled dataset $\mathcal{D}$, we split $\mathcal{D}$ into two disjoint folds, a

[3] Here, as is typical in the conformal inference literature, the guarantee is marginal over the randomness of both the calibration losses and the test loss.

training set $\mathcal{D}_{\text{train}}$ and a calibration set $\mathcal{D}_{\text{calib}}$. For notational simplicity, we write $\mathcal{D}_{\text{train}} := \{z_i\}_{i=1}^{n_{\text{train}}}$, $\mathcal{D}_{\text{calib}} = \{x_i\}_{i=1}^{N}$. Given a test text x_{N+1} and its decomposition, we seek to identify the set of incorrect claims within it.

We apply the K-decomposer to $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$, and x_{N+1}. Using the training texts $\mathcal{D}_{\text{train}}$ and the ground-truth labels of their claims, we fit a *scoring function* $s : \mathcal{S} \times \mathcal{X} \mapsto [0, 1]$, where for a given claim c and parent text x, the quantity $s(c; x)$ measures the evidence that claim c in text x is incorrect, with larger values indicating stronger evidence. The scoring function is then held fixed during calibration and deployment.

Given the pretrained scorer s, for any $\lambda \in [0, 1]$, we define the decision process $\mathcal{A}_{\lambda,s}$, which selects exactly those claims with scores of at least $1-\lambda$: given $x \in \mathcal{X}$ with decomposition $x = (c_1, \ldots, c_K)$, we define the set of selected indices to be

$$\mathcal{A}_{\lambda,s}(x) := \{i \in [K] : s(c_i; x) \geq 1 - \lambda\}. \tag{2}$$

Given a target FNR level $\alpha \in (0, 1)$, we conformalize the decision process by setting the threshold to be

$$\hat{\lambda} = \inf\left\{\lambda \in [0, 1] : \frac{1}{N+1}\sum_{i=1}^{N} \text{FNP}(\mathcal{A}_{\lambda,s}, x_i) + \frac{1}{N+1} \leq \alpha\right\}, \tag{3}$$

with $\hat{\lambda} = 1$ if the argument to the infimum is empty. Finally, at test time, we apply the calibrated decision process $\mathcal{A}_{\hat{\lambda},s}$ to the test text x_{N+1} to obtain the set of claims it flags as incorrect. The complete algorithm is summarized in Algorithm 1.

Algorithm 1 Conformal factuality with textual relations (ConFacTR)

Require: Nominal level α, labeled dataset $\mathcal{D}$, test text x_{N+1}, K-decomposer ϕ_K, score-fitting algorithm $\mathcal{B}$, parameter grids Λ.

1: Split the labeled data $\mathcal{D}$ into two disjoint sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{calib}}$.
2: Apply the decomposer $\phi_K(\cdot)$ to each piece of text in $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{calib}}$, and x_{N+1} to obtain their atomic claims.
3: Apply $\mathcal{B}$ to $\{(z_i, \mathcal{H}_1(z_i))\}_{z_i \in \mathcal{D}_{\text{train}}}$ to obtain the scoring function s.
4: **for** $\lambda \in \Lambda$ **do**
5: **for** $i = 1$ to N **do**
6: Compute $\mathcal{A}_{\lambda,s}(x_i)$ according to (2) and its corresponding $\text{FNP}(\mathcal{A}_{\lambda,s}, x_i)$.
7: **end for**
8: **end for**
9: Set $\hat{\lambda}$ as in (3).
10: Compute the rejection set $\mathcal{A}_{\hat{\lambda},s}(x_{N+1})$ according to (2).
11: **return** $\mathcal{A}_{\hat{\lambda},s}(x_{N+1})$.

The following theorem, whose proof is provided in Appendix A.1, states that the ConFacTR algorithm controls the FNR at the nominal level.

Theorem 1. *Suppose that $\mathcal{D}$ and x_{N+1} are jointly exchangeable. Then, for any target level $\alpha \in (0, 1)$, ConFacTR (Algorithm 1) controls FNR at level α:*

$$\text{FNR}\big(\mathcal{A}_{\hat{\lambda},s}, x_{N+1}\big) \leq \alpha.$$

By Theorem 1, the FNR control holds *regardless* of the choice of the pretrained score $s(\cdot)$. While its choice does not affect the validity guarantee, a more informative scoring function can improve the classification of claims, thereby yielding higher retention rates. We discuss this crucial design choice in the following sections.

3.2 Scoring function construction

In previous work, Mohri & Hashimoto (2024) use an LLM to estimate an incorrectness probability $\pi(c) := \widehat{\mathbb{P}}(c \text{ is incorrect})$ for each claim c, and use this estimate as the score $s(c; x)$. Here, $\pi(c)$ is

to be interpreted as the strength of the "belief" of the LLM that the claim c is incorrect, rather than a probability with a frequentist interpretation.

Consequently, the algorithm introduced by Mohri & Hashimoto (2024) scores each claim within a piece of text separately, without accounting for the relationships between claims. In practice, however, claims within a piece of text may very well exhibit logical relationships that provide additional information about their correctness. To leverage these relationships, we combine ideas from NLI with tools from variational inference (see e.g., Jaakkola, 1997; Wainwright & Jordan, 2008) to improve the claim-level scores.

3.2.1 Natural Language Inference scores

We begin by introducing the basic categories of NLI relationships between pairs of claims (Rubin-Toles et al., 2025; Marinescu et al., 2025; Shihab et al., 2026; Hou et al., 2025).

Definition 2. *Given two claims, a premise P and a hypothesis H, we define the following three directional natural language inference (NLI) relationships:*

(1) **Entailment** *($P \Rightarrow H$): P entails H if and only if H must be true given P.*

(2) **Contradiction** *($P \perp H$): P contradicts H if and only if P and H cannot both be true.*

(3) **Neutrality** *($P \approx H$): P is neutral with respect to H if and only if P provides insufficient information to establish either H or its negation.*

In what follows, we assume that every pair of claims satisfies exactly one of these three conditions. In Table 1, we provide concrete examples of the three NLI relationships introduced above.

Relationship	Premise	Hypothesis
Entailment	A dog is sleeping in the living room.	A dog is sleeping indoors.
Contradiction	A dog is sleeping in the living room.	A dog is running around the house.
Neutrality	A dog is sleeping in the living room.	A dog is dreaming of electric sheep.

Table 1: Examples of Natural Language Inference relationships.

Given a text $x = (c_1, \ldots, c_K)$, we define a $K \times K \times 3$ relationship tensor $\mathbf{M}$ that encodes all pairwise relationships between claims. Specifically, for any indices $i, j \in [K]$, we define

$$\mathbf{M}_{i,j,0} = \mathbb{1}\{c_i \text{ entails } c_j\}, \quad \mathbf{M}_{i,j,1} = \mathbb{1}\{c_i \text{ is neutral to } c_j\}, \quad \mathbf{M}_{i,j,2} = \mathbb{1}\{c_i \text{ contradicts } c_j\}.$$

Since exactly one condition is true for each pair, for all $i, j \in [K]$, we have

$$\mathbf{M}_{i,j,0} + \mathbf{M}_{i,j,1} + \mathbf{M}_{i,j,2} = 1.$$

Next, we let $\mathbf{w} \in \{0,1\}^K$ denote a logical state for the correctness of each of the claims, where $\mathbf{w}_i = 1$ if and only if the claim i is incorrect. Logical consistency imposes restrictions on $\mathbf{w}$: two claims i and j cannot both be correct when $i \perp j$, and a correct claim i cannot entail incorrect claim j. For each $i, j \in [K]$ with $i \neq j$, define

$$\Delta_{ij} := \mathbf{M}_{i,j,1} + \mathbf{M}_{i,j,2}\mathbb{1}\{(\mathbf{w}_i, \mathbf{w}_j) \neq (0,0)\} + \mathbf{M}_{i,j,0}\mathbb{1}\{(\mathbf{w}_i, \mathbf{w}_j) \neq (0,1)\},$$

and the state restrictor $T_{\mathbf{M}}(\mathbf{w}) := \prod_{i \neq j} \Delta_{ij}$. Then $T_{\mathbf{M}}(\mathbf{w}) = 1$ if $\mathbf{w}$ is logically consistent with the NLI relationships encoded by $\mathbf{M}$ and $T_{\mathbf{M}}(\mathbf{w}) = 0$ otherwise. We further let $\Omega_{\mathbf{M}} = \{\mathbf{w} \in \{0,1\}^K : T_{\mathbf{M}}(\mathbf{w}) = 1\}$ denote the set of logically consistent states.

Based on this idea, we use a working statistical model to derive our scores. This is not an assumption that we believe to be true, but rather an operational assumption that we use to derive our analysis. As already discussed, the final results do not depend on the choice of this working model. Specifically, we will derive our updated scores by considering that the claim states are independent Bernoulli random variables with success probabilities π_i (where we are implicitly working conditionally on all information used to construct π_i); thus we treat π_i *as if* they were correct. Then, the probability

of observing a configuration of states $\mathbf{w}$ is $\mathbb{P}(\mathbf{W} = \mathbf{w}) = \prod_{i=1}^{K} \pi_i^{w_i}(1-\pi_i)^{1-w_i}$. The vector of marginal probabilities, conditioning on logical feasibility, is:

$$\mathbb{E}(\mathbf{W} \mid \mathbf{W} \in \Omega_{\mathbf{M}}) = \frac{\sum_{\mathbf{w}\in\{0,1\}^K} \mathbb{P}(\mathbf{W}=\mathbf{w}) T_{\mathbf{M}}(\mathbf{w})\mathbf{w}}{\sum_{\mathbf{w}\in\{0,1\}^K} \mathbb{P}(\mathbf{W}=\mathbf{w}) T_{\mathbf{M}}(\mathbf{w})}.$$

The conditional requires knowledge the oracle $\mathbf{M}$, which is not available. In practice, we may estimate it by $\widehat{\mathbf{M}}$, where we instead condition on the information provided by the $\mathbb{P}_i$s and the NLI model data, encoded in the sigma-field $\mathcal{G} = \sigma(\pi_1, \cdots, \pi_K, \text{NLI model data})$. We then compute posterior expectations with respect to the distribution induced by the working probability model:

$$\widehat{\mathbf{M}}_{i,j,0} = \mathbb{E}(\mathbf{M}_{i,j,0} \mid \mathcal{G}), \quad \widehat{\mathbf{M}}_{i,j,1} = \mathbb{E}(\mathbf{M}_{i,j,1} \mid \mathcal{G}), \quad \widehat{\mathbf{M}}_{i,j,2} = \mathbb{E}(\mathbf{M}_{i,j,2} \mid \mathcal{G}).$$

Clearly, $\widehat{\mathbf{M}}_{i,j,0} + \widehat{\mathbf{M}}_{i,j,1} + \widehat{\mathbf{M}}_{i,j,2} = 1$. Given these probabilities, we can define a *pseudo state restrictor* $T_{\widehat{\mathbf{M}}}(\mathbf{w}) \in [0,1]$, which acts as a weighting function:

$$\mathbf{e}_i = \begin{pmatrix} \mathbb{1}\{\mathbf{w}_i = 0\} \\ \mathbb{1}\{\mathbf{w}_i = 1\} \end{pmatrix}, \quad \mathbf{A}_{ij} = \begin{pmatrix} 1-\widehat{\mathbf{M}}_{i,j,2} & 1-\widehat{\mathbf{M}}_{i,j,0} \\ 1 & 1 \end{pmatrix}, \quad T_{\widehat{\mathbf{M}}}(\mathbf{w}) = \prod_{i\neq j} \mathbf{e}_i^T \mathbf{A}_{ij} \mathbf{e}_j.$$

Thus, a state assignment receives a smaller weight when it conflicts with an NLI relationship to which the model assigns high probability. For example, if the estimated probability of contradiction $\widehat{\mathbf{M}}_{i,j,2}$ is close to one, then the configuration $(w_i, w_j) = (0,0)$ receives a weight close to zero. The product over all ordered pairs aggregates these local compatibility scores. Consequently, $T_{\widehat{\mathbf{M}}}(\mathbf{w})$ is large when the configuration $\mathbf{w}$ is broadly compatible with the estimated NLI relations and small otherwise.

We posit a working observation model under which the estimated NLI relationships $\widehat{\mathbf{M}}$ are generated conditionally on the latent claim states $\mathbf{W}$. Specifically, we consider the working model $\mathbb{P}(\widehat{\mathbf{M}} \mid \mathbf{W} = \mathbf{w}) \propto T_{\widehat{\mathbf{M}}}(\mathbf{w})$, where the proportionality constant may depend on $\widehat{\mathbf{M}}$ but not on $\mathbf{w}$. Thus, $T_{\widehat{\mathbf{M}}}(\mathbf{w})$ acts as a working likelihood[4] measuring the compatibility between the latent truth configuration $\mathbf{w}$ and the observed NLI information $\widehat{\mathbf{M}}$. Combining this working likelihood with the independent Bernoulli working prior on $\mathbf{W}$, Bayes' rule gives the posterior mean

$$\mathbf{q} \equiv \mathbb{E}(\mathbf{W} \mid \widehat{\mathbf{M}}) = \frac{\sum_{\mathbf{w}\in\{0,1\}^K} \mathbb{P}(\mathbf{W}=\mathbf{w}) T_{\widehat{\mathbf{M}}}(\mathbf{w})\mathbf{w}}{\sum_{\mathbf{w}\in\{0,1\}^K} \mathbb{P}(\mathbf{W}=\mathbf{w}) T_{\widehat{\mathbf{M}}}(\mathbf{w})}.$$

This posterior mean provides the chances that the claims are incorrect, given all the working model assumptions and observed information, and thus would be a reasonable choice for the scores. However, computing this posterior exactly requires summing over all 2^K configurations, resulting in computational complexity $O(K^2 2^K)$. Due to the high computational complexity, we next consider approximation methods. We first introduce a baseline approximation, followed by our proposed variational inference approach.

Baseline: star-structured approximation. A baseline inspired by Hou et al. (2025) is to impose the working model that a claim is directly affected only by its "neighbors"; formally in our model: for three claims $i, j, k \in [K], i \neq j, i \neq k, j \neq k$,

$$\begin{aligned} &\mathbb{P}(j \text{ incorrect}, k \text{ incorrect} \mid \text{claim } i \text{ incorrect}) \\ &\quad = \mathbb{P}(j \text{ incorrect} \mid i \text{ incorrect}) \times \mathbb{P}(k \text{ incorrect} \mid i \text{ incorrect}) \end{aligned} \tag{4}$$

Under the working model (4), we can derive an upper bound on the exact posterior probability of each claim, given in the following theorem, with a proof in Section A.4.

[4]Again we do not assume that T represents the correct data-generating model but rather we only use it as a device in order to derive a reasonable score function.

Theorem 2. *Assume the working model* (4) *and* $\mathbb{P}(\widehat{\mathbf{M}} \mid \mathbf{W} = \mathbf{w}) \propto T_{\widehat{\mathbf{M}}}(\mathbf{w})$. *For two claims* $i, j \in [K]$*, with* $i \neq j$*, define*

$$q^{star}_{i|j} = \mathbb{P}(i \text{ is incorrect} \mid \widehat{\mathbf{M}}_{i,j,\cdot}, \widehat{\mathbf{M}}_{j,i,\cdot}),$$

where $\pi_i = \mathbb{P}(i \text{ is incorrect})$ *denotes the marginal probability of incorrectness. We then have*

$$q_i \leq \left[1 + \left(\frac{\pi_i}{1-\pi_i}\right)^{K-2} \prod_{j \neq i} \left(\frac{1 - \mathbf{v}_{ij}^T U_{ij} \mathbf{v}_{ji}}{\mathbf{v}_{ij}^T U_{ij} \mathbf{v}_{ji}}\right)\right]^{-1} =: q_i^{star}.$$

The definition of U_{ij} *and* $\mathbf{v}$ *are provided in A.4.*

Remark 1. *While this procedure can work well in some cases, as our experiments on the WiCE (Kamoi et al., 2023) dataset show, it can struggle in more complex cases, which involve potential contradictions. For instance, consider the following claims:* ***John lives in Paris, John does not live in Paris, John lives in London****; assuming that John lives in Paris and sometimes visits London, so the ground truth is* $(0, 1, 0)$*. Suppose a model returns the scores* $(0.1, 0.9, 0.5)$*, which are reasonable (lower scores for correct claims). Our procedure will transform the scores to* $(0.414, 1, 0.870)$*, which is less reasonable (Details in Section A.5).*

3.2.2 Variational inference approximation

As an alternative, we use *belief propagation* (Yedidia et al., 2005) to approximate the posterior. Define $\mathbf{B}_{ij}$ to be a bidirectional 2×2 probability matrix, where each entry corresponds to probability assigned to an event when the first claim is right/wrong and second entry is right/wrong:

$$\mathbf{B}_{ij} = \begin{pmatrix} (1 - \widehat{\mathbf{M}}_{i,j,2})(1 - \widehat{\mathbf{M}}_{j,i,2}) & 1 - \widehat{\mathbf{M}}_{i,j,0} \\ 1 - \widehat{\mathbf{M}}_{j,i,0} & 1 \end{pmatrix}.$$

It turns out that the belief propagation algorithm, which constructs a sequence of update rules $m^r_{P_{i,j} \to i}$, $r = 0, 1, \ldots$, has the following form.

Lemma 1. *The Belief Propagation algorithm translates to the following update rule,*

$$m^{(r+1)}_{P_{i,j} \to i}(w_i) = \sum_{w_j \in \{0,1\}} \mathbf{e}_i^T \mathbf{B}_{i,j} \mathbf{e}_j \, \pi_j^{w_j} \, (1 - \pi_j)^{1 - w_j} \prod_{k \neq i,j} m^{(r)}_{P_{j,k} \to j}(w_j), \quad \text{for } w_i \in \{0, 1\}.$$

See the proof provided in Section A.3. Moreover, we also observe empirically that the theoretical fixed point approximations for belief propagation are quite accurate in our case. These take the form, for a large iteration number M,

$$\frac{\pi_i \prod_{j \neq i} m^{(M)}_{P_{i,j} \to i}(1)}{(1 - \pi_i) \prod_{j \neq i} m^{(M)}_{P_{i,j} \to i}(0) + \pi_i \prod_{j \neq i} m^{(M)}_{P_{i,j} \to i}(1)}.$$

4 Experiments

We conduct experiments on FELM (Zhao et al., 2023), WiCE (Kamoi et al., 2023), FABLES (Kim et al., 2024) datasets, and a human-annotated dataset we construct specifically for this problem. For each of the datasets, we report averages and standard errors across 50 train-test splits; see Section A.2 for details.

Our first dataset is FELM (also used in Rubin-Toles et al. (2025); Hou et al. (2025)). It consists of 189 records, which cover mathematics, reasoning, world knowledge, among others. The second dataset is FABLES (Kim et al., 2024), comprised of 130 book summaries generated by various LLMs (gpt-4, gpt-3.5-turbo, claude-3-opus, etc.) with each summary consisting of around 20 human-annotated claims. The third dataset is WiCE (Kamoi et al., 2023), which consists of 121 claims from Wikipedia articles.

We also generate our dataset using GPT-5.6-sol max by prompting it to extract claims from 125 small Wikipedia articles from (Foundation, 2023). We take $\sim$20 claims from each article and change them adversarially, so that we introduce entailment and contradictory relationships. We then manually assemble the text from claims, so it preserves both the human-like style and correctness.

4.1 RESULTS

We run ConFacTR using our scores, comparing with baseline scores from Hou et al. (2025) and Rubin-Toles et al. (2025), as well as those obtained from GPT-5-nano, with FNR target set at $\alpha = 0.25$ (due to space limitations, details are presented in Section A.6).

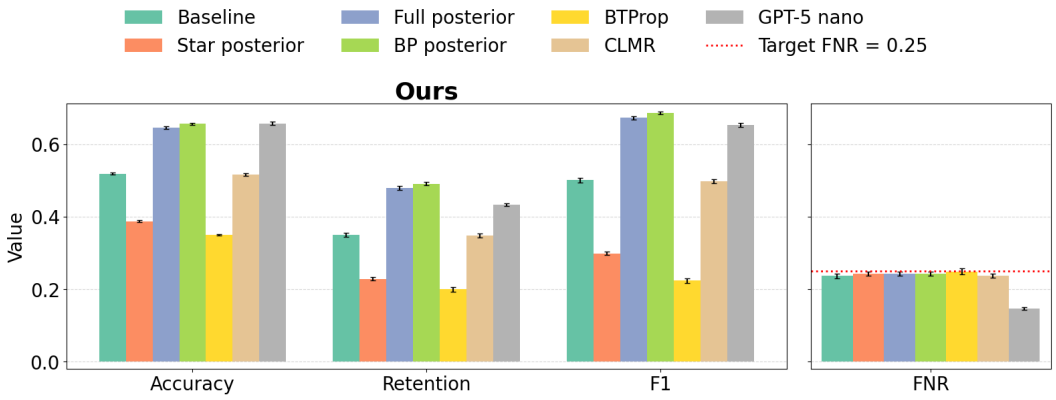


Figure 2: Visualization of performance of different methods applied to dataset generated by us.

We observe that all scores control the FNR, as expected. Posterior and Belief Propagation (BP) posterior scores perform similarly on all four metrics, outperforming GPT-5-nano on accuracy, retention and F1. Baseline scores are inferior to GPT-5-nano, indicating that our scores improve performance. As expected, on texts with a large number of claims, the performance of the star posterior scores can degrade. Moreover, the other datasets lead to similar conclusions (see Section A.6).

5 DISCUSSION

In this work, we have proposed a method for factuality retention with false negative rate control; consisting of a generic method as well as approaches to improve scores. We argued that our scores empirically improve performance.

Limitations. Many descriptive texts contain mostly atomic claims, which are neutral to the surrounding text. We expect our scores would perform similarly to the baseline scores. When the number of claims is large, it becomes computationally taxing to score all possible NLI relationships, which grow as $O(K^2)$, so further approximation and sampling ideas are needed to leverage our methods.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, implementing methods.

We have not used generative AI tools for formulating mathematical claims, cleaning and reformatting datasets, helping to develop theoretical models or conceptual frameworks, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments. Interpreting results and assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work.

Additionally, we used generative AI tools for creating or editing software code, summarizing or analyzing existing literature, sourcing/searching for information, and identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by two authors, mathematical proofs have been verified by three authors, synthetic data was manually checked by one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4685–4697, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1475. URL https://aclanthology.org/D19-1475/.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, volume 37, pp. 114812–114842. Curran Associates, Inc., 2024. doi: 10.52202/079017-3645. URL https://papers.nips.cc/paper_files/paper/2024/hash/d02ff1aeaa5c268dc34790dd1ad21526-Abstract-Conference.html.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. FacTool: Factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=hJkQL9VtWT.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Gal Elidan, Ian McGraw, and Daphne Koller. Residual belief propagation: Informed scheduling for asynchronous message passing. *arXiv preprint arXiv:1206.6837*, 2012.

Wikimedia Foundation. Wikimedia downloads, 2023. URL https://dumps.wikimedia.org.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454. URL https://aclanthology.org/2022.tacl-1.11/.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. ClaimBuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948, 2017. doi: 10.14778/3137765.3137815. URL https://doi.org/10.14778/3137765.3137815.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. A probabilistic framework for llm hallucination detection via belief tree propagation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3076–3099, 2025.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.395. URL https://aclanthology.org/2024.emnlp-main.395/.

Alexander T. Ihler, John W. III Fisher, Alan S. Willsky, and David Maxwell Chickering. Loopy belief propagation: convergence and effects of message errors. *Journal of Machine Learning Research 6*, pp. 905, 2005.

Tommi Sakari Jaakkola. *Variational methods for inference and estimation in graphical models*. PhD thesis, Massachusetts Institute of Technology, 1997.

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

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. Fables: Evaluating faithfulness and content selection in book-length summarization., 2024.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. Summac: Re-visiting nli-based models for inconsistency detection in summarization, 2021. URL https://arxiv.org/abs/2111.09525.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli, June 2022. URL https://osf.io/74b8k.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. Vera: A general-purpose plausibility estimation model for commonsense statements. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1264–1287, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.81. URL https://aclanthology.org/2023.emnlp-main.81/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Radu Marinescu, Debarun Bhattacharjya, Junkyu Lee, Tigran T Tchrakian, Javier Carnerero-Cano, Yufang Hou, Elizabeth M Daly, and Alessandra Pascale. Factreasoner: A probabilistic approach to long-form factuality assessment for large language models. In *EMNLP (Findings)*, pp. 14547–14577, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36029–36047. PMLR, 2024. URL https://proceedings.mlr.press/v235/mohri24a.html.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pp. 345–356. Springer, 2002.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Maxon Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, and Surbhi Goel. Conformal language model reasoning with coherent factuality, 2025. URL https://arxiv.org/abs/2505.17126.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *IJCAI*, 1999.

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. Cgric: Compositional risk certification for structured llm outputs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, 1999.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001. URL https://people.eecs.berkeley.edu/~jordan/papers/wainwright-jordan-fnt.pdf.

Ting Wang, Yuanjie Shi, Yan Yan, and Huan Zhang. Inference-time conformal reasoning with valid factuality control for large language models. *arXiv preprint arXiv:2606.08831*, 2026.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. DnDScore: Decontextualization and decomposition for factuality verification in long-form text generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23609–23626, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1205. URL https://aclanthology.org/2025.emnlp-main.1205/.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80756–80827. Curran Associates, Inc., 2024. doi: 10.52202/079017-2567. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/937ae0e83eb08d2cb8627fe1def8c751-Abstract-Conference.html.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, July 2005. doi: 10.1109/TIT.2005.850085. URL https://www.cs.princeton.edu/courses/archive/spr06/cos598C/papers/YedidaFreemanWeiss2004.pdf.

Yiran Zhao, Jinghan Zhang, I. Chern, Siyang Gao, Pengfei Liu, and Junxian He. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems 36*, pp. 44502–44523, 2023.

A Appendix

A.1 Proof of Theorem 1

The proof mostly follows from Angelopoulos et al. (2024, Theorem 1). We nevertheless present a self-contained proof for completeness. Throughout, we condition on $\mathcal{D}_{\text{train}}$ and therefore treat the scoring function s as fixed. By the joint exchangeability, $\mathcal{D}_{\text{calib}}$ and x_{N+1} are jointly exchangeable conditional on $\mathcal{D}_{\text{train}}$.

For any $\lambda \in [0,1]$ and $i \in [N+1]$, define

$$L_i(\lambda) = \text{FNP}(\mathcal{A}_{\lambda,s}, x_i).$$

Next, we define an "oracle" threshold

$$\tilde{\lambda} = \inf\left\{\lambda \in [0,1] : \frac{1}{N+1}\sum_{i=1}^{N+1} L_i(\lambda) \le \alpha\right\}.$$

Since $L_{N+1}(\lambda) \le 1$, we have $\tilde{\lambda} \le \hat{\lambda}$ by definition. As a result,

$$\text{FNR}(\mathcal{A}_{\hat{\lambda},s}, x_{N+1}) = \mathbb{E}\big[L_{N+1}(\hat{\lambda})\big] \le \mathbb{E}\big[L_{N+1}(\tilde{\lambda})\big],$$

where the last step is due to the monotonicity of $L_{N+1}(\lambda)$ in λ. Let E denote the order statistics of $L_1, \ldots, L_{N+1}$ and $\hat{\lambda}$ is deterministic given E. Therefore, by exchangeability,

$$\mathbb{E}[L_{N+1}(\tilde{\lambda}) \mid E] = \frac{1}{N+1}\sum_{i=1}^{N+1} \mathbb{E}[L_i(\tilde{\lambda}) \mid E] = \mathbb{E}\Big[\frac{1}{N+1}\sum_{i=1}^{N+1} L_i(\tilde{\lambda}) \,\Big|\, E\Big] \le \alpha,$$

where the last step follows from the definition of $\tilde{\lambda}$ and the right continuity of L_i's. Taking expectation over E completes the proof.

A.2 Data Generation

Wikipedia passages. We use source passages from the following Hugging Face dataset: Foundation (2023). This dataset contains Wikipedia passages in different languages that are extracted from official Wikipedia webpages from 2023. We focus on the "20231101.en" subset which consists of English passages. We randomly sampled 1000 passages from the subset and downloaded the first two paragraphs of each passage. We treat the downloaded passages as factual texts.

Injection by prompting LLMs. We feed the 1000 downloaded passages and a prompt to an LLM (in our case, ChatGPT-5.6 Sol) to create a ready-to-use dataset of 125 texts. The prompt does the following to each passage:

1. Determines if the content of the passage is appropriate. This step is necessary because there are Wikipedia passages which consist of only dates and person names. We only need to select 125 texts out of 1000 passages, so an LLM filters out the outliers.
2. Compresses the passage to $15 - 20$ claims.
3. Decomposes each passage into atomic claims and assigns labels of factuality to each claim.
4. Modifies the claims, such that there is at least one entailment and one contradiction among the claims. It also ensures that there is a certain number of incorrect claims and assigns labels to claims accordingly.

The exact prompt is included in the GitHub repository.

Manual Validation. After creating the injection, we manually validate it for accuracy with respect to the original Wikipedia passage. Then, we manually alter the text to ensure that it resembles human-generated one. Although we only processed 125 texts, the size of the dataset is comparable to FELM and FABLES, which all have about 100 texts.

A.3 Details on Belief Propagation

In our proof we will follow Wainwright & Jordan (2008) and Marinescu et al. (2025). Assume a graphical model of K nodes. Let the graph be $G = (V, E)$ where $V = \{w_1, \cdots, w_K\}$. Each node w_i is a random variable with state space $\{0, 1\}$. We suppose that the joint distribution is of the following form:

$$p(w_1, \cdots, w_K) = \frac{1}{Z} \prod_C f_C(w_C).$$

where Z is the normalization constant, $C \subseteq [K]$ is a subset of nodes and w_C is a tuple of all w_i with $i \in C$. We call each C a factor. To compute the final distribution, we define the following factors of our model:

$$f_{U_i}(w_i) = \pi_i^{w_i}(1-\pi_i)^{1-w_i}, \quad f_{P_{i,j}}(w_i, w_j) = \mathbf{e}_i^T \mathbf{B}_{i,j} \mathbf{e}_j.$$

By following the definitions of messages and message sending, we have that

$$m_{U_i \to i}(w_i) = f_{U_i}(w_i) = \pi_i^{w_i}(1-\pi_i)^{1-w_i}.$$

$$\begin{aligned} m_{P_{i,j}\to i}(w_i) &= \sum_{w_j \in \{0,1\}} f_{P_{i,j}}(w_i, w_j)\, m_{j \to P_{i,j}}(w_j), \\ &= \sum_{w_j \in \{0,1\}} f_{P_{i,j}}(w_i, w_j)\, m_{U_j \to j}(w_j) \prod_{k \neq i,j} m_{P_{k,j}\to j}(w_j), \\ &= \sum_{w_j \in \{0,1\}} \mathbf{e}_i^T \mathbf{B}_{i,j} \mathbf{e}_j\, \pi_j^{w_j}(1-\pi_j)^{1-w_j} \prod_{k \neq i,j} m_{P_{k,j}\to j}(w_j). \end{aligned}$$

This would define the update rule. Let's put the superscript (r) to denote the iteration number.

$$m^{(r+1)}_{P_{i,j}\to i}(w_i) = \sum_{w_j \in \{0,1\}} \mathbf{e}_i^T \mathbf{B}_{i,j} \mathbf{e}_j\, \pi_j^{w_j}(1-\pi_j)^{1-w_j} \prod_{k \neq i,j} m^{(r)}_{P_{j,k}\to j}(w_j).$$

After they converge, we then compute that

$$b_i(w_i) \propto m_{U_i \to i}(w_i) \times \prod_{j \neq i} m_{P_{i,j}\to i}(w_i) \propto \pi_i^{w_i}(1-\pi_i)^{1-w_i} \prod_{j \neq i} m_{P_{i,j}\to i}(w_i).$$

Transferring it to probability, we have that the aggregate score vector is $\tilde{\mathbf{q}}$ with each element being:

$$\tilde{q}_i = \frac{\pi_i \prod_{j \neq i} m_{P_{i,j}\to i}(1)}{(1-\pi_i)\prod_{j \neq i} m_{P_{i,j}\to i}(0) + \pi_i \prod_{j \neq i} m_{P_{i,j}\to i}(1)}.$$

A.4 Proof for Star Scores

Upper bound for a star score We will prove that there's an upper bound $q^{\text{star}}_{A|B} \leq \mathbf{v}_{AB}^T U_{AB} \mathbf{v}_{BA}$, where

$$\mathbf{v}_{AB} := \begin{pmatrix} \mathbb{P}(A \perp B \mid \text{data}) \\ \mathbb{P}(A \approx B \mid \text{data}) \\ \mathbb{P}(A \Rightarrow B \mid \text{data}) \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{M}}_{A,B,2} \\ \widehat{\mathbf{M}}_{A,B,1} \\ \widehat{\mathbf{M}}_{A,B,0} \end{pmatrix}$$

and U_{AB} is a linear operator. We will consider 9 different computations to compute the upper bound on the posterior probability, using basic set theory properties.

1. $\mathbf{M}_{B,A,2} = 1\,(B \perp A), \mathbf{M}_{A,B,0} = 1\,(A \Rightarrow B)$. It is easy to verify that $q^{\text{star}}_{A|B} = 1$.
2. $\mathbf{M}_{B,A,1} = 1\,(B \approx A), \mathbf{M}_{A,B,1} = 1\,(A \approx B)$. It is also not difficult to show $q^{\text{star}}_{A|B} = \pi_A$.
3. $\mathbf{M}_{B,A,0} = 1\,(B \Rightarrow A), \mathbf{M}_{A,B,0} = 1\,(A \Rightarrow B)$. By Bayes' rule the score is going to be:

$$q^{\text{star}}_{A|B} = \frac{\mathbb{P}(\text{A is incorrect}, \text{B is incorrect})}{\mathbb{P}(\text{A is correct}, \text{B is correct}) + \mathbb{P}(\text{A is incorrect}, \text{B is incorrect})}.$$

We can note that $1 - \pi_A - \pi_B + \max(0, \pi_A + \pi_B - 1) \leq \mathbb{P}(\text{A is correct, B is correct}) \leq 1 - \pi_A - \pi_B + \min(\pi_A, \pi_B)$, so by taking the derivative with respect to $s = \mathbb{P}(\text{A is incorrect, B is incorrect})$ of $q^{\text{star}}_{A|B} = \frac{s}{1-\pi_A-\pi_B+2s}$ in three cases $\pi_A + \pi_B \lessgtr 1$ (could be verified through basic algebra), we get an upper bound on the probability, which gives us most conservative pairwise score:

$$q^{\text{star}}_{A|B} \leq \begin{cases} 1, & \pi_A + \pi_B > 1. \\ 1/2, & \pi_A + \pi_B = 1. \\ \frac{\min(\pi_A, \pi_B)}{1-\pi_A-\pi_B+2\min(\pi_A,\pi_B)}, & \pi_A + \pi_B < 1. \end{cases}$$

4. $\mathbf{M}_{B,A,0} = 1\ (B \Rightarrow A), \mathbf{M}_{A,B,1} = 1\ (A \approx B)$. Using the information given:

$$q^{\text{star}}_{A|B} = \frac{\mathbb{P}(\text{A is incorrect, B is incorrect})}{1 - \mathbb{P}(\text{A is incorrect, B is correct})}.$$

We know that $\max(0, \pi_A + \pi_B - 1) \leq \mathbb{P}(\text{A is incorrect, B is incorrect}) \leq \min(\pi_A, \pi_B)$, and $\mathbb{P}(\text{A is incorrect, B is correct}) = \pi_A - \mathbb{P}(\text{A is incorrect, B is incorrect})$. Since the function is increasing w.r.t. $\mathbb{P}(\text{A is incorrect, B is incorrect})$, we take the upper bound:

$$q^{\text{star}}_{A|B} \leq \frac{\min(\pi_A, \pi_B)}{1 - \pi_A + \min(\pi_A, \pi_B)}.$$

5. $\mathbf{M}_{B,A,1} = 1\ (B \approx A), \mathbf{M}_{A,B,0} = 1\ (A \Rightarrow B)$. Using the information given:

$$q^{\text{star}}_{A|B} = \frac{\mathbb{P}(\text{A is incorrect})}{1 - \mathbb{P}(\text{A is correct, B is incorrect})}.$$

We know that if $s = \mathbb{P}(\text{A is incorrect, B is incorrect})$, then $\max(0, \pi_A + \pi_B - 1) \leq s \leq \min(\pi_A, \pi_B)$, and $\mathbb{P}(\text{A is correct, B is incorrect}) = \pi_B - s$. Since the function is decreasing w.r.t. s, we take the lower bound $s = \max(0, \pi_A + \pi_B - 1)$:

$$q^{\text{star}}_{A|B} \leq \begin{cases} 1, \pi_A + \pi_B > 1. \\ \frac{\pi_A}{1-\pi_B}, \pi_A + \pi_B \leq 1. \end{cases}$$

6. Cases $\mathbf{M}_{B,A,1} = 1\ (B \approx A), \mathbf{M}_{A,B,2} = 1\ (A \perp B)$, $\mathbf{M}_{B,A,2} = 1\ (B \perp A), \mathbf{M}_{A,B,1} = 1\ (A \approx B)$, and $\mathbf{M}_{B,A,2} = 1\ (B \perp A), \mathbf{M}_{A,B,2} = 1\ (A \perp B)$ all lead to the same result. Using given information, we get:

$$q^{\text{star}}_{A|B} = \frac{\mathbb{P}(\text{A is incorrect})}{1 - \mathbb{P}(\text{A is correct, B is correct})}.$$

We know that $1 - \pi_A - \pi_B + \max(0, \pi_A + \pi_B - 1) \leq \mathbb{P}(\text{A is correct, B is correct}) \leq 1 - \pi_A - \pi_B + \min(\pi_A, \pi_B)$. Thus, we will get:

$$q^{\text{star}}_{A|B} \leq \frac{\pi_A}{\pi_A + \pi_B - \min(\pi_A, \pi_B)} = \frac{\pi_A}{\max(\pi_A, \pi_B)}.$$

7. $\mathbf{M}_{B,A,0} = 1\ (B \Rightarrow A), \mathbf{M}_{A,B,2} = 1\ (A \perp B)$. In this case, we have:

$$q^{\text{star}}_{A|B} = \frac{\mathbb{P}(\text{A is incorrect, B is incorrect})}{\mathbb{P}(\text{B is incorrect})} = \frac{s}{\pi_B} \leq \frac{\min(\pi_A, \pi_B)}{\pi_B}.$$

Note that we can summarize all these computations in the form of a linear operator U_{AB}, which we then are going to use in computation of an exact posterior upper bound:

	$\mathbf{M}_{B,A,2} = 1$	$\mathbf{M}_{B,A,1} = 1$	$\mathbf{M}_{B,A,0} = 1$
$\mathbf{M}_{A,B,2} = 1$	$\frac{\pi_A}{\max(\pi_A, \pi_B)}$	$\frac{\pi_A}{\max(\pi_A, \pi_B)}$	$\frac{\min(\pi_A, \pi_B)}{\pi_B}$
$\mathbf{M}_{A,B,1} = 1$	$\frac{\pi_A}{\max(\pi_A, \pi_B)}$	π_A	$\frac{\min(\pi_A, \pi_B)}{1 - \pi_A + \min(\pi_A, \pi_B)}$
$\mathbf{M}_{A,B,0} = 1$	1	$\begin{cases} 1, & \pi_A + \pi_B > 1, \\ \frac{\pi_A}{1-\pi_B}, & \pi_A + \pi_B \leq 1 \end{cases}$	$\begin{cases} 1, & \pi_A + \pi_B > 1, \\ \frac{1}{2}, & \pi_A + \pi_B = 1, \\ \frac{\min(\pi_A, \pi_B)}{1-\pi_A-\pi_B+2\min(\pi_A,\pi_B)}, & \pi_A + \pi_B < 1 \end{cases}$

Exact posterior We assume that graph is a star, ignoring any relationships beyond the node i and then use the Bayesian update on the scores. Before doing so, we should remind ourselves of an object we have previously defined, specifically $q^{\text{star}}_{i|j} = \mathbb{P}(i \text{ is wrong } \mid \widehat{\mathbf{M}}_{i,j,0}, \widehat{\mathbf{M}}_{i,j,1}, \widehat{\mathbf{M}}_{i,j,2}, \widehat{\mathbf{M}}_{j,i,0}, \widehat{\mathbf{M}}_{j,i,1}, \widehat{\mathbf{M}}_{j,i,2})$. Although we don't know it exactly, we know its upper bound, which we will use as its value:

$$q^{\text{star}}_{i|j} \leq \begin{pmatrix} \widehat{\mathbf{M}}_{i,j,2} & \widehat{\mathbf{M}}_{i,j,1} & \widehat{\mathbf{M}}_{i,j,0} \end{pmatrix} U_{ij} \begin{pmatrix} \widehat{\mathbf{M}}_{j,i,2} \\ \widehat{\mathbf{M}}_{j,i,1} \\ \widehat{\mathbf{M}}_{j,i,0} \end{pmatrix}.$$

Note that under the star model, we have *conditional independence* of all nodes given the node i or more formally, we consider the tuple of evidence $(\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.})$ for each j in the neighborhood. Then, we know that the joint distribution has to be the product of marginals where w_i is the correctness of claim i, such that $w_i = 1$ means that the claim is incorrect:

$$\mathbb{P}((\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.})_{j \neq i} | w_i = w) = \prod_{j \neq i} \mathbb{P}((\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.}) | w_i = w), \quad w \in \{0, 1\}.$$

Now, let's apply the Bayes rule, knowing that we always have a fully connected graph with K nodes:

$$\begin{aligned} \mathbb{P}(w_i = w \mid \widehat{\mathbf{M}}) &\propto \mathbb{P}(w_i = w) \mathbb{P}((\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.})_{j \neq i} \mid w_i = w), \\ &\propto \mathbb{P}(w_i = w) \prod_{j \neq i} \mathbb{P}((\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.}) \mid w_i = w), \\ &\propto \frac{\prod_{j \neq i} \mathbb{P}(w_i = w \mid (\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.}))}{\mathbb{P}(w_i = w)^{K-2}} \prod_{j \neq i} \mathbb{P}((\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.})). \end{aligned}$$

At the same time, a more trivial result always holds true, as $w_i \in \{0, 1\}$:

$$\mathbb{P}(w_i = 0 \mid \widehat{\mathbf{M}}) + \mathbb{P}(w_i = 1 \mid \widehat{\mathbf{M}}) = 1.$$

Thus, we can conclude that

$$\begin{aligned} q^{\text{star}}_i &= \mathbb{P}(w_i = 1 \mid \widehat{\mathbf{M}}), \\ &= \frac{\mathbb{P}(w_i = 1 \mid \widehat{\mathbf{M}})}{\mathbb{P}(w_i = 0 \mid \widehat{\mathbf{M}}) + \mathbb{P}(w_i = 1 \mid \widehat{\mathbf{M}})}, \\ &= \frac{\frac{\prod_{j \neq i} \mathbb{P}(w_i = 1 | (\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.}))}{\mathbb{P}(w_i = 1)^{K-2}}}{\frac{\prod_{j \neq i} \mathbb{P}(w_i = 0 | (\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.}))}{\mathbb{P}(w_i = 0)^{K-2}} + \frac{\prod_{j \neq i} \mathbb{P}(w_i = 1 | (\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.}))}{\mathbb{P}(w_i = 1)^{K-2}}}. \end{aligned}$$

Note that we have already computed the probabilities necessary to evaluate this formula (as per assumption of an upper bound):

$$\mathbb{P}(w_i = 1) = \pi_i, \quad \mathbb{P}(w_i = 1 \mid (\widehat{\mathbf{M}}_{i,j,.}, \widehat{\mathbf{M}}_{j,i,.})) = q^{\text{star}}_{i|j}.$$

Thus, we get the following result for the score:

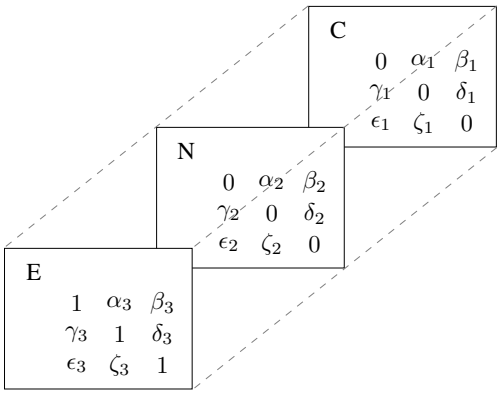
$$q^{\text{star}}_i = \frac{\frac{\prod_{j \neq i} q^{\text{star}}_{i|j}}{\pi_i^{K-2}}}{\frac{\prod_{j \neq i} (1 - q^{\text{star}}_{i|j})}{(1 - \pi_i)^{K-2}} + \frac{\prod_{j \neq i} q^{\text{star}}_{i|j}}{\pi_i^{K-2}}} = \left(1 + \left(\frac{\pi_i}{1 - \pi_i}\right)^{K-2} \prod_{j \neq i} \left(\frac{1 - q^{\text{star}}_{i|j}}{q^{\text{star}}_{i|j}}\right)\right)^{-1}.$$

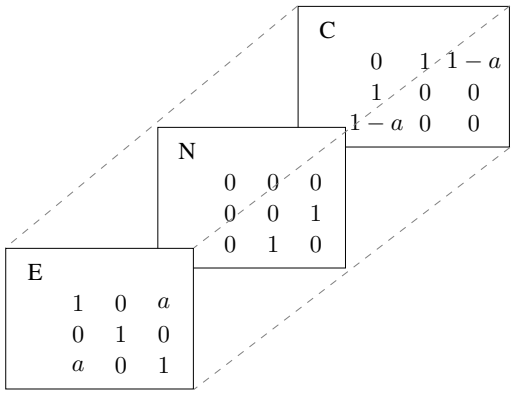
Which could also be rewritten as:

$$q^{\text{star}}_i = \left[1 + \left(\frac{\pi_i}{1 - \pi_i}\right)^{K-2} \prod_{j \neq i} \left(\frac{1 - \mathbf{v}_{ij}^T U_{ij} \mathbf{v}_{ji}}{\mathbf{v}_{ij}^T U_{ij} \mathbf{v}_{ji}}\right)\right]^{-1}$$

A.5 PROOF FOR THE COUNTEREXAMPLE

In this section we define both a general form of $\widehat{\mathbf{M}}$ and a specific one we optimize. We also assume that the score triplet that we take in is $(\pi_1, \pi_2, \pi_3) = (0.1, 0.9, 0.5)$:





By plugging in the values into U_{AB}, we get the following set of linear operators:

$$U_{12} = \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/10 & 1/10 \\ 1 & 1 & 1/2 \end{pmatrix}, \qquad U_{13} = \begin{pmatrix} 1/5 & 1/5 & 1/5 \\ 1/5 & 1/10 & 1/10 \\ 1 & 1/5 & 1/6 \end{pmatrix}, \qquad U_{21} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 9/10 & 1/2 \\ 1 & 1 & 1/2 \end{pmatrix}.$$

$$U_{23} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 9/10 & 5/6 \\ 1 & 1 & 1 \end{pmatrix}, \qquad U_{32} = \begin{pmatrix} 5/9 & 5/9 & 5/9 \\ 5/9 & 1/2 & 1/2 \\ 1 & 1 & 1 \end{pmatrix}, \qquad U_{31} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1/2 & 1/6 \\ 1 & 5/9 & 1/6 \end{pmatrix}.$$

Then, let's compute all the conditional scores:

$$q^{\text{star}}_{1|2} = \alpha^T U_{12} \gamma, q^{\text{star}}_{1|3} = \beta^T U_{13} \epsilon, \quad q^{\text{star}}_{2|1} = \gamma^T U_{21} \alpha, \quad q^{\text{star}}_{2|3} = \delta^T U_{23} \zeta, \quad q^{\text{star}}_{3|1} = \epsilon^T U_{31} \beta, \quad q^{\text{star}}_{3|2} = \zeta^T U_{32} \delta.$$

One pathway to obtain our goal is to solve the following optimization problem:

$$\begin{aligned} \min_{\alpha, \cdots, \zeta, q_1^{\text{star}}, q_2^{\text{star}}, q_3^{\text{star}}} \quad & (1 - q_1^{\text{star}})^2 + (q_2^{\text{star}})^2 + (1 - q_3^{\text{star}})^2 \\ \text{s.t.} \quad & q_1^{\text{star}} = \frac{1}{1 + \left(\frac{\pi_1}{1-\pi_1}\right)\left(\frac{1-q^{\text{star}}_{1|2}}{q^{\text{star}}_{1|2}}\right)\left(\frac{1-q^{\text{star}}_{1|3}}{q^{\text{star}}_{1|3}}\right)}, \\ & q_2^{\text{star}} = \frac{1}{1 + \left(\frac{\pi_2}{1-\pi_2}\right)\left(\frac{1-q^{\text{star}}_{2|1}}{q^{\text{star}}_{2|1}}\right)\left(\frac{1-q^{\text{star}}_{2|3}}{q^{\text{star}}_{2|3}}\right)}, \\ & q_3^{\text{star}} = \frac{1}{1 + \left(\frac{\pi_3}{1-\pi_3}\right)\left(\frac{1-q^{\text{star}}_{3|1}}{q^{\text{star}}_{3|1}}\right)\left(\frac{1-q^{\text{star}}_{3|2}}{q^{\text{star}}_{3|2}}\right)}, \\ & x^\top \mathbf{1} = 1, \quad x \geq \mathbf{0} \qquad \forall x \in \{\alpha, \beta, \gamma, \delta, \epsilon, \zeta\}. \end{aligned}$$

Using a specific matrix we defined at the start with one parameter a, we can reformulate the general problem into a parametric one. Then:

$$q_1^{\text{star}} = \frac{9(-25a^2 + 24a + 6)}{-25a^2 + 24a + 246}, \quad q_2^{\text{star}} = 1, \quad q_3^{\text{star}} = 1 - \frac{5}{6}a^2.$$

Thus, we need to minimize the following function with respect to a:

$$\underset{a \in [0,1]}{\text{minimize}} \left(1 - \frac{9(-25a^2 + 24a + 6)}{-25a^2 + 24a + 246}\right)^2 + \left(\frac{5}{6}a^2\right)^2.$$

Solving it numerically, we get $a = 0.3946$. Thus, $q_1^{\text{star}} = 0.414$ and $q_3^{\text{star}} = 0.870$ and our operation transformed $(0.1, 0.9, 0.5) \mapsto (0.414, 1, 0.870)$ with the ground truth being $(0, 1, 0)$. For example,

if we make a reasonable assumption that $\hat{\lambda} = 0.6$, the rejection set would be defined as $\mathcal{C} = \{c_j : \pi_j \geq 1 - \hat{\lambda}\}$. Then, before the update, we would reject an incorrect claim 2 with score 0.9 and correct claim with score 0.5. After the update, we reject all claims. Thus, before the update, our accuracy was 66% and after it is 33%.

A.6 Details on the experiment

Metrics We consider four different metrics for evaluating the performance of all models. Here is the comprehensive list with definitions, assuming we adopt the notation of $\mathcal{H}_{0,i}$ corresponding to the set of correct claims in the text x_i:

$$\text{precision} = \frac{|([K] \setminus \mathcal{A}_{\widehat{\lambda},s}(x_i)) \cap \mathcal{H}_{0,i}|}{|[K] \setminus \mathcal{A}_{\widehat{\lambda},s}(x_i)| \vee 1}, \quad \text{recall} = \frac{|([K] \setminus \mathcal{A}_{\widehat{\lambda},s}(x_i)) \cap \mathcal{H}_{0,i}|}{|\mathcal{H}_{0,i}| \vee 1}.$$

Metric	Formula	Source
F1	$2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$	(Marinescu et al., 2025)
Retention	$\frac{\lvert[K] \setminus \mathcal{A}_{\widehat{\lambda},s}(x_i)\rvert}{\lvert\phi_K(x_i)\rvert}$	(Rubin-Toles et al., 2025)
Accuracy	$\frac{\left\lvert\mathcal{A}_{\widehat{\lambda},s}(x_i) \cap \mathcal{H}_{1,i}\right\rvert + \left\lvert([K] \setminus \mathcal{A}_{\widehat{\lambda},s}(x_i)) \cap \mathcal{H}_{0,i}\right\rvert}{\lvert\phi_K(x_i)\rvert}$	(Hou et al., 2025)
FNR	$\frac{\left\lvert\mathcal{H}_{1,i} \setminus \mathcal{A}_{\widehat{\lambda},s}(x_i)\right\rvert}{\lvert\mathcal{H}_{1,i}\rvert \vee 1}$	(Angelopoulos et al., 2024)

Table 2: Evaluation metrics used in our experiments.

Dataset experiment details In this section we provide a brief description of how simulations have been conducted. Note that we start with seed number 42 and then each consecutive experiment takes a consecutive number as a seed:

Dataset	Train	Test	# of seeds
FELM	150	39	50
FABLES	100	30	50
WiCE	91	30	50
Ours	100	25	50

After running each experiment on 50 different splits of the data, we calculate the standard error (SE) and mean, which are both reported in the table (as mean $\pm$ SE) and the graph (SE is reported as confidence bounds around the bar).

Correctness scorer details As a scorer of claim correctness we chose out of the box Liu et al. (2023) scorer. It is based on FLAN-T5 (Chung et al., 2024) open-source model.

NLI scorer details For scoring NLI relations between claims we chose Laurer et al. (2022) DeBERTa-v3 scorer.

Decomposer details Since for all publicly available datasets (Zhao et al., 2023; Kamoi et al., 2023; Kim et al., 2024) claim decomposition is already provided, we only needed to perform claim splitting when working with our dataset. We used GPT-5.6-sol max to perform claim splitting given only the original text and a prompt, which is provided on GitHub.

Ablation studies We have considered two ablation studies, following procedures in Hou et al. (2025), Rubin-Toles et al. (2025). While it was necessary to alter the former one, we have implemented the latter one without any alterations.

Belief Tree Propagation While the original BTProp considers claim decomposition and then prompting a language model to generate child nodes, we cannot access outside knowledge by using a language model. Thus, we consider an ablation study, where we state that for a claim $c_j, j \in [K]$, claims $c_i, \forall i \in [K] \setminus \{j\}$, are child nodes for an update and assume that every child node is a leaf.

Experiment results While in the main paper we provide the graphical illustration of model's performance, here we show the detailed output after creating $N = 50$ train/test splits.

Method	Acc. ↑	Ret. ↑	FNR ↓	F1 ↑
Ours				
Baseline	0.5189 ± 0.0035	0.3500 ± 0.0061	0.2365 ± 0.0061	0.5011 ± 0.0065
Star posterior	0.3875 ± 0.0022	0.2289 ± 0.0050	0.2430 ± 0.0058	0.2989 ± 0.0052
Full posterior	0.6447 ± 0.0037	0.4792 ± 0.0050	0.2430 ± 0.0054	0.6724 ± 0.0048
BP posterior	$\mathbf{0.6557 \pm 0.0031}$	$\mathbf{0.4912 \pm 0.0046}$	0.2432 ± 0.0056	$\mathbf{0.6865 \pm 0.0040}$
BTProp	0.3501 ± 0.0026	0.1999 ± 0.0068	0.2490 ± 0.0086	0.2242 ± 0.0065
CLMR	0.5157 ± 0.0033	0.3475 ± 0.0058	0.2376 ± 0.0061	0.4978 ± 0.0062
GPT-5 nano	0.6567 ± 0.0046	0.4332 ± 0.0043	$\mathbf{0.1467 \pm 0.0039}$	0.6528 ± 0.0060
FELM				
Baseline	0.5425 ± 0.0070	0.2860 ± 0.0091	0.2503 ± 0.0119	0.2672 ± 0.0089
Star posterior	0.5467 ± 0.0063	0.2457 ± 0.0079	0.2381 ± 0.0105	0.2312 ± 0.0081
Full posterior	$\mathbf{0.5608 \pm 0.0062}$	$\mathbf{0.3248 \pm 0.0090}$	0.2565 ± 0.0111	$\mathbf{0.3145 \pm 0.0078}$
BP posterior	0.5590 ± 0.0061	0.3241 ± 0.0092	0.2580 ± 0.0113	0.3089 ± 0.0079
BTProp	0.5171 ± 0.0056	0.2319 ± 0.0080	0.2413 ± 0.0099	0.1960 ± 0.0077
CLMR	0.5501 ± 0.0073	0.2847 ± 0.0090	0.2468 ± 0.0120	0.2732 ± 0.0089
GPT-5 nano	0.5217 ± 0.0071	0.000 ± 0.000	$\mathbf{0.000 \pm 0.000}$	0.000 ± 0.000
FABLES				
Baseline	0.4158 ± 0.0059	0.3866 ± 0.0071	0.2449 ± 0.0091	0.5183 ± 0.0066
Star posterior	0.3653 ± 0.0048	0.3312 ± 0.0061	0.2508 ± 0.0107	0.4560 ± 0.0061
Full posterior	$\mathbf{0.5036 \pm 0.0051}$	$\mathbf{0.4880 \pm 0.0065}$	0.2339 ± 0.0097	$\mathbf{0.6128 \pm 0.0053}$
BP posterior	0.4990 ± 0.0053	0.4797 ± 0.0071	0.2351 ± 0.0095	0.6066 ± 0.0058
BTProp	0.3764 ± 0.0062	0.3385 ± 0.0079	0.2640 ± 0.0114	0.4327 ± 0.0082
CLMR	0.4089 ± 0.0057	0.3794 ± 0.0068	0.2456 ± 0.0088	0.5105 ± 0.0065
GPT-5 nano	0.2883 ± 0.0040	0.2403 ± 0.0038	$\mathbf{0.1252 \pm 0.0063}$	0.3474 ± 0.0050
WiCE				
Baseline	0.5426 ± 0.0061	0.2023 ± 0.0080	0.2370 ± 0.0094	0.1200 ± 0.0061
Star posterior	0.5360 ± 0.0061	0.1918 ± 0.0079	0.2365 ± 0.0094	0.1049 ± 0.0061
Full posterior	0.5331 ± 0.0061	$\mathbf{0.2093 \pm 0.0074}$	0.2437 ± 0.0093	$\mathbf{0.1192 \pm 0.0057}$
BP posterior	0.5331 ± 0.0061	$\mathbf{0.2093 \pm 0.0074}$	0.2437 ± 0.0093	$\mathbf{0.1192 \pm 0.0057}$
BTProp	$\mathbf{0.5894 \pm 0.0063}$	0.1413 ± 0.0057	0.1408 ± 0.0054	0.1125 ± 0.0072
CLMR	0.5387 ± 0.0061	0.1988 ± 0.0080	0.2376 ± 0.0093	0.1119 ± 0.0063
GPT-5 nano	0.5957 ± 0.0057	0.0000 ± 0.0000	$\mathbf{0.0000 \pm 0.0000}$	0.0000 ± 0.0000

Table 3: Results across our datasets. Values are reported as mean ± standard error.

We also provide a visualization of the results:

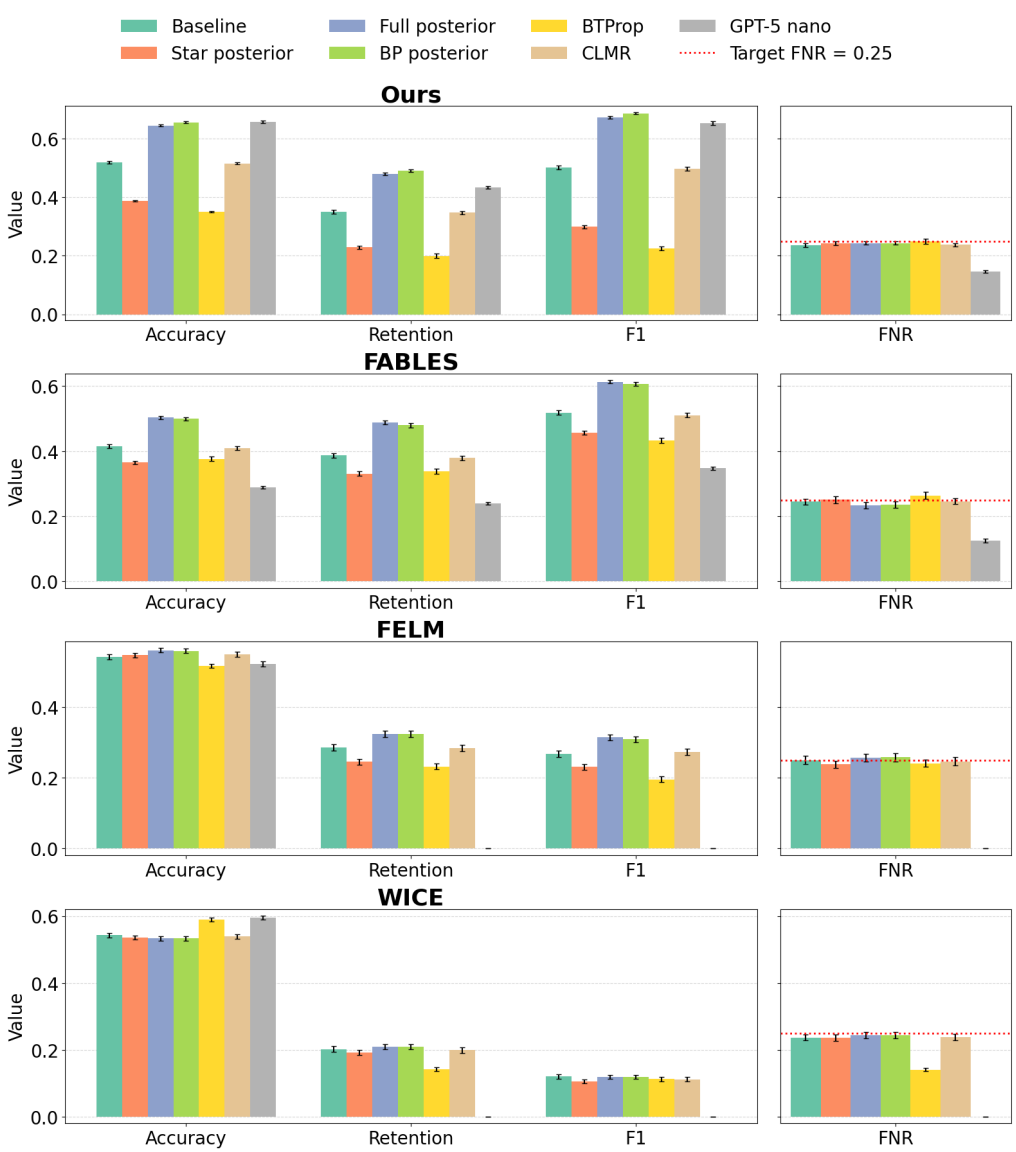


Figure 3: Visualization of performance of different methods we applied to different datasets.

Since our dataset has been previously discussed, we will focus on the remaining three. On FABLES, both BP and Full posterior significantly improve the performance of the baseline, dominating all other methods on all metrics. On FELM, the improvement is not as dramatic, but we can see that both BP and Full Posterior are the best methods. It's worth noting that in both cases GPT-5-nano performs worse than the baseline scores on accuracy and, on FELM, gpt-5-nano rejects every single claim, leading to 0 F1 and Retention. Due to the small number of claims, star posterior performs almost as well as full posterior on WiCE. Moreover, the BTProp method, which is based on tree belief propagation, performs better than all other methods except gpt-5-nano, potentially indicating that assumption 4 holds on this dataset.