L-FRESCO: Factual Recall Evaluation Score for Legal Analysis Generation

Anonymous Authors[1]

1. Introduction

Existing automatic tools evaluate the factuality of text generations based on *factual precision*, which measures the fraction of generated information being factually accurate (Min et al., 2023; Chern et al., 2023; Yue et al., 2023). However, comprehensiveness and precision are both crucial aspects of reliable and verifiable text generation (Liu et al., 2023). In this work, we show that precision-based factuality metrics are limited in evaluating the comprehensiveness of text generations from certain domains, especially legal texts (Magesh et al., 2024). We propose L-FRESCO, **F**actual **R**ecall **E**valuation **Sco**re for **L**egal analysis generation. Inspired by FActScore (Min et al., 2023), which decomposes generated text into atomic facts and then verifies their factuality, L-FRESCO follows a *Decompose-Then-Compare* framework to compute similarity between the reference atomic claim and the generated atomic claim. Moreover, we explore a generalized variant, FRESCO and discuss its potentials to be applied across text domains.

2. Background

2.1. Legal Analysis Generation Task

Lawyers write legal analysis for a variety of professional tasks. In Common Law jurisdictions, claims in legal analysis are supported by prior case rulings, also known as case law. Citing appropriate cases from millions of documents is a strenuous task. Empowering lawyers to write analysis with large langauge models (LLMs) can be a promising direction.

In one of our upcoming papers focused on legal generation, we evaluate the ability of LLMs to generate legal analysis with common evaluation metrics such as ROUGE (Lin, 2004), BARTScore (Yuan et al., 2021), and Citation Recall and Precision (Liu et al., 2023). However, we find that existing metrics are limited in reflecting the quality and soundness of legal analysis generation.

Formally, suppose that a case document consists of paragraphs $\{p_1, p_2, \ldots, p_N\}$, and we sample a reference paragraph p_t from the case document. Given the set of relevant passages cited by p_t, $R_t = \{r_1, r_2, ...r_T\}$, we define input $s = f_t(p_1, p_2, \ldots, p_{t-1}, R_t)$, where $f_t(\cdot)$ maps input context into prompts to LLMs. The task is to generate the next paragraph $p_t^{'} \sim P_{\text{LLM}}(\cdot|s)$, such that $\text{MET}(p_t, p_t^{'})$ is maximized for some evaluation metric $\text{MET}(\cdot)$.

2.2. Limitations of Existing Metrics for Legal Analysis Generation

Prompting LLMs to generate legal analysis, we find that outputs can have high metric scores without being analytical sound and factually comprehensive. Legal analysis texts follow regular patterns. Each claim is almost always supported by one or more case citations. For instance, from the case *Board of Trustees of the Airconditioning and Refrigeration Industry Health and Welfare Trust Fund et al v. Innovative Air, Inc. et al*:

> *Divers' status as a corporate officer does not exempt him from liability.* See Kayes v. Pacific Lumber Co., 51 F.3d 1449, 1459 (9th Cir.1995)

where the *red italicized* texts are the claim, and the blue texts are the citation that supports the claim.

However, prompting Meta-Llama-3-8B-Instruct [1] to generate, we obtain the following output:

> *As previously stated, 29 U.S.C. § 1002(21)(A) defines a fiduciary as a person who exercises any authority or control respecting management or disposition of [a plan's] assets.* P51 F.3d 1449, 1459 (9th Cir.1995).

While the generated claim defines what a fiduciary is, the significance of the citation should be showing that "corpoate officer has liability". This citation is a total mismatch of the generated claim and should deserve a low factuality score.

Existing metrics can overestimate the quality and factuality of legal analysis generation for the following reasons. First, generation can have an overall high lexical and semantic overlap with the reference text but is factually imprecise. Second, generation can have a high factual precision by generating simple and correct claims but makes incomprehensive citations, i.e., a low citation recall. Third, generation can have a high citation recall but supports claims with in-

[1] https://github.com/meta-llama/llama3

appropriate citations. This motivates a factual recall metric for legal and general long-form generation.

3. L-FRESCO

The high-level idea of L-FRESCO follows a *Decompose-Then-Compare* framework: we respectively **decompose** a paragraph of generated legal text into pairs of citations and claims; we repeat the same decomposition for its reference paragraph. We assume that in the reference text, the citation always properly supports the claim. Given the same citation, we **compare** the generated claim with the reference claim. If the two claims are similar enough, we hypothesize that the citation also properly supports the generated claim, and thus should be considered as high quality generation.

3.1. Decompose

As discussed in 2.2, legal analysis follows predictable patterns of attaching citations to every claim made. The citations and claims are in this sense, *atomic*, since each claim can be supported by a single case. Unlike FActScore (Min et al., 2023) that uses InstructGPT (Ouyang et al., 2022) to decompose the claims, L-FRESCO can use simple rule-based programs to extract pairs of claims and citations from case law data with the `eyecite` parser (Cushman et al., 2021).

3.2. Compare

The goal of the *Compare* step is to assess, given the same citation, whether the claim generated is close enough to the reference claim. Since we assume that the reference claim is appropriately supported by its citation, we hypothesize that enough closeness to the reference claim implies generated claim also being correctly supported by the citation.

3.3. Formulation

Similar to Min et al. (2023), L-FRESCO can be mathematically formulated as follows:

$$\text{L-FRESCO} = \frac{1}{|C_r|} \sum_{(c_i, c_r) \in (A_c, C_r)} \text{B-SIM}(c_i, c_r)$$

where A_c is the set of generated atomic claims, and C_r is the set of reference claims. Two claims are paired if they share the same case citation. $\text{B-SIM}(\cdot)$ is a binary similarity measure that outputs 1 if the two claims are close enough and 0, vice versa. This measure can be binary similarity metrics based on lexical or semantic overlaps, such as ROUGE with a threshold to make its output binary. Alternatively, we can use a zero-shot or fine-tuned LLM for evaluating binary relevance. In the full paper, we plan to benchmark the performance of different similarity measures and determine the optimal scoring component for L-FRESCO.

3.4. Plans for Evaluation

We plan to evaluate the effectiveness of L-FRESCO by comparing its correlation with human ratings in terms of analytical and writing quality, given by evaluators who have had training in law. We will also assess the correlation of existing metrics with human ratings and contrast the effectiveness of L-FRESCO to other metrics. For our *Decompose* step, we will explore two options to extract the claim and citation pairs. First, develop a precise rule-based program. Second, use a bootstrapped Llama-3-8B or GPT-4 (Josh Achiam et al., 2024; Wanner et al., 2024). For our *Compare* step, we will use ROUGE, GPT-4, zero-shot Llama-3-8B, and fine-tuned Llama-3-8B as different similarity functions to measure the closeness between claims.

4. Generalized FRESCO

L-FRESCO is not only useful for the legal domain, but also potentially any text domain that cites to external sources and values comprehensiveness of citations. For instance, when generating a biography (Min et al., 2023), a comprehensive biography that covers all the keypoints of one's life is desired. Likewise, the summarized keypoints of one's life can be seen as befitting a general notion of "citation", and the elaborate stories that support the keypoint are the "claim".

However, there are many assumptions about the case law data that would not hold on other domains. First, case law data has extractable and atomic citation and claims. Second, citations can be matched exactly. Third, there exists a unique solution to the bipartite matching problem of citation and claims. Thus, we also define a generalized L-FRESCO, FRESCO, extending its application across domains:

$$\text{FRESCO} = \frac{1}{|R|} \sum_{c_i \in C, r_j \in R} x_{ij} \text{B-SIM}(c_i, r_j)$$

Where the binary indicator x_{ij} can be found by the following integer program of maximum bipartite matching:

$$\begin{aligned} \text{Maximize} \quad & \sum_{c_i \in C, r_j \in R} x_{ij} \text{C-SIM}(c_i, r_j), \\ \text{subject to} \quad & \sum_{c_i \in C} x_{ij} = 1 \quad \forall r_j \in R, \\ & \sum_{r_j \in R} x_{ij} = 1 \quad \forall c_i \in C, \\ & x_{ij} \in \{0, 1\} \quad \forall c_i \in C, r_j \in R, \end{aligned}$$

where $x_{ij} = 1$ indicates that the i-th generated claim is matched with the j-th reference claim. Since we do not assume that citations can be exactly matched, FRESCO computes based on the best citation match, where $\text{C-SIM}(\cdot)$ is a low-cost, continuous similarity function for matching citations.

References

Chern, I.-C., Chern, S., Chen, S., Yuan, W., Feng, K., Zhou, C., He, J., Neubig, G., Liu, P., et al. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*, 2023.

Cushman, J., Dahl, M., and Lissner, M. eyecite: A tool for parsing legal citations. *Journal of Open Source Software*, 6(66):3617, 2021. URL https://doi.org/10.21105/joss.03617.

Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. GPT-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL https://api.semanticscholar.org/CorpusID:964287.

Liu, N. F., Zhang, T., and Liang, P. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848, 2023. URL https://api.semanticscholar.org/CorpusID:258212854.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-free? assessing the reliability of leading ai legal research tools. 2024. URL https://api.semanticscholar.org/CorpusID:269976547.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, 2023. URL https://arxiv.org/abs/2305.14251.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L. E., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. J. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://api.semanticscholar.org/CorpusID:246426909.

Wanner, M., Ebner, S., Jiang, Z., Dredze, M., and Durme, B. V. A closer look at claim decomposition. *ArXiv*, abs/2403.11903, 2024. URL https://api.semanticscholar.org/CorpusID:268531293.

Yuan, W., Neubig, G., and Liu, P. Bartscore: Evaluating generated text as text generation. *ArXiv*, abs/2106.11520, 2021. URL https://api.semanticscholar.org/CorpusID:235593404.

Yue, X., Wang, B., Chen, Z., Zhang, K., Su, Y., and Sun, H. Automatic evaluation of attribution by large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.307. URL https://aclanthology.org/2023.findings-emnlp.307.