BELIEF LEDGERS: POSTERIOR ACCOUNTING FOR AUDITABLE, FAITHFUL EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable explanations for evidence-driven NLP systems should reflect how a prediction is formed, not merely identify plausible evidence. We introduce **Belief Ledgers**, a probabilistic framework that turns retrieved passages, extracted claims, and numerical comparisons into an auditable sequence of posterior updates. Belief Ledgers group dependent evidence into clusters, compute cluster-level log Bayes factors for binary decisions, and use a separate conjugate Gaussian model for continuous margins, ensuring that all displayed contributions remain in a consistent accounting space. The resulting explanation contains provenance-linked evidence entries that exactly sum to the model's posterior log-odds and supports controlled counterfactual interventions. Across scientific claim verification, multi-sentence reading comprehension, and fact verification, Belief Ledgers improve predictive performance while substantially improving calibration. On SciProp, they achieve 75.8% accuracy and 0.034 ECE, versus 74.2% and 0.120 ECE for a strong NLI-aggregation baseline. Matched deletion/insertion tests and ablations further show that ledger-ranked evidence is causally predictive of model confidence and that dependence-aware clustering is critical. Belief Ledgers offer a practical path to calibrated, testable explanations for evidence-aggregation pipelines.

1 INTRODUCTION

Explanations for modern NLP systems are often judged by plausibility: whether selected text looks relevant to a person. This is insufficient for evidence-driven pipelines, where a prediction is produced by retrieval, extraction, aggregation, and a final decision. An explanation is useful for auditing only when it exposes how the system actually combined evidence and predicts what will happen when that evidence changes (Jacovi & Goldberg, 2020; DeYoung et al., 2020).

We study pipelines that decide a query from discrete evidence items—sentences, extracted micro-claims, or reported numerical comparisons. This setting makes a stronger kind of explanation possible. Rather than attach a post-hoc score to each input, we define an inference rule whose intermediate updates are the explanation. The central object is a *Belief Ledger*: a ranked, provenance-linked record of evidence clusters and their quantified contributions to a posterior belief.

The difficulty is that evidence items are commonly redundant or correlated. Treating every retrieved sentence as independent can double-count repeated findings and produce overconfident, visually convincing explanations. Moreover, binary decisions and numerical performance margins are often conflated despite living in different mathematical spaces. Belief Ledgers address both issues by clustering dependent evidence, aggregating one contribution per cluster, and maintaining separate models for posterior log-odds and continuous margins.

Our contributions are: (i) a posterior-accounting representation whose ledger entries exactly decompose the posterior log-odds for evidence-driven binary decisions; (ii) dependence-aware clustering and uncertainty-aware likelihoods, together with a separate conjugate model for numerical margins; and (iii) an intervention-based evaluation that tests not only task accuracy, but calibration, accounting error, and whether explanations predict posterior shifts under matched evidence edits. Experiments on scientific proposition verification, MultiRC, and FEVER show improved calibration and competitive task performance; on SciProp, the method improves accuracy by 1.6 points while reducing ECE from 0.120 to 0.034.

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.

2 Related work

Post-hoc attribution and rationales. LIME and SHAP estimate feature importance from perturbations (Ribeiro et al., 2016; Lundberg & Lee, 2017); gradient-based methods attribute predictions in differentiable models (Sundararajan et al., 2017). Rationale models select text used by a downstream predictor (Lei et al., 2016; DeYoung et al., 2020). These methods can be useful, but a high-ranked span need not provide a quantitative account of a pipeline's confidence. Attention weights likewise do not by themselves establish faithful reasoning (Jain & Wallace, 2019; Abnar & Zuidema, 2020).

Faithfulness and verification. Deletion, insertion, and counterfactual tests are standard tools for probing whether explanations track a model's computation (Jacovi & Goldberg, 2020; Wiegreffe & Pinter, 2019). Fact-verification systems retrieve evidence and predict support labels (Thorne et al., 2018; Wadden et al., 2020). Our work builds on these settings but treats explanation as posterior accounting: the system exposes the units and updates by which it aggregates evidence, allowing direct tests of predicted versus observed interventions.

3 Belief Ledgers

3.1 Task and evidence

A query q induces a binary hypothesis $H \in \{0, 1\}$. For example, $H = 1$ can denote that method A outperforms method B on dataset D under metric M; in MultiRC it denotes that an answer option is correct. An evidence extractor returns $\mathcal{E} = \{E_i\}_{i=1}^{n}$. Each item contains a source identifier, a support/refute polarity, optional numeric statistic x_i, and an uncertainty estimate σ_i^2.

3.2 Dependence-aware evidence clustering

Evidence from the same source, experimental setup, or paraphrased result should not be counted repeatedly. We form clusters $C_{1:m} = g(\mathcal{E})$ using shared source identifiers, normalized entity tuples, and semantic similarity of micro-claim text. In the experiments, agglomerative clustering uses cosine threshold $\tau = 0.78$, selected on the SciProp development split to minimize accounting error while maintaining cluster purity. A cluster statistic s_j is a robust aggregate (median or winsorized mean), with uncertainty

$$\tau_j^2 = \mathrm{median}_\ell(\sigma_{j,\ell}^2) + \lambda \,\mathrm{Var}_\ell(x_{j,\ell}). \tag{1}$$

This construction turns a potentially dependent set of mentions into an exchangeable unit for posterior aggregation.

3.3 Binary posterior accounting

Let $\pi = P(H = 1)$ be a development-calibrated prior. For each cluster, we specify class-conditional likelihoods $p(s_j \mid H = h)$ and define its log Bayes factor

$$W_j = \log \frac{p(s_j \mid H = 1)}{p(s_j \mid H = 0)}. \tag{2}$$

Under conditional independence across clusters given H,

$$\log \frac{P(H = 1 \mid C_{1:m})}{P(H = 0 \mid C_{1:m})} = \log \frac{\pi}{1 - \pi} + \sum_{j=1}^{m} W_j. \tag{3}$$

Equation 3 is the accounting guarantee: every displayed ledger entry uses the same unit (log-odds), and the entries exactly reconstruct the posterior. The decision confidence is the sigmoid of the right-hand side. We display clusters in decreasing $|W_j|$ and retain the smallest prefix accounting for at least $\rho = 0.90$ of total absolute contribution.

Table 1: Main-task performance and calibration (mean ± SD over five seeds).

Dataset	Metric	NLI-Agg.	Ledger	Δ
SciProp	Acc. / ECE	74.2±0.8 / .120	75.8±0.4 / .034	+1.6 / -.086
MultiRC	F1a / ECE	71.1±1.2 / .150	72.4±0.6 / .054	+1.3 / -.096
FEVER	Acc. / ECE	86.5±0.4 / .110	87.1±0.3 / .040	+0.6 / -.070

3.4 Continuous margin accounting

When evidence provides numerical margins, we use a separate model rather than mixing scores with log-odds:

$$\delta \sim \mathcal{N}(\mu_{\delta,0}, \sigma^2_{\delta,0}), \qquad x_j \mid \delta \sim \mathcal{N}(\delta, \sigma_j^2). \tag{4}$$

The posterior is conjugate,

$$\sigma^{-2}_{\text{post}} = \sigma^{-2}_{\delta,0} + \sum_{j \in J} \sigma_j^{-2}, \tag{5}$$

$$\mu_{\text{post}} = \sigma^2_{\text{post}} \left(\sigma^{-2}_{\delta,0} \mu_{\delta,0} + \sum_{j \in J} \sigma_j^{-2} x_j \right). \tag{6}$$

With $\alpha_j = \sigma^2_{\text{post}} \sigma_j^{-2}$, the posterior-mean shift decomposes as $\mu_{\text{post}} - \mu_{\delta,0} = \sum_j \alpha_j (x_j - \mu_{\delta,0})$. Thus a margin ledger is also additive, but it is never combined with the decision ledger.

4 Experimental setup

Tasks. SciProp contains NLP/ML performance propositions paired with candidate scientific evidence. We enforce paper-disjoint and time-disjoint splits; its leakage audit finds 0.2% exact-match overlap and 1.1% entity leakage. MultiRC provides sentence-level rationales for multi-sentence question answering (Khashabi et al., 2018). FEVER tests retrieval-based claim verification (Thorne et al., 2018); SNLI is used as an additional controlled entailment check (Bowman et al., 2015).

Baselines and metrics. We compare against a strong unconstrained NLI aggregator, LIME, Integrated Gradients where applicable, rationale models, and retrieval-plus-NLI systems. We report accuracy or task F1, expected calibration error (ECE, ten equal-width bins), and matched-budget deletion and insertion AUC. We also measure posterior accounting error Δ_{acct} and intervention alignment error (IAE), the absolute discrepancy between the ledger-predicted and re-run posterior shift. All reported primary results are mean ± standard deviation over five seeds.

Matched interventions. A fair comparison must control for evidence-unit size. Our primary protocol deletes or inserts evidence to identical subword-token budgets. For Belief Ledgers, units are ranked clusters; for sentence-level baselines, units are ranked sentences. A secondary protocol maps all methods to the same clusters. Numeric value-edit interventions perturb SciProp evidence by $\pm 1\sigma$ and $\pm 2\sigma$.

5 Results

Table 1 shows that the ledger improves both decision quality and calibration. The largest gain occurs on SciProp, whose multiple partially redundant findings make dependence control particularly important. The posterior trajectories also stabilize rapidly: the mean top-cluster update is $+2.1$ bits and the average absolute contribution falls below 0.1 bits after rank four.

Faithfulness. On SciProp, Belief Ledgers obtain token-budgeted deletion AUC of 0.152 and insertion AUC of 0.589; on MultiRC the corresponding values are 0.224 and 0.514. Table 2 shows low accounting and intervention discrepancies. Unlike a rationale that merely identifies text, the ledger quantitatively predicts the confidence drop caused by removing high-contribution evidence.

Table 2: Accounting and intervention alignment (mean ± SD; lower is better except coverage).

Dataset	Δ_{acct}	IAE (Del.)	IAE (Pert.)	CI cov.
SciProp	.002±.001	.041±.012	.038±.009	.942
MultiRC	.005±.002	.062±.015	–	–
FEVER	.003±.001	.055±.014	–	–
SNLI	.008±.003	.089±.021	–	–

Table 3: SciProp ablations. Dependence-aware clustering is central to calibration and intervention faithfulness.

Variant	ECE ↓	Del.-AUC ↓
Belief Ledger	.034	.152
– clustering	.089	.174
– uncertainty	.045	.161
+ fixed prior	.039	.155

Ablations and residual dependence. Removing clustering sharply worsens calibration, consistent with double-counting redundant evidence. Removing uncertainty also worsens faithfulness. Residual between-cluster dependence is low on SciProp (mean mutual information 0.081; 95th percentile 0.140), compared with within-cluster item dependence (mean 0.64), supporting but not proving the conditional-independence approximation.

Human evaluation. In a blinded within-subject pilot (15 annotators, 300 items), users rated ledger explanations as more verifiable (5.8 versus 4.2 on a 1–7 scale), ranked them as more faithful (1.3 versus 2.1; lower is better), and verified decisions faster (28.1 versus 42.5 seconds) than saliency maps. Inter-annotator agreement was Krippendorff's $\alpha = 0.76$.

6 Discussion and limitations

Belief Ledgers are designed for systems whose decisions explicitly aggregate discrete evidence. They are not a universal replacement for explanations of arbitrary end-to-end neural computations. Their validity depends on extraction quality, likelihood specification, and the extent to which clustering captures dependence. Residual cross-cluster correlation and likelihood misspecification can affect posterior calibration; reporting accounting error and intervention alignment makes such failures visible rather than silently converting them into persuasive narratives. Future work should evaluate larger and more diverse human populations and extend the dependence model beyond hard clustering.

7 Conclusion

Belief Ledgers turn evidence aggregation into a transparent posterior-accounting process. By clustering dependent evidence, separating binary log-odds from numeric margins, and exposing provenance-linked additive updates, they provide explanations that are auditable by construction and testable through interventions. The results show that this structure improves calibration while supporting competitive prediction and measurable faithfulness in verification and reading-comprehension pipelines.

AI use statement

We used generative AI tools to assist with language editing, paper structure, title and keyword suggestions, and LaTeX formatting. The authors reviewed and revised all AI-assisted text, verified the mathematical statements and reported experimental results against the underlying work, and take full responsibility for the final manuscript, claims, and artifacts.

ETHICS AND REPRODUCIBILITY STATEMENT

Belief Ledgers are intended to improve auditability, but provenance-linked explanations can still be over-trusted when the underlying evidence extractor is wrong. We therefore report uncertainty and recommend retaining source links for human verification. The paper specifies the inference equations, clustering threshold, evaluation metrics, split controls, and intervention protocols. The supplementary material contains additional hyperparameters, sensitivity analyses, and evaluation details; code and data should be released anonymously where permitted.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

Jay DeYoung et al. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*, 2020.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems. In *ACL*, 2020.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL*, 2019.

Daniel Khashabi et al. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *NAACL*, 2018.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *EMNLP*, 2016.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 2016.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

James Thorne et al. Fever: A large-scale dataset for fact extraction and verification. In *NAACL*, 2018.

David Wadden et al. Fact or fiction: Verifying scientific claims. In *EMNLP*, 2020.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP*, 2019.

A ADDITIONAL EXPERIMENTAL DETAIL

A.1 CLUSTERING SENSITIVITY AND LIKELIHOODS

The development sensitivity sweep ranges from $\tau = .50$ ($\Delta_{acct} = .015$, deletion AUC .224, accuracy 74.8) to $\tau = .95$ (.025, .254, 74.1); the selected $\tau = .78$ gives .002, .152, and 75.8, respectively. The selected threshold balances accounting accuracy and intervention faithfulness. We additionally compared a robust default likelihood, a mixture model for noisy extractions, and a product-of-experts variant for SciProp numerical margins. The mixture achieves $\Delta_{acct} = .001$, perturbation IAE $.038$, and CI coverage $.951$; the product-of-experts reaches $.004$, $.049$, and $.915$, respectively. We retain the default because it is the simplest stable model.

A.2 QUALITATIVE EXAMPLE

For a SciProp query, the first cluster can contain a primary reported result and raise a prior near $.50$ to $.78$. A later cluster from a correction or contradictory setting can lower it to $.44$, after which an independent replication can yield a final posterior near $.55$. The ledger records these opposing changes separately, rather than attributing the final uncertainty to a diffuse set of similar sentences. This example illustrates the intended auditing behavior; individual evidence entries retain source provenance.