From “Thinking” to “Justifying”: Aligning High-Stakes Explainability with Professional Communication Standards

Anonymous ACL submission

Abstract

Explainable AI (XAI) in high-stakes domains should help stakeholders trust and verify system outputs. Yet Chain-of-Thought methods reason before concluding, and logical gaps or hallucinations can yield conclusions that do not reliably align with their rationale (Turpin et al., 2023). Thus, we propose “Result → Justify”, which constrains the *output communication* to present a conclusion before its structured justification. We introduce SEF (Structured Explainability Framework), operationalizing professional conventions (e.g., CREAC, BLUF) via six metrics for structure and grounding. Experiments across four tasks in three domains validate this approach: all six metrics correlate with correctness ($r = 0.20$–0.42; $p < 0.001$), and SEF achieves 83.9% accuracy (+5.3 over CoT). These results suggest structured justification can improve verifiability and may also improve reliability. Code and data will be released upon acceptance.

1 Introduction

As large language models (LLMs) are deployed in high-stakes domains, explainable AI (XAI) has become essential for helping humans assess model outputs (Bommasani et al., 2021). Prior XAI work has progressed from post-hoc attribution (Ribeiro et al., 2016; Lundberg and Lee, 2017) to process-oriented approaches like Chain-of-Thought (CoT) (Wei et al., 2022; Yao et al., 2023). While CoT prompting elicits step-by-step reasoning before a conclusion, hallucinations or logical gaps can yield conclusions that misalign with their rationale (Turpin et al., 2023; Lanham et al., 2023). Such misaligned or unfaithful explanations can hinder stakeholders’ understanding of the model’s reasoning.

To address this challenge, we draw on professional communication practices as a source of *justification structure*, and use them to format explanations as verifiable justifications. CREAC (Conclusion-Rule-Explanation-Analysis-Conclusion) in legal writing (Mangan et al., 2022) and BLUF (Bottom Line Up Front), a U.S. military communication standard widely adopted in business (Sehgal, 2016), are widely used templates for presenting a defensible conclusion and its supporting evidence. Legal theory frames this as the distinction between the *context of discovery* (how one arrives at an answer) and the *context of justification* (how one defends it) (Burton, 2007). We hypothesize that AI explanations in high-stakes domains may benefit from adopting this justification-oriented structure: “**Result** → **Justify**.” Crucially, SEF targets the *context of justification* by constraining output communication rather than internal deliberation. Such perspective is supported by mathematical writing, where results precede polished proofs that may differ from the original derivation, emphasizing clarity and evaluability (Lamport, 2012; Aigner and Ziegler, 2010).

We propose SEF (Structured Explainability Framework), which operationalizes CREAC and BLUF conventions through six metrics capturing *plausibility* (structural clarity) and *faithfulness* (evidence grounding) (Jacovi and Goldberg, 2020) (Figure 1). While mechanistic interpretability examines internal activations, SEF acts as a behavioral probe: by measuring output structure, we infer how prompting constraints shape reasoning organization (Lanham et al., 2023; Lyu et al., 2023). Experiments across four tasks in three domains validate this approach:

1. All six metrics correlate with correctness ($r = 0.20$–0.42; $p < 0.001$), suggesting that structured justifications align with more reliable outputs.
2. SEF achieves 83.9% accuracy (+5.3 over CoT). Notably, Direct (80.9%) outperforms CoT (78.6%), consistent with concerns about unconstrained reasoning.

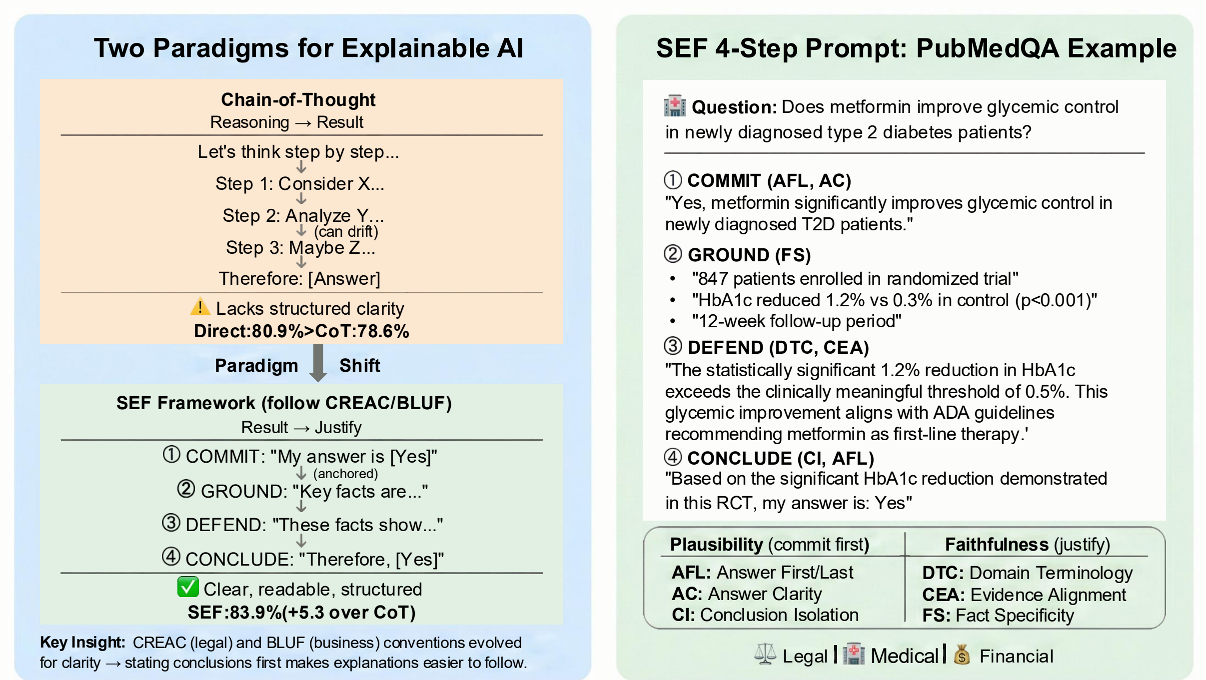

Figure 1: **SEF: Structured Explainability via CREAC/BLUF Conventions. Left:** Paradigm comparison showing CoT ("Reasoning → Result") vs. SEF ("Result → Justify"). CoT traces can be hard to verify (Direct 80.9% > CoT 78.6%); SEF achieves 83.9%. **Right:** PubMedQA example demonstrating the 4-step prompt (Commit, Ground, Defend, Conclude) with metric mappings.

3. Ablations show the conclusion-first presentation scaffold is the main driver of accuracy: removing plausibility constraints causes the largest drop (−32.9), while removing faithfulness constraints reduces accuracy modestly (−4.3) but weakens domain grounding.

2 Related Work

Process-Oriented Reasoning. CoT (Wei et al., 2022) follows a "Reasoning → Result" paradigm but has known failures. Turpin et al. (2023) show CoT can be *unfaithful*, offering plausible reasoning unrelated to actual computation. Similarly, Lanham et al. (2023) note models may reach correct answers via flawed steps. Remedies like self-consistency (Wang et al., 2022), Tree-of-Thought (Yao et al., 2023), and Chain-of-Verification (Dhuliawala et al., 2024) improve accuracy but retain this paradigm without structural constraints, which reduces verifiability in high-stakes settings. We propose "Result → Justify" (like CREAC and BLUF): state the conclusion first, then build a structured, verifiable defense.

Explainability Evaluation. Explanation evaluation separates faithfulness (actual model reasoning) from plausibility (human reasonableness) (Jacovi and Goldberg, 2020; Wiegreffe and Marasović, 2021). Existing metrics (e.g., LIME, SHAP) focus on *token-level* attribution (Ribeiro et al., 2016; Lundberg and Lee, 2017). While these identify *what* inputs influence predictions, they ignore *how* the model structures arguments. Because current methods cannot distinguish coherent justification from a list of salient tokens, we propose metrics grounded in professional communication (e.g., evidence linking) to assess if AI explanations mirror expert practice.

Domain-Specific NLP. Benchmarks exist for legal (Guha et al., 2023), medical (Jin et al., 2019), and financial (Malo et al., 2013) NLP. We complement them with metrics that operationalize justification-oriented structure, which these domains value for review and accountability.

3 SEF Framework

SEF is dual-purpose: a prompting specification and a heuristic metric suite. Concretely, we use the metrics both to evaluate all methods and to design a constrained, sectioned output format for generation. Full heuristic definitions and the prompt template are in Appendix A (App. A.1–A.2).

Six Justification Metrics. We score each explanation with six heuristic metrics in $[0, 1]$, grouped into two dimensions:

Plausibility (commitment and sectioning).

- *Answer First/Last (AFL)*: answer appears in the first/last 200 characters (1.0 both, 0.5 one, 0.0 neither).
- *Answer Clarity (AC)*: explicit answer statements via regex patterns (strong 1.0, medium 0.7, weak 0.3).
- *Conclusion Isolation (CI)*: conclusion structurally separated (explicit conclusion headers 1.0; discourse markers 0.6).

Faithfulness (justify with evidence).

- *Domain Terminology Consistency (DTC)*: counts curated domain terms (thresholded score 0.2–1.0).
- *Conclusion–Evidence Alignment (CEA)*: detects evidence-linking language (e.g., "based on", "according to") plus analysis cues.
- *Fact Specificity (FS)*: counts specificity indicators (numbers/quotes, enumerations) against vagueness markers.

Metric-informed prompting. The metrics induce a 4-step schema: **Commit** (AFL, AC), **Ground** (FS), **Defend** (DTC, CEA), **Conclude** (CI, AFL). Unlike CoT, SEF commits first and constrains format to reduce drift.

4 Experiments

4.1 Setup

Models. We evaluate four open-weight 12–14B instruction-tuned models: DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025), Gemma 3 12B (Gemma Team et al., 2024), Ministral-3-14B (Mistral AI, 2025), and Qwen 2.5 14B (Bai et al., 2023). We use greedy decoding (temperature=0) via vLLM (Kwon et al., 2023).

Tasks. We evaluate on four Yes/No classification tasks spanning legal (Guha et al., 2023), medical (Jin et al., 2019), and financial (Malo et al., 2013) domains (1,618 test samples total; see Appendix A.4 for details). The binary format makes accuracy a sufficient metric while enabling comparison of explanation quality through our six metrics.

Baselines. We compare against six prompting strategies: (1) **Direct**, answer without explanation; (2) **Chain-of-Thought** (CoT; Wei et al., 2022), "think step by step"; (3) **Tree-of-Thought** (ToT; Yao et al., 2023), explore multiple reasoning paths; (4) **Chain-of-Verification** (CoVe; Dhuliawala et al., 2024), verify reasoning steps; (5) **Vanilla RAG** (V-RAG; Lewis et al., 2020), retrieve relevant context;

Dimension	Metric	Pearson r	p-value
Plausibility	AFL	0.36	$< .001$
	AC	0.42	$< .001$
	CI	0.31	$< .001$
Faithfulness	DTC	0.20	$< .001$
	CEA	0.23	$< .001$
	FS	0.36	$< .001$

Table 1: Pearson correlation between each metric and prediction correctness (n=90,608 samples across all models, methods, and tasks). All metrics show significant correlations with accuracy ($p < 0.001$). Plausibility metrics tend to show stronger effects ($r = 0.31$–0.42) than faithfulness metrics ($r = 0.20$–0.36).

(6) **Self-RAG** (Asai et al., 2023), iterative retrieval with self-reflection. All baselines use the same greedy decoding and are evaluated on identical test sets; full prompt templates are in Appendix A.3.

Ablation Protocol. For systematic ablation, we create eight SEF variants: six single-component ablations (removing AFL, AC, CI, DTC, CEA, or FS instructions), plus two dimension-level ablations (removing all Presentation or all Domain instructions). Each variant runs on all four models across all four tasks (64 experimental conditions total).

4.2 Metric-Accuracy Correlation

We first examine whether the metrics correlate with prediction accuracy. Crucially, we compute each metric on 90,608 outputs from *all* methods (4 models $\times$ 7 methods $\times$ 4 tasks $\times$ variable samples). The correlations reflect general associations between structure and correctness, not artifacts of SEF-specific formatting.

Table 1 shows Pearson correlations between each metric and prediction correctness. All six metrics correlate significantly with accuracy ($p < 0.001$), supporting our hypothesis that CREAC/BLUF-style justification structure is associated with more reliable outputs. Both dimensions contribute: plausibility metrics ($r = 0.31$–0.42) capture commitment and sectioning, while faithfulness metrics ($r = 0.20$–0.36) capture domain grounding.

4.3 Main Results

Table 2 shows that SEF achieves 83.9% average accuracy, outperforming CoT by 5.3 points (+9.0 on PubMedQA, +4.6 on ConsumerQA, +2.8 on FPB). On Hearsay, SEF matches the best baseline (CoVe) at 54.5%.

Notably, Direct prompting (80.9%) outperforms CoT (78.6%) and most reasoning methods, a

Method	FPB	CQA	Hear.	PMQ	Avg
Direct	94.3	93.9	49.5	86.1	80.9
CoT	92.7	91.5	49.7	80.5	78.6
ToT	93.6	89.8	45.0	82.4	77.7
CoVe	88.9	90.2	54.5	83.9	79.4
V-RAG	95.3	86.6	45.0	81.0	76.9
Self-RAG	76.5	88.0	48.1	76.9	72.4
SEF	**95.5**	**96.1**	**54.5**	**89.5**	**83.9**

Table 2: Accuracy (%) averaged over four models. SEF achieves highest average accuracy (83.9%), outperforming all baselines including CoT (+5.3 points). **Tasks:** FPB = Financial PhraseBank, CQA = ConsumerQA, Hear. = Hearsay, PMQ = PubMedQA.

Variant	FPB	CQA	Hear.	PMQ	Avg	Δ
SEF (Full)	95.5	96.1	54.5	89.5	83.9	–
w/o AFL	94.5	92.3	41.3	86.9	78.8	−5.1
w/o AC	90.5	91.0	47.8	85.0	78.6	−5.3
w/o CI	70.9	68.9	49.3	56.7	61.4	−22.5
w/o DTC	94.0	91.4	48.3	84.8	79.6	−4.3
w/o CEA	90.5	90.5	54.7	84.5	80.0	−3.9
w/o FS	90.8	91.6	50.4	84.8	79.4	−4.5
w/o Pres.	35.1	67.5	53.8	47.8	51.0	−32.9
w/o Domain	92.3	90.5	50.9	84.5	79.6	−4.3

Table 3: Ablation results: accuracy (%) when removing individual components or entire dimensions. Plausibility components (especially CI) tend to cause larger drops than faithfulness components, consistent with correlation patterns in Table 1. **Notation:** w/o = without; Pres. = Plausibility (AFL+AC+CI); Domain = Faithfulness (DTC+CEA+FS).

counter-intuitive finding consistent with recent work showing that CoT's free-form reasoning traces can introduce errors through drift or unfaithful explanations (Turpin et al., 2023). SEF outperforms even Direct (+3.0 points), suggesting that *structured* justification adds value beyond mere answer commitment.

4.4 Ablation Study

Table 3 reveals how constraints contribute to verifiable justifications:

Plausibility enables structured presentation. Without plausibility constraints, SEF degrades to unstructured reasoning, causing a massive accuracy drop (−32.9). This underscores that the explicit structural constraints are the primary driver of performance.

Faithfulness enables domain grounding. Removing faithfulness constraints reduces accuracy modestly (−4.3) but results in generic explanations. The impact is task-dependent: DTC is vital for Hearsay (−6.2) due to legal precision, while FS drives performance on PubMedQA by enforcing clinical evidence citation.

Both dimensions are necessary. Plausibility provides the structural scaffold; faithfulness fills it with domain-appropriate content. Neither alone yields fully verifiable justifications.

5 Discussion

Accuracy as behavioral validation. Direct (80.9%) outperforming CoT (78.6%) is consistent with work questioning CoT faithfulness (Turpin et al., 2023), suggesting that CoT's free-form reasoning traces may sometimes hurt rather than help. SEF (83.9%) combines answer commitment (like Direct) with structured justification (unlike Direct's lack of explanation, and unlike CoT's unconstrained traces). This accuracy pattern validates that CREAC/BLUF-style structure (Mangan et al., 2022; Sehgal, 2016) reflects effective output organization, not just convention.

What output structure reveals. The systematic correlation between structure and accuracy ($r = 0.20$–0.42) suggests prompting constraints shape how models organize their outputs. Imposing conclusion-first structure is associated with higher accuracy on average. This behavioral analysis complements mechanistic interpretability by highlighting *which output organization patterns* coincide with reliable outputs.

Implications. For high-stakes domains, explanations may benefit from mirroring professional practice: structured defenses of committed conclusions, rather than reasoning traces hoping to find answers. Our results suggest this improves both interpretability and reliability.

6 Conclusion

We introduced SEF, which structures AI explanations via CREAC and BLUF conventions. This "Result → Justify" paradigm states the conclusion upfront, then builds a structured defense as a *presentation* scaffold. SEF achieves 83.9% accuracy (+5.3 over CoT), suggesting verifiable justifications may also improve reliability. Additionally, our six metrics provide scalable signals of quality to enable verification-oriented evaluation. For high-stakes domains, conclusion-first justification supports more trustworthy model use.

Limitations

Proxy metrics. Our six metrics are scalable heuristics that approximate professional communication conventions, not deep domain soundness. They measure *structural* compliance (e.g., presence of conclusion headers, domain terms) rather than *semantic* correctness of the justification itself. A response can score well while containing factual errors. We mitigate this concern by computing correlations across *all* methods (not just SEF), showing that structure correlates with accuracy generally. Importantly, even when the answer is incorrect, structured justifications improve readability: the human-friendly format makes it easier for reviewers to identify errors in the reasoning. These automatic metrics serve as a scalable first-pass filter; human expert evaluation to assess semantic faithfulness remains future work.

Statistical interpretation. Our per-sample correlations quantify association (effect size r), not causality, and samples may not be independent due to shared prompts, models, and datasets. We therefore interpret them descriptively and complement them with ablations that intervene on prompting constraints.

Scope and artifacts. We validate on Yes/No tasks across three domains with 12–14B instruction-tuned models. Extending to multi-class settings, open-ended generation, additional domains, or larger models remains future work. We use public datasets and open-weight models under their stated licenses/terms and do not redistribute third-party datasets or model weights.

Ethical Considerations

Persuasive hallucinations. Structured justifications may make incorrect answers more convincing. Any use of SEF should include human oversight and avoid presenting generated justifications as professional advice or as evidence of correctness.

Professional boundaries. SEF should not replace professional judgment in legal, medical, or financial decisions. Our 83.9% accuracy implies substantial error rates.

Privacy and data use. We do not collect new personal data or deploy systems. We use publicly released research benchmarks; users should respect dataset terms and avoid applying the framework to sensitive or identifiable records without appropriate approvals and safeguards.

Environmental impact. We do not train new models. Experiments use existing 12–14B models with deterministic decoding, but compute remains non-trivial; releasing code and configurations is intended to reduce redundant reruns.

References

Martin Aigner and Günter M. Ziegler. 2010. *Proofs from THE BOOK*. Springer Berlin Heidelberg.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint*.

Steven J Burton. 2007. *An introduction to law and legal reasoning*. Wolters Kluwer Law & Business.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, page 3563–3578. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint*.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. Legalbench: a collaboratively built benchmark for measuring legal reasoning in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. 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)*, page 2567–2577. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626. ACM.

Leslie Lamport. 2012. How to write a 21st century proof. *Journal of Fixed Point Theory and Applications*, 11(1):43–63.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint*.

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. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2013. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Jean Mangan, Brittany Blanchard, Gabrielle Gravel, Chase Lyndale, and Connely Doizé. 2022. 1.10: CREAC legal writing paradigm. In *Legal Writing Manual*. University System of Georgia via Affordable Learning Georgia. LibreTexts.

Mistral AI. 2025. Introducing mistral 3. *Mistral AI Technical Blog*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Kabir Sehgal. 2016. How to write email with military precision. *Harvard Business Review*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Implementation Details

Reproducibility note. SEF's metrics are **rule-based heuristics** (regex/pattern matching), not model-based judges. This makes the evaluation lightweight and deterministic. These metrics assess structural compliance rather than semantic correctness; their value lies in producing human-readable outputs that facilitate expert review. We summarize the exact scoring rules below.

A.1 Metric Computation (Regex/Heuristics)

Inputs. Each metric consumes an explanation string e and a predicted binary answer $a \in \{\texttt{Yes}, \texttt{No}\}$. All matching is performed on a lowercased copy of e.

AFL (Answer First/Last). Let e_{head} be the first 200 characters of e and e_{tail} the last 200. AFL is 1.0 if a occurs in both e_{head} and e_{tail}, 0.5 if it occurs in exactly one, else 0.0. Occurrence is checked by direct substring match or a word-boundary match for `Yes`/`No`.

AC (Answer Clarity). We detect explicit answer statements with regexes. If any *strong* pattern matches, AC=1.0; else if any *medium* pattern matches, AC=0.7; else if a appears anywhere in e, AC=0.3; else 0.0. Strong patterns capture explicit declarations (e.g., "My answer is: Yes", "Final answer: No", "The answer is: Yes", or "Answer: Yes"); medium patterns capture discourse markers that directly precede a Yes/No conclusion (e.g., "therefore, Yes" or "in conclusion, No").

CI (Conclusion Isolation). CI=1.0 if the output contains an explicit conclusion header (e.g., "CONCLUSION:" / "Conclusion:"); CI=0.6 if it contains common concluding markers (e.g., "in conclusion", "to conclude", "in summary", "therefore", "thus"); else 0.0.

DTC (Domain Terminology Consistency). We count occurrences of curated, domain-specific terms (legal/medical/financial lexicons). If no lexicon is defined, DTC=0.5. Otherwise, if the term-count is ≥ 5, DTC=1.0; if ≥ 3, DTC=0.8; if ≥ 1, DTC=0.5; else DTC=0.2.

CEA (Conclusion–Evidence Alignment). We count evidence-linking cues and analysis cues. If link-cue count ≥ 3 and analysis-cue count ≥ 1, CEA=1.0; else if link-cue count ≥ 2, CEA=0.8; else if link-cue count ≥ 1, CEA=0.5; else if analysis-cue count ≥ 1, CEA=0.3; else 0.0. Cues include phrases such as "based on", "according to", "this suggests", and explicit "ANALYSIS" sectioning.

Metric	Heuristic scoring rule (summary)
AFL	Answer appears in both the first and last 200 characters (1.0); only one side (0.5); neither (0.0).
AC	Explicit answer declaration (1.0); weaker conclusion marker preceding Yes/No (0.7); answer appears anywhere (0.3); otherwise (0.0).
CI	Explicit conclusion header present (1.0); concluding discourse markers present (0.6); otherwise (0.0).
DTC	Domain term count ≥ 5 (1.0); ≥ 3 (0.8); ≥ 1 (0.5); otherwise (0.2).
CEA	Link-cue count ≥ 3 and analysis-cue count ≥ 1 (1.0); link cues ≥ 2 (0.8); ≥ 1 (0.5); analysis cues ≥ 1 (0.3); otherwise (0.0).
FS	Specificity count ≥ 4 and vagueness count ≤ 1 (1.0); specificity ≥ 3 (0.8); ≥ 2 (0.6); ≥ 1 (0.4); otherwise (0.2).

Table 4: Summary of the six SEF metrics and their heuristic scoring rules.

FS (Fact Specificity). We count specificity indicators and vagueness indicators. FS=1.0 if specificity count ≥ 4 and vagueness count ≤ 1; else if specificity count ≥ 3, FS=0.8; else if ≥ 2, FS=0.6; else if ≥ 1, FS=0.4; else 0.2. Specificity cues include numbers and quotes, enumerations (e.g., "first/second"), and explicit fact headers (e.g., "KEY FACTS"); vagueness cues include hedges (e.g., "generally", "possibly").

A.2 SEF Prompt Template

Overview. Our SEF prompting method enforces a conclusion-first structured format with four labeled sections. Table 5 shows the prompt template (domain role/instructions + sectioned format) used in our experiments; ablations remove the corresponding instructions/sections.

A.3 Baseline Prompt Templates

For reproducibility, we document the core prompt structures for each baseline. All methods receive the same `Context` and `Question` inputs; only the instruction wrapper differs.

Direct. Minimal prompt requesting only the answer:

```
Answer the following question.
Context: [context]
Question: [question]
Provide only the answer (Yes/No).
Answer:
```

Prompt component	Template text (verbatim headers)	Metrics targeted
Role instruction	`[Domain role] Analyze the following question with precision and clarity.`	–
Domain terminology (optional)	`Use precise [legal/medical/financial] terminology consistently.`	DTC
Fact specificity (optional)	`Cite specific facts from the context, not vague generalizations.`	FS
Context + question	`Context: ...` `Question: ...`	–
Section 1: Commit	`ANSWER PREVIEW: State your answer upfront (Yes or No).`	AFL (first), AC
Section 2: Ground	`KEY FACTS: List 2-3 specific facts from the context that are most relevant.`	FS

Section 3: Defend	`ANALYSIS:`
	`- Use precise domain terminology`
	`- Explain how each fact supports your answer`
		DTC, CEA

Section 4: Conclude	`CONCLUSION:`
	`- State your final answer clearly and unambiguously`
	`- Summarize the key evidence supporting your answer`
	`- End with: My answer is: [Yes/No]`
		CI, AFL (last), AC

Table 5: SEF prompt template used in our experiments.

Chain-of-Thought (CoT). Free-form reasoning triggered by "think step by step":

```
Given the following context and question,
let's think step by step.
Context: [context]
Question: [question]
Let's think step by step:
1. First, let me understand the context
and relevant rules.
2. Then, I'll analyze how they apply to
this question.
3. Finally, I'll determine the correct
answer.
Step-by-step reasoning:
```

Tree-of-Thought (ToT). Three-stage process: (1) generate k=3 initial approaches, (2) develop each into full analysis, (3) evaluate and select best path.

Stage 1: `Propose 3 different analytical approaches (1-2 sentences each).`
Stage 2: `For each approach: "Develop this into a complete analysis... then state your final answer."`
Stage 3: `"Which path (1, 2, or 3) provides the most thorough and accurate analysis?"`

Chain-of-Verification (CoVe). Four-stage process with self-verification:

Stage 1: `"Answer the question with reasoning."`
Stage 2: `"Generate 3 verification questions to check accuracy."`
Stage 3: `Answer each verification question against the context.`
Stage 4: `"Based on initial analysis and verification, provide your final answer (incorporate any corrections)."`

Vanilla RAG (V-RAG). Retrieve-then-generate with top-k=3 passages:

```
Answer the question using the provided
context.
Retrieved Context: [top-k passages by
keyword overlap]
Question: [question]
Based on the retrieved context, provide
your reasoning and answer.
```

Self-RAG. Multi-stage retrieval with reflection and self-critique:

Stage 1: `Assess if retrieval is needed (RETRIEVE or USE_FULL).`
Stage 2: `If retrieving, filter passages by relevance (RELEVANT or NOT_RELEVANT).`
Stage 3: `Generate with reflection: "[Reflection: ...] [Generation: ...] [Answer: ...]"`
Stage 4: `Self-critique for accuracy (NEEDS_REFINEMENT or SUFFICIENT).`
Stage 5: `If needed, refine based on critique.`

A.4 Datasets and Input Formatting

Tasks and sources. We evaluate on four binary classification tasks spanning three high-stakes domains, totaling 1,618 test samples:

- **Legal** (Guha et al., 2023): ConsumerQA (396 samples), identifying unfair terms in consumer contracts; Hearsay (94 samples), applying Federal Rule of Evidence 801(c).
- **Medical**: PubMedQA (Jin et al., 2019) (1,000 samples), answering biomedical research questions based on abstracts.

- **Financial**: FPB (Malo et al., 2013) (128 samples), classifying sentiment in financial news phrases.

Evaluation splits. We use the evaluation splits provided by each benchmark.

Input formatting. Each example is formatted into a `Context` field (the provided passage/document) and a `Question` field. All prompting methods receive the same inputs; only the instruction/prompt wrapper differs across methods.

A.5 Exact Model Identifiers

We report paper-friendly names in the main text; the exact vLLM/HF identifiers used by our experiment runner are:

- DeepSeek-R1-Distill-Qwen-14B: `deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`
- Gemma 3 12B: `google/gemma-3-12b-it`
- Ministral-3-14B: `mistralai/Ministral-3-14B-Instruct-2512`
- Qwen 2.5 14B Instruct: `Qwen/Qwen2.5-14B-Instruct`