Complete Evidence Extraction with Model Ensembles: A Case Study on Medical Coding

Katharina Beckh[1,2], **Sven Heuser**[1], **Stefan Rüping**[1]
[1]Fraunhofer IAIS
[2]Lamarr Institute
Correspondence: katharina.beckh@iais.fraunhofer.de

Abstract

High-stakes decisions informed by decision support systems require explicit evidence. While prior work focuses on short sufficient evidence, regulatory compliance and medical billing call for *complete* evidence: all relevant input tokens that support a decision. We formulate complete evidence extraction as a task and study it in a medical coding setting. Motivated by the Rashomon effect, we aggregate token-level evidence from multiple language models to increase evidence completeness. We perform a case study using existing equally-performing models, feature attributions, and a dataset with human-annotated evidence. Our results show that Rashomon ensembles significantly increase evidence recall while incurring only a small token overhead over individual models. Ensembles of only three models already outperform the best single model and recover information that individual models miss.

1 Introduction

In many high-stakes settings, stakeholders must be able to justify decisions. As decision support systems are more widely used for classification, these evidentiary requirements extend to the systems themselves. Evidence extraction methods address this need by identifying input features that support a model prediction.[1] Most prior work focuses on minimal *sufficient* justifications (Lei et al., 2016; Thorne et al., 2018; Luss and Dhurandhar, 2024), where evidence is short: often a single word or phrase at one position in the text.

However, in critical settings, such as regulatory compliance (Beckh et al., 2024; Schmitz et al., 2024) and clinical medicine (Noll et al., 2023; Li et al., 2024), stakeholders need *complete* evidence, which identifies all possible tokens at different positions in the document. For example, in psychiatric care, the number and severity of indicators, such as self-harm and aggression, determine the need for hospitalization and impact billing (Noll et al., 2023). If a decision-support system fails to retrieve relevant indicators and the threshold for an extended stay is not met, a patient might be discharged from the clinic too early – a risk for the patient and others.

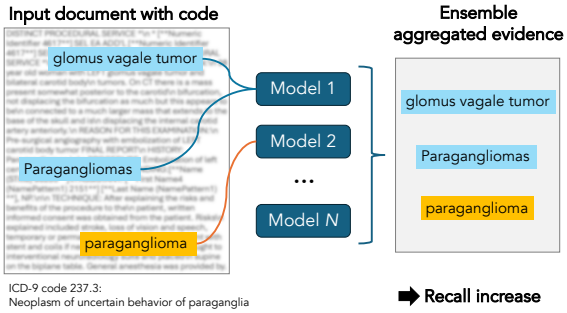


Figure 1: Illustration of complete evidence extraction on the medical coding task. An ensemble approach, aggregating the evidence of several models, leads to higher recall and, thus, more complete evidence.

While sufficient evidence extraction is model-centric and concerned with faithfully explaining model predictions, completeness is data-centric and demands the full set of cues available in the input, which is especially hard to realize in long text. Rather than tackling this objective with a single model, we draw inspiration from work utilizing multiple models (Kobylińska et al., 2024; Cavus et al., 2025; Hsu et al., 2025): Different models achieve similar classification performance while relying on distinct solutions, a phenomenon referred to as the Rashomon effect (Breiman, 2001; Müller et al., 2023; Rudin et al., 2024). If individual models attend to different valid cues, aggregating their evidence may increase coverage of human-annotated evidence. This intuition also connects to observed benefits of diversity in clinical

[1]Evidence does not necessarily provide insight into *how* a prediction was reached (Tan, 2022), hence we use the more specific term evidence instead of explanation or rationale.

decision making (Yuan et al., 2026). We therefore explore whether ensembles can recover more complete evidence than single models.

For this, we present a case study investigating complete evidence extraction for a medical coding task (Figure 1). The dataset consists of long clinical notes with human-annotated evidence. Using existing domain-specific models and feature attribution scores (Edin et al., 2024), we build a Rashomon ensemble formed by multiple equally-accurate models and aggregate evidence extracted from all models in the ensemble.

Our contributions are as follows:

- We formulate and empirically study complete evidence extraction in a medical coding task with human-annotated evidence.
- We demonstrate that Rashomon ensembles significantly improve evidence recall over single models with only a small token overhead.
- We provide empirical guidance on ensemble design: evidence-guided training improves precision but not recall; ensembles of already three models outperform any single model.

2 Background

Feature attribution methods assign importance scores to input features with respect to a model prediction (Ribeiro et al., 2016; Sundararajan et al., 2017). Li et al. (2024) formalize completeness as the extent to which all predictive features are captured. Their study was conducted in controlled synthetic settings where predictive features are known. In text classification, the true predictive features are latent. We therefore approximate them with human-annotated evidence and focus on complete evidence as introduced by Cheng et al. (2023): all text spans that support a label.[2]

Operationally, we equate completeness with high recall of human-annotated complete evidence. Prior work exploits model diversity for more trustworthy explanations or clinical decisions (Kobylińska et al., 2024; Yuan et al., 2026), but does not study completeness of textual evidence. Here, we study Rashomon ensembles of same-architecture language models that differ only by random seed, holding the attribution method fixed, and analyze whether aggregating evidence improves completeness in an applied setting.

[2]The notion of *comprehensiveness* (DeYoung et al., 2020) is similar, but completeness as understood here does not directly assume that evidence removal leads to a reduction in model confidence.

3 Methods

We perform a case study on clinical coding investigating evidence retrieval from several models. We compare ensembles to single models measuring recall of human-annotated evidence and evidence similarity between models. The evaluation is on evidence recall, not on classification performance.

Task and Dataset We consider the classification task of assigning medical codes to free-text clinical notes as the basis task and focus our evaluation on the newly framed complete evidence extraction task. MDACE (Cheng et al., 2023) is a medical dataset based on MIMIC-III (Goldberger et al., 2000; Johnson et al., 2016). It contains electronic health records in the English language with clinical codes (ICD-9). For each code, evidence is provided in the form of annotated text spans, e.g., code 416.8 (other pulmonary heart disease) has 'pulmonary hypertension' as one evidence span.

MDACE documents are annotated in a sufficient and also a complete style, in which *all* text that is relevant for a code is annotated. The complete subset has roughly three times more evidence. Discharge summaries (6,000 tokens on average) serve as data basis. We follow the dataset split by Cheng et al. (2023) of 181 train, 60 validation, 61 test hospital admissions (586 test documents). In contrast to most prior work, we utilize the subset with complete evidence for testing. Filtering for cases with more than one evidence span yields 17 test cases for recall analysis, limiting generalizability. Due to the small amount of data, we additionally use the whole test set with 586 test cases (with only sufficient evidence) to analyze how stable the results are measuring evidence similarity of models and number of unique tokens. To our knowledge, MDACE is the only available dataset offering complete evidence annotations for long clinical text.

Models We use existing model weights from Edin et al. (2024). The underlying architecture is a transformer, encoder-based, trained on medical text and fine-tuned on MDACE. From the models, we selected two training regimes to investigate whether training style affects ensemble performance:

- Input Gradient Regularization (IGR): no use of evidence labels, but gradient penalty to reduce irrelevant token importance.
- Evidence-Guided Training (EGT): uses human evidence spans as supervision to guide the model's attention to those tokens.

Train type	Metric	Single model	Ensemble
IGR	*recall*	0.60 (±0.25)	0.87 (±0.23)
	precision	0.70 (±0.21)	0.49 (±0.18)
EGT	*recall*	0.63 (±0.26)	0.86 (±0.24)
	precision	0.74 (±0.22)	0.57 (±0.24)

Table 1: Mean recall and precision values (+ standard deviation) for single models and the Rashomon ensemble comprising 10 models, for input gradient regularization (IGR) and evidence-guided training (EGT).

Each approach contains 10 seeds from random initializations, leading to a total of 20 models.

Ensembles For each approach, IGR and EGT, Rashomon ensembles consist of the 10 respective models. We aggregate the evidence of all models, i.e., the union of extracted tokens. For the exploration of optimal ensemble size, we report scores for all possible model combinations.

Evidence Extraction To retrieve evidence, a feature attribution method is employed. We use AttInGrad scores from Edin et al. (2024) which had the highest faithfulness and plausibility metrics on MDACE in prior work (Edin et al., 2024; Beckh et al., 2025). The decision threshold was set based on a validation set.

Metrics Using human-annotated evidence as ground truth, we compute token-level recall $R = \frac{|G \cap M|}{|G|}$ and precision $P = \frac{|G \cap M|}{|M|}$, where G and M are the sets of ground-truth and predicted token IDs. Since missing evidence is costlier than checking extra tokens, we focus on recall as our completeness measure. We assess recall improvements via 10,000-fold bootstrap resampling of test cases, reporting mean recall differences between the ensemble and the best single model with percentile-based 95% confidence intervals and Bonferroni-corrected p-values. Evidence overlap of models in ensembles is measured by pairwise Jaccard similarity.

4 Results

Single model vs. ensemble Table 1 displays mean recall and precision scores for the single models and the ensemble approach under both training regimes. The ensembles achieve higher recall than the single models. Bootstrap resampling shows that these gains are statistically significant: $\Delta_{\text{IGR}} = 0.16$ (95% CI [0.06, 0.26], $p_{\text{corr}} = 0.0036$) and $\Delta_{\text{EGT}} = 0.22$ (95% CI [0.11, 0.32], $p_{\text{corr}} < 0.0002$).

	Evidence similarity		Unique tokens	
	full test set	*complete test set*	*full test set*	*complete test set*
IGR	0.52 (±0.12)	0.52 (±0.15)	10.58 (±4.9)	11.06 (±3.98)
EGT	0.55 (±0.06)	0.57 (±0.07)	10.27 (±5.69)	9.88 (±4.30)

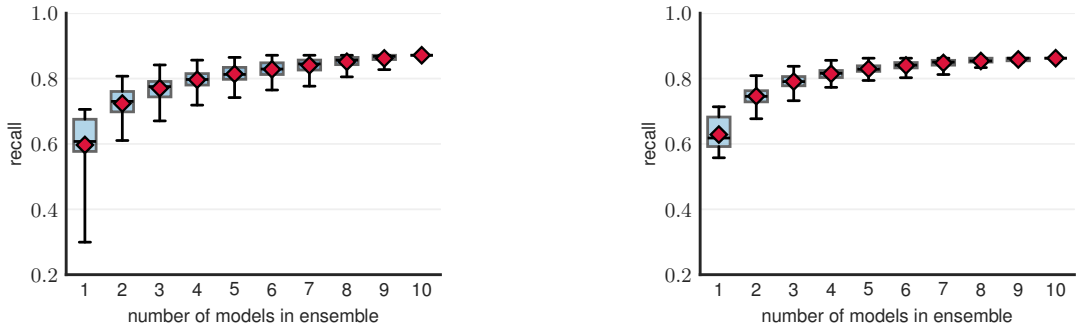
Table 2: Evidence similarity of the 10 models in the ensemble measured by mean pairwise Jaccard similarity, number of unique evidence tokens in the ensemble on full MDACE test set and subset with complete evidence.

As evidence is aggregated in the ensemble approach, precision is reduced due to the additional evidence tokens. When comparing the training regimes, EGT shows higher precision scores, but no recall improvement for the ensemble. Thus, using evidence cues during training may improve precision but has no notable effect on recall.

Ensemble size Figure 2 shows mean, minimum, and maximum recall values for each ensemble size. For each size, we report all possible model combinations, e.g., for size 2 we analyze 45 model pairs. With each additional model, more evidence information is retrieved, unless it is a very low-performing combination. The marginal utility gains of added models decline, i.e., they are most pronounced in the beginning and reduce with increasing ensemble size. Adding only one model already leads to a recall increase of roughly 0.1. IGR and EGT show similar recall patterns. One difference is that IGR has lower minimum values which is likely due to an outlier model with worse performance. From 3 models onward (4 in the case of IGR), even the lowest performing combination has higher recall than the best single model.

Evidence similarity Table 2 shows evidence similarity measured by pairwise Jaccard similarity and unique token counts for the full test set and the subset with complete evidence. Evidence similarity is between 0.52 and 0.57 indicating a substantial evidence overlap between the models. The agreement and the number of unique tokens are similar on both test sets, confirming the same evidence patterns for more data.

Qualitative insights When inspecting data points, we anecdotally find that EGT models extract more clinically relevant tokens, whereas IGR models more frequently highlight function words



(a) IGR (b) EGT

Figure 2: Evidence recall for different ensemble sizes for all possible model combinations. Red diamonds show mean value, whiskers indicate minimum and maximum value, i.e., worst and best performing model combination.

and punctuation. In one case of low recall, the human-annotated evidence 'glomus vagale tumor' (code 237.3) could not be reliably retrieved. Nearly all models identified meaningful tokens 'paragangli', 'vagus', and subword 'omus', but none captured 'tumor'. This may be because 'tumor' occurs in many codes and is a *descriptive*, not a *discriminative* feature. We also observed under-annotation in the human-annotated spans: in several cases, the ensemble discovered valid supporting spans that were not present in the annotation, suggesting imperfect human-annotated evidence (also see Khadka et al., 2026). Thus, ensemble precision may be underestimated and higher than reported.

5 Discussion

Our main result is that aggregated evidence from Rashomon ensembles significantly increased evidence recall in a medical coding task. This validates that different models retrieve meaningful evidence that single models could not. The diversity in evidence patterns supports this: pairwise Jaccard similarity between models is moderate (0.52–0.57), suggesting substantial overlap in retrieved tokens but also systematic differences. Aggregating evidence across models leverages this diversity. Our findings are an empirical instance of benefiting from the Rashomon effect (Breiman, 2001).

Practically, the results imply that when exhaustive coverage is required, Rashomon ensembles are a useful strategy, provided that measures are implemented to handle the increase in false positives. For this task, an ensemble consisting of 10 models retrieves 10-11 tokens per document. This number is small relative to the typical length of discharge summaries (around 6,000 tokens). The overhead therefore is acceptable in environments where it is costly to miss evidence. Our results are in line with similar work on mixed-vendor systems (Yuan et al., 2026). While prior work focused on LLMs, our case study shows recall benefits on small domain-specific language models from different initializations.

Limitations and future work Here, we simply aggregate model evidence by taking the union of evidence tokens. For consensus finding, we envision more sophisticated methods which keep relevant cues while also improving precision. Our study focused on encoder models and a single feature attribution method, which may not be representative of the whole landscape of possible evidence extraction approaches for transformer architectures. The analysis was limited by the amount of available data, restricting generalizability. We addressed this by applying the ensemble approach to more test data with only sufficient evidence, showing that the same patterns of evidence similarity and number of tokens emerge. We hypothesize that this transfers to similar settings with long texts and potentially beyond: for scientific discovery, ensembles may help surface weak or rare signals. A promising direction for future work is to steer model diversity directly during training (Pang et al., 2019).

6 Conclusion

We presented the first systematic case study of complete evidence extraction for clinical coding, focusing on recovering all relevant evidence tokens. Aggregating evidence from multiple models significantly increased recall compared to single models with only a small token overhead. Ensembles of already three models outperform any single model, suggesting that exploiting model diversity is a suitable strategy when exhaustive coverage is required.

Acknowledgments We would like to thank Elisa Studeny for support in data processing. We are grateful to Sebastian Müller, Vanessa Toborek, and the NLU team for valuable discussions.

References

Katharina Beckh, Joann Rachel Jacob, Adrian Seeliger, Stefan Rüping, and Najmeh Mousavi Nejad. 2024. Limitations of feature attribution in long text classification of standards. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 10–17.

Katharina Beckh, Elisa Studeny, Sujan Sai Gannamaneni, Dario Antweiler, and Stefan Rueping. 2025. The anatomy of evidence: An investigation into explainable ICD coding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16840–16851, Vienna, Austria. Association for Computational Linguistics.

Leo Breiman. 2001. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231.

Mustafa Cavus, Jan N. van Rijn, and Przemysław Biecek. 2025. Beyond the single-best model: Rashomon partial dependence profile for trustworthy explanations in AutoML. In *International Conference on Discovery Science*, volume 16090 of *Lecture Notes in Computer Science*, pages 445–459. Springer.

Hua Cheng, Rana Jafari, April Russell, Russell Klopfer, Edmond Lu, Benjamin Striner, and Matthew Gormley. 2023. MDACE: MIMIC documents annotated with code evidence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7534–7550, Toronto, Canada. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. 2024. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4869–4890, Miami, Florida, USA. Association for Computational Linguistics.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Ethan Hsu, Tony Cao, Lesia Semenova, and Chudi Zhong. 2025. The Rashomon set has it all: Analyzing trustworthiness of trees under multiplicity. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA*.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

Supriya Khadka, Xiaorui Jiang, and Vasile Palade. 2026. LLM-assisted clinical coding audit through an interpretable coding pipeline. *International Journal of Medical Informatics*, page 106605.

Katarzyna Kobylińska, Mateusz Krzyziński, Rafał Machowicz, Mariusz Adamek, and Przemysław Biecek. 2024. Exploration of the Rashomon set assists trustworthy explanations for medical data. *IEEE Journal of Biomedical and Health Informatics*, 28(11):6454–6465.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Yawei Li, Yang Zhang, Kenji Kawaguchi, Ashkan Khakzar, Bernd Bischl, and Mina Rezaei. 2024. A dual-perspective approach to evaluating feature attribution methods. *Transactions on Machine Learning Research*.

Ronny Luss and Amit Dhurandhar. 2024. When stability meets sufficiency: Informative explanations that do not overwhelm. *Transactions on Machine Learning Research*.

Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. 2023. An empirical evaluation of the Rashomon effect in explainable machine learning. In *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy*, pages 462–478. Springer.

Samuel Noll, Sarah Haag, Rémi Guidon, and Simon Hölzer. 2023. A new case-mix based payment system for the psychiatric day care sector in Switzerland: proposed methods for developing the tariff structure. *Health Policy*, 131:104797.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. 2019. Improving adversarial robustness via promoting ensemble diversity. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4970–4979. PMLR.

Marco Túlio 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, San Francisco, CA, USA*, pages 1135–1144. ACM.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo I. Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. 2024. Position: Amazing things come from having many good models. In *Forty-first International Conference on Machine Learning, ICML 2024*.

Anna Schmitz, Rebekka Görge, Elena Haedecke, Marion Borowski, Adrian Seeliger, and Maximilian Poretschkin. 2024. *Towards Formalising AI Readiness of Standards*, pages 209–231. T.M.C. Asser Press, The Hague.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW*, pages 3319–3328. PMLR.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. 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)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Grace Chang Yuan, Xiaoman Zhang, Sung Eun Kim, and Pranav Rajpurkar. 2026. Do mixed-vendor multi-agent LLMs improve clinical diagnosis? In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 1–18, Rabat, Morocco. Association for Computational Linguistics.