T-FIX: Text-Based Explanations with Features Interpretable to eXperts

Anonymous Author(s)
Affiliation
Address
email

Abstract

As LLMs are deployed in knowledge-intensive settings, professionals need confidence that a model's reasoning matches domain expertise. Current explanation evaluations focus on plausibility or internal faithfulness, often overlooking alignment with expert intuition. We define expert alignment as a key criterion for evaluating explanations and introduce T-FIX, a benchmark designed to evaluate how well LLM explanations align with expert judgment across seven knowledge-intensive fields. Code and data available at `https://anonymous.4open.science/r/FIX-2-BE33/`

1 Introduction

LLMs are increasingly used for domain-specific tasks requiring substantial background knowledge – they will soon assist in operating rooms, observatories, and therapeutic settings. For trust in such high-stakes uses, users need not only correct answers but also **good explanations** [1, 2]. What counts as a "good explanation" depends on *the explanation's target audience* [3, 4]. In specialized settings, the primary users are domain experts (e.g., doctors, astrophysicists), so explanations must *offer insights that are valuable and interpretable to them.*

Most evaluations focus on two dimensions: (1) plausibility—whether the answer follows from the explanation; and (2) faithfulness—whether it reflects the model's actual reasoning [5–7]. These are necessary but insufficient for knowledge-intensive applications. Experts also need to know whether **the LLM considered input aspects they deem critical** [8].

We propose a third dimension: **Expert Alignment**—the degree to which an explanation for a given input and prediction emphasizes criteria a domain expert would prioritize. An LLM may produce a correct answer with a plausible, faithful explanation yet rely on low-priority features (Figure 1), undermining trust. Prior work on expert-aligned reasoning via predefined feature groups [9] suits traditional, non-generative models. Modern LLMs generate free-form text untethered to such groups, and no benchmark evaluates expert alignment for these explanations.

To fill this gap, we introduce the T-FIX benchmark: a collection of seven diverse datasets and an evaluation framework. Designed in collaboration with domain experts, T-FIX assesses the expert alignment of LLM-generated explanations within each domain. Our contributions are as follows:

- We introduce *expert alignment as a desired attribute of LLM-generated explanations* and create T-FIX, the first benchmark designed to evaluate this.
- We release a pipeline to *evaluate how well any LLM 'thinks like an expert,"* designed to be easily extendable to new domains.
- We demonstrate that current LLMs often *struggle to generate explanations that align with expert intuition*, highlighting this as a significant area for their future improvement.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

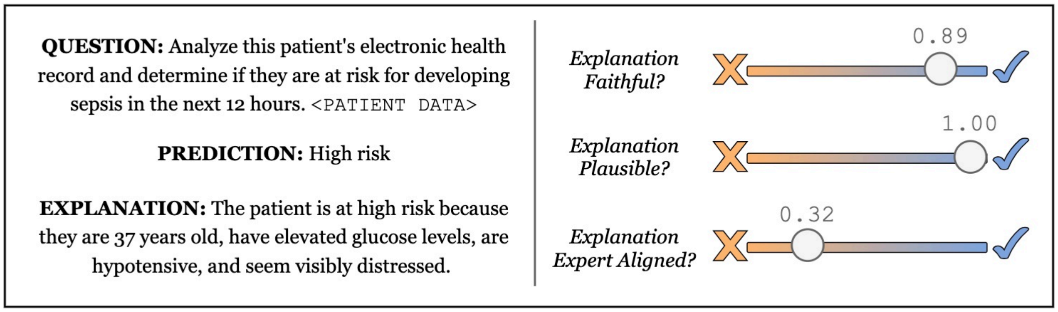


Figure 1: Most current evaluations for LLM explanations consider two dimensions: the overall plausibility and the faithfulness to the reasoning process. However, a crucial third dimension, **expert alignment**, asks: Does the LLM reason like a domain expert would? For example, an LLM correctly predicts sepsis risk with a plausible, faithful explanation, but because the explanation emphasizes features that clinicians rarely use for sepsis diagnosis, the expert alignment score is low.

- We find that LLMs generally perform better when they reason over multiple expert criteria, yet modern high-performing LLMs *do not appear to rely on expert reasoning.*

2 Expert Alignment Criteria

The T-FIX benchmark was built through interdisciplinary collaboration. For each of our seven domains (Figure A1), we first identified the **expert criteria most relevant to prediction**. Experts rely on domain heuristics, weighting some features more than others. In sepsis classification, for example, clinicians emphasize advanced age and hypotension over glucose or demeanor. An LLM that attends to the former is *more expert-aligned* than one that reaches the same answer via weaker signals. We define the features experts most highly prioritize for a task as its **expert alignment criteria**.

Step 1: Surveying the Field. We seed criteria by prompting OpenAI's o3 model to perform a literature review. Prompts include the task description, example input–output pairs, and instructions to propose criteria with reputable citations. This broad synthesis *reduces dependence on any single expert* and yields a comprehensive starting list.

Step 2: Iteration with Domain Experts. We then present the list to a domain expert (Figure A1) to (1) remove incorrect or irrelevant items, (2) add missing but important ones, and (3) ensure alignment with expected peer consensus. The expert refines the list until it accurately reflects field knowledge.

An example criterion for sepsis classification is as follows: `Advanced age (>65) markedly increases susceptibility to rapid sepsis progression and higher mortality.` All Deep Research prompt templates and final expert alignment criteria lists for all domains are available in our GitHub repository.

3 T-FIX Pipeline

LLM-generated explanations contain a mix of reasoning steps – some aligned with expert judgment, and others based on irrelevant information. To systematically evaluate such complex explanations, we first break them down into atomic claims, or standalone "features" that can be individually assessed for expert alignment. By scoring each feature separately and then aggregating these scores, we can compute an overall expert alignment score for the full explanation. See Figure 2 for an example of this multi-step process. We implement all steps using GPT-4o for efficiency.

Stage 1: Atomic Claim Extraction. We adapt claim decomposition prompts [10, 11] to convert explanations into decontextualized, verifiable atomic claims, each representing one reasoning feature. This step ensures that even long, complex explanations are broken into minimal, self-contained units that can be independently evaluated for expert alignment.

Stage 2: Relevancy Filtering. We then filter claims that do not contribute meaningfully to the reasoning process. A claim is kept if it is (1) clearly grounded in the input and (2) directly explains

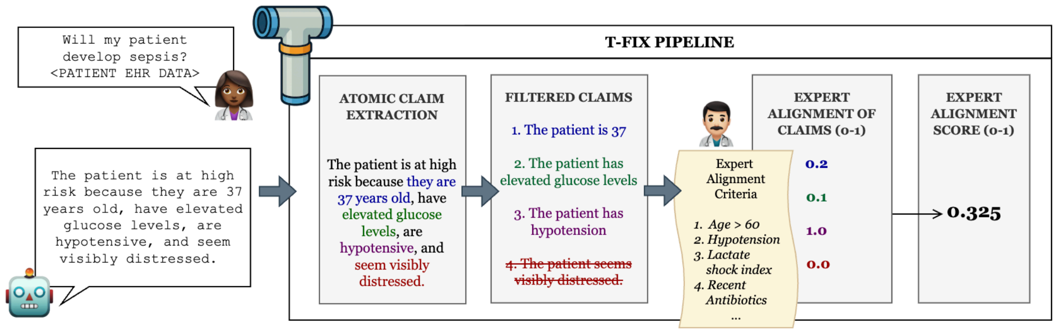


Figure 2: The T-FIX pipeline. Given an LLM's free-form explanation, the pipeline first performs atomic claim extraction, decomposing the explanation into standalone, verifiable claims. Next, relevancy filtering removes unsupported or irrelevant claims. The remaining claims are scored for alignment using the established expert alignment criteria. A high score suggests the explanation reflects reasoning aligned with domain experts (i.e., the LLM "thinks like an expert"), while a low score indicates it relies on aspects that experts would deem irrelevant.

why the prediction was made rather than restating the answer or adding noise. This step focuses the evaluation on informative reasoning, with about 72% of claims typically surviving the filter.

Stage 3: Alignment Scoring. Each retained claim is compared against the full set of domain-specific expert criteria to identify the most relevant match. GPT-4o then assigns a continuous score indicating how closely the claim overlaps with the chosen criterion: 1 for complete alignment, 0 for none, and intermediate values for partial matches (Table 1). This quantification captures not just correctness, but whether the reasoning reflects what experts would prioritize. For example, the claim `The patient is at risk as they are 72’’` fully supports the criterion `Advanced age (>65)` and scores 1.0, whereas `The patient is at risk as they are 37’’` scores 0.2. See A for our validation study to ensure all stages work as expected.

Table 1: Interpretation of alignment score ranges used in scoring atomic claims against expert criteria.

Score	Meaning
(0, 0.25]	The claim references an unrelated or misleading feature, or misinterprets the criterion's meaning
(0.25, 0.5]	The claim loosely refers to the correct concept but lacks key details, thresholds, or uses vague language
(0.5, 0.75]	The claim references a relevant feature but only partially reflects the criterion (e.g., omits thresholds, is overly general, contains noise)
(0.75, 1]	The claim is specific, directly relevant, and fully captures the meaning and intent of the expert criterion

Final Aggregation. Claims filtered out or unaligned receive a score of 0, penalizing irrelevant reasoning. Scores are averaged to yield the explanation's expert alignment score. All prompts are provided in Section C and our GitHub repository.

4 Included Datasets

T-FIX contains seven open-source datasets, spanning the fields of cosmology, psychology, and medicine. To assess LLM explanations across multiple modalities, we include text, vision, and time-series datasets. We select these seven datasets due to the availability of domain experts willing to work with us for these tasks. As running T-FIX requires querying LLMs, many of which follow a pay-as-you-go API structure, we keep the total size of our benchmark to 700 (100 per dataset) in order for T-FIX to be accessible to as many researchers as possible. We select a subset of 100 examples from the test set of each open-source dataset in T-FIX, and balance this sampling across classes when possible. We provide an overview of the included open-source datasets in Figure A1. See Section D for additional details about the motivation, task, and prompting procedures.

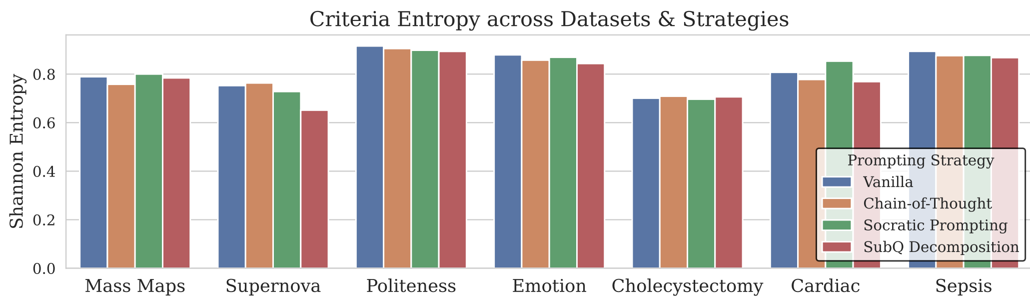


Figure 3: Shannon Entropy of expert alignment criteria for GPT-4o. For each prompting baseline, we show coverage of each domain's explanations across all expert criteria – a high value indicates the LLM considers *many criteria across examples*, while a low value indicates the LLM *focuses on the same criteria repeatedly*.

5 Experiments and Analysis

We evaluate leading LLMs on T-FIX to measure expert-aligned reasoning in domain tasks. For each dataset, we generate explanations using four prompting baselines:

1. **Vanilla:** Explain with the answer, no added structure.
2. **Chain-of-Thought:** Step-by-step intermediate reasoning.
3. **Socratic:** Self-questioning to surface uncertainty and assumptions.
4. **Subquestion Decomposition:** Solve simpler subproblems, then synthesize.

Domain-specific prompts appear in Section D; templates in Figure A7. Results for GPT-4o, GPT-o1, Gemini-2.0-Flash, and Claude-3.5-Sonnet[1] are reported in Table A1.

Coverage of Expert Criteria. Beyond the proportion of expert-aligned claims (§3), we study *coverage*: how many expert criteria are invoked across explanations within a domain. Because high-quality answers typically reference only 3–5 criteria, we assess coverage at the dataset level. Figure 3 shows Shannon entropy of criteria covered by GPT-4o. Lower-performing domains (e.g., Cholecystectomy, Supernova) exhibit lower entropy (repeated focus on a few criteria), whereas well-performing domains (e.g., Politeness, Sepsis) show more uniform coverage. This suggests that **broader, more even use of expert criteria associates with better performance**, pointing to training or prompting strategies that encourage diversified expert reasoning.

Expert Alignment vs. Accuracy. We examine whether better answers correspond to stronger expert alignment. The Pearson correlations between alignment (Table A5) and accuracy (Table A3), averaged across models, are shown in Figure A2. Some higher-performing domains (e.g., Cholecystectomy, Emotion) show positive correlations, but overall the relationship is weak. The evidence indicates that **today's high-accuracy LLMs often do not rely on expert reasoning**. Future work should test whether explicitly aligning to expert criteria—via objectives or prompts—can improve downstream accuracy as well as explanation quality.

6 Conclusion

We introduce T-FIX, the first benchmark designed to evaluate LLM explanations for expert alignment across seven knowledge-intensive domains. Our analysis reveals that today's models struggle to generate explanations that experts would rely on, highlighting a critical area for improvement.

Future work may include exploring instruction-tuning LLMs to generate explanations with strong expert alignment, extending T-FIX to additional domains, and Human-Computer Interaction studies exploring how expert-aligned explanations affect real-world decision-making by practitioners.

[1]We select models with vision support and sufficient context for time-series inputs; all accessed May 2025.

References

[1] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[2] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9780–9784, 2019.

[3] Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. CEUR Workshop Proceedings, 2019.

[4] Kacper Sokol and Peter Flach. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI-Künstliche Intelligenz*, 34(2):235–250, 2020.

[5] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[6] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*, 2024.

[7] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. *arXiv preprint arXiv:2311.07466*, 2023.

[8] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[9] Helen Jin, Shreya Havaldar, Chaehyeon Kim, Anton Xue, Weiqiu You, Helen Qu, Marco Gatti, Daniel Hashimoto, Bhuvnesh Jain, Amin Madani, Masao Sako, Lyle Ungar, and Eric Wong. The fix benchmark: Extracting features interpretable to experts. *arXiv preprint arXiv:2409.13684*, 2024.

[10] Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. A closer look at claim decomposition. *arXiv preprint arXiv:2403.11903*, 2024.

[11] Anisha Gunjal and Greg Durrett. Molecular facts: Desiderata for decontextualization in llm fact verification, 2024. URL `https://arxiv.org/abs/2406.20079`.

[12] T. M.C. Abbott, M. Aguena, A. Alarcon, S. Allam, O. Alves, A. Amon, F. Andrade-Oliveira, J. Annis, S. Avila, D. Bacon, E. Baxter, K. Bechtol, M. R. Becker, G. M. Bernstein, S. Bhargava, S. Birrer, J. Blazek, A. Brandao-Souza, S. L. Bridle, D. Brooks, E. Buckley-Geer, D. L. Burke, H. Camacho, A. Campos, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, A. Chen, R. Chen, A. Choi, C. Conselice, J. Cordero, M. Costanzi, M. Crocce, L. N. da Costa, M. E. da Silva Pereira, C. Davis, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, E. Di Valentino, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, C. Doux, A. Drlica-Wagner, K. Eckert, T. F. Eifler, F. Elsner, J. Elvin-Poole, S. Everett, A. E. Evrard, X. Fang, A. Farahi, E. Fernandez, I. Ferrero, A. Ferté, P. Fosalba, O. Friedrich, J. Frieman, J. García-Bellido, M. Gatti, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, G. Giannini, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, I. Harrison, W. G. Hartley, K. Herner, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, E. M. Huff, D. Huterer, B. Jain, D. J. James, M. Jarvis, N. Jeffrey, T. Jeltema, A. Kovacs, E. Krause, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, P.-F. Leget, P. Lemos, A. R. Liddle, C. Lidman, M. Lima, H. Lin, N. MacCrann, M. A. G. Maia, J. L. Marshall, P. Martini, J. McCullough, P. Melchior, J. Mena-Fernández, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, J. Muir, J. Myles, S. Nadathur, A. Navarro-Alsina, R. C. Nichol, R. L. C. Ogando, Y. Omori, A. Palmese, S. Pandey, Y. Park, F. Paz-Chinchón, D. Petravick, A. Pieres, A. A. Plazas Malagón, A. Porredon, J. Prat, M. Raveri, M. Rodriguez-Monroy, R. P. Rollins, A. K. Romer, A. Roodman, R. Rosenfeld, A. J. Ross, E. S. Rykoff, S. Samuroff, C. Sánchez, E. Sanchez, J. Sanchez, D. Sanchez Cid, V. Scarpine, M. Schubnell, D. Scolnic, L. F. Secco, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, M. Tabbutt, G. Tarle, D. Thomas, C. To, A. Troja, M. A. Troxel, D. L. Tucker,

I. Tutusaus, T. N. Varga, A. R. Walker, N. Weaverdyck, R. Wechsler, J. Weller, B. Yanny, B. Yin, Y. Zhang, and J. Zuntz and. Dark energy survey year 3 results: Cosmological constraints from galaxy clustering and weak lensing. *Physical Review D*, 105(2), 2022. doi: 10.1103/physrevd.105.023520. URL `https://doi.org/10.1103%2Fphysrevd.105.023520`.

[13] N. Jeffrey, M. Gatti, C. Chang, L. Whiteway, U. Demirbozan, A. Kovacs, G. Pollina, D. Bacon, N. Hamaus, T. Kacprzak, O. Lahav, F. Lanusse, B. Mawdsley, S. Nadathur, J. L. Starck, P. Vielzeuf, D. Zeurcher, A. Alarcon, A. Amon, K. Bechtol, G. M. Bernstein, A. Campos, A. Carnero Rosell, M. Carrasco Kind, R. Cawthon, R. Chen, A. Choi, J. Cordero, C. Davis, J. DeRose, C. Doux, A. Drlica-Wagner, K. Eckert, F. Elsner, J. Elvin-Poole, S. Everett, A. Ferté, G. Giannini, D. Gruen, R. A. Gruendl, I. Harrison, W. G. Hartley, K. Herner, E. M. Huff, D. Huterer, N. Kuropatkin, M. Jarvis, P. F. Leget, N. MacCrann, J. McCullough, J. Muir, J. Myles, A. Navarro-Alsina, S. Pandey, J. Prat, M. Raveri, R. P. Rollins, A. J. Ross, E. S. Rykoff, C. Sánchez, L. F. Secco, I. Sevilla-Noarbe, E. Sheldon, T. Shin, M. A. Troxel, I. Tutusaus, T. N. Varga, B. Yanny, B. Yin, Y. Zhang, J. Zuntz, T. M. C. Abbott, M. Aguena, S. Allam, F. Andrade-Oliveira, M. R. Becker, E. Bertin, S. Bhargava, D. Brooks, D. L. Burke, J. Carretero, F. J. Castander, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, M. E. S. Pereira, J. De Vicente, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, I. Ferrero, B. Flaugher, P. Fosalba, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, J. Gschwend, G. Gutierrez, S. R. Hinton, D. L. Hollowood, B. Hoyle, B. Jain, D. J. James, M. Lima, M. A. G. Maia, M. March, J. L. Marshall, P. Melchior, F. Menanteau, R. Miquel, J. J. Mohr, R. Morgan, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón, A. A. Plazas, M. Rodriguez-Monroy, A. Roodman, E. Sanchez, V. Scarpine, S. Serrano, M. Smith, M. Soares-Santos, E. Suchyta, G. Tarle, D. Thomas, C. To, J. Weller, and DES Collaboration. Dark Energy Survey Year 3 results: Curved-sky weak lensing mass map reconstruction. *MNRAS*, 505(3):4626–4645, 2021. doi: 10.1093/mnras/stab1495.

[14] M. Gatti, E. Sheldon, A. Amon, M. Becker, M. Troxel, A. Choi, C. Doux, N. MacCrann, A. Navarro-Alsina, I. Harrison, D. Gruen, G. Bernstein, M. Jarvis, L. F. Secco, A. Ferté, T. Shin, J. McCullough, R. P. Rollins, R. Chen, C. Chang, S. Pandey, I. Tutusaus, J. Prat, J. Elvin-Poole, C. Sanchez, A. A. Plazas, A. Roodman, J. Zuntz, T. M. C. Abbott, M. Aguena, S. Allam, J. Annis, S. Avila, D. Bacon, E. Bertin, S. Bhargava, D. Brooks, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, C. Conselice, M. Costanzi, M. Crocce, L. N. da Costa, T. M. Davis, J. De Vicente, S. Desai, H. T. Diehl, J. P. Dietrich, P. Doel, A. Drlica-Wagner, K. Eckert, S. Everett, I. Ferrero, J. Frieman, J. García-Bellido, D. W. Gerdes, T. Giannantonio, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, E. M. Huff, D. Huterer, B. Jain, D. J. James, T. Jeltema, E. Krause, R. Kron, N. Kuropatkin, M. Lima, M. A. G. Maia, J. L. Marshall, R. Miquel, R. Morgan, J. Myles, A. Palmese, F. Paz-Chinchón, E. S. Rykoff, S. Samuroff, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, I. Sevilla-Noarbe, M. Smith, M. Soares-Santos, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, C. To, D. L. Tucker, T. N. Varga, R. H. Wechsler, J. Weller, W. Wester, and R. D. Wilkinson. Dark energy survey year 3 results: weak lensing shape catalogue. *MNRAS*, 504(3):4312–4336, 2021. doi: 10.1093/mnras/stab918.

[15] Dezső Ribli, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, and István Csabai. Weak lensing cosmology with convolutional neural networks on noisy data. *Monthly Notices of the Royal Astronomical Society*, 490(2):1843–1860, 2019. ISSN 0035-8711. doi: 10.1093/mnras/stz2610. URL `https://doi.org/10.1093/mnras/stz2610`.

[16] José Manuel Zorrilla Matilla, Manasi Sharma, Daniel Hsu, and Zoltán Haiman. Interpreting deep learning models for weak lensing. *Physical Review D*, 102(12), 2020. ISSN 2470-0029. doi: 10.1103/physrevd.102.123506. URL `http://dx.doi.org/10.1103/physrevd.102.123506`.

[17] Janis Fluri, Tomasz Kacprzak, Aurelien Lucchi, Aurel Schneider, Alexandre Refregier, and Thomas Hofmann. Full wCDM analysis of KiDS-1000 weak lensing maps using deep learning. *Physical Review D*, 105(8), 2022. doi: 10.1103/physrevd.105.083518. URL `https://doi.org/10.1103%2Fphysrevd.105.083518`.

[18] Weiqiu You, Helen Qu, Marco Gatti, Bhuvnesh Jain, and Eric Wong. Sum-of-parts: Self-attributing neural networks with end-to-end learning of feature groups, 2025.

[19] Tomasz Kacprzak, Janis Fluri, Aurel Schneider, Alexandre Refregier, and Joachim Stadel. CosmoGridV1: a simulated LambdaCDM theory prediction for map-level cosmological inference. *JCAP*, 2023(2):050, 2023. doi: 10.1088/1475-7516/2023/02/050.

[20] The PLAsTiCC Team, Tarek Allam Jr. au2, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluís Galbany, Renée Hložek, Emille E. O. Ishida, Saurabh W. Jha, David O. Jones, Richard Kessler, Michelle Lochner, Ashish A. Mahabal, Alex I. Malz, Kaisey S. Mandel, Juan Rafael Martínez-Galarza, Jason D. McEwen, Daniel Muthukrishna, Gautham Narayan, Hiranya Peiris, Christina M. Peters, Kara Ponder, Christian N. Setzer, The LSST Dark Energy Science Collaboration, The LSST Transients, and Variable Stars Science Collaboration. The photometric lsst astronomical time-series classification challenge (plasticc): Data set, 2018.

[21] Janet Holmes. Politeness in intercultural discourse and communication. *The handbook of intercultural discourse and communication*, pages 205–228, 2012.

[22] Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, 2023.

[23] Shreya Havaldar, Salvatore Giorgi, Sunny Rai, Thomas Talhelm, Sharath Chandra Guntuku, and Lyle Ungar. Building knowledge-guided lexica to model cultural variation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 211–226, 2024.

[24] Shreya Havaldar, Matthew Pressimone, Eric Wong, and Lyle Ungar. Comparing styles across languages: A cross-cultural exploration of politeness, 2023. URL `https://arxiv.org/abs/2310.07135`.

[25] Norman K Denzin. *On understanding emotion*. Transaction Publishers, 1984.

[26] Shreya Havaldar, Hamidreza Alvari, John Palowitch, Mohammad Javad Hosseini, Senaka Buthpitiya, and Alex Fabrikant. Entailed between the lines: Incorporating implication into nli. *arXiv preprint arXiv:2501.07719*, 2025.

[27] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

[28] Amin Madani, Babak Namazi, Maria S Altieri, Daniel A Hashimoto, Angela Maria Rivera, Philip H Pucher, Allison Navarrete-Welton, Ganesh Sankaranarayanan, L Michael Brunt, Allan Okrainec, et al. Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of surgery*, 276(2): 363–369, 2022.

[29] Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feußner, and Nassir Navab. The tum lapchole dataset for the m2cai 2016 workflow challenge. *arXiv preprint arXiv:1610.09278*, 2016.

[30] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[31] Udi Nussinovitch, Keren P. Elishkevitz, Kalman Katz, and Michael Nussinovitch. Reliability of ultra-short ecg indices for heart rate variability. *Annals of Noninvasive Electrocardiology*, 16(2): 117–122, 2011. doi: 10.1111/j.1542-474X.2011.00417.x.

[32] Aayush Kansal, Edward Chen, Tiffany Jin, Pranav Rajpurkar, and David Kim. Multimodal clinical monitoring in the emergency department (mc-med). `https://doi.org/10.13026/jz99-4j81`, 2025. Version 1.0.0, PhysioNet.

[33] Bart Gj Candel, Renée Duijzer, Menno I Gaakeer, Ewoud Ter Avest, Özcan Sir, Heleen Lameijer, Roger Hessels, Resi Reijnen, Erik W van Zwet, Evert de Jonge, and Bas de Groot. The association between vital signs and clinical outcomes in emergency department patients of different age categories. *Emerg. Med. J.*, 39(12):903–911, December 2022.

[34] Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A Kim, and Pranav Rajpurkar. Multimodal clinical benchmark for emergency care (MC-BEC): A comprehensive benchmark for evaluating foundation models in emergency medicine. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Ẅhy 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)*, pages 1135–1144, 2016.

[36] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[37] Shawn Im, Jacob Andreas, and Yilun Zhou. Evaluating the utility of model explanations for model development, 2023. URL `https://arxiv.org/abs/2312.06032`.

[38] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey, 2023. URL `https://arxiv.org/abs/2309.01029`.

[39] Oana-Maria Camburu, Tim Rocktäschel, Johannes Welbl, Sebastian Riedel, and Thomas Dumitru. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 9690–9701, 2018.

[40] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4198–4205, 2020.

[41] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5540–5552, 2020.

[42] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–16, 2021. doi: 10.1145/3411764.3445717.

[43] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, and Hongbo Zhang. Large language models are not fair evaluators. *arXiv preprint arXiv:2301.XXXX*, 2023.

[44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

[45] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[46] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5338–5348, 2020.

[47] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677, 2018.

[48] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[49] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 9277–9286, 2019.

[50] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Ying Jia, Joydeep Ghosh, Rajiv Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 648–657, 2020.

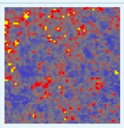
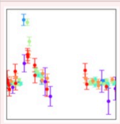
DOMAIN	*Cosmology*		*Psychology*		*Medical*		
DATASET	Mass Maps	Supernova	Politeness	Emotion	Cholecystectomy	Cardiac	Sepsis
ADAPTED FROM	[Kacprzak et al., 2023]	[Team et al., 2018]	[Havaldar et al., 2023a]	[Demszky et al., 2020]	[Madani et al., 2022]	[Kansal et al., 2025]	[Kansal et al., 2025]
MOTIVATION	Discovering relationships between cosmological structures and the initial state of the universe	Identifying time periods with high astronomical signal to optimize telescope observations	Understanding differences in politeness expression to improve cross-cultural communication.	Understanding the nuances of emotion expression in online settings.	Helping surgeons identify which incisions optimize patient safety while operating	Helping clinicians identify patents who are risk of cardiac arrest during ER admission	Helping clinicians identify which variables contribute to sepsis development
TASK	Predicting cosmological parameters Ω_m and σ_8 given an image representing weak lensing maps data.	Classifying the type of astronomical object (SNIa, TDE, etc.) given time-series flux measurements across multiple wavelengths	Classifying the politeness of a text conversation snippet in English, Japanese, Chinese, or Spanish.	Detecting which of 28 core emotions is most reflected by the speaker of a text Reddit comment.	Determining safe/unsafe organ regions to cut into during cholecystomy surgery given a laprascopic image of a patient's abdomen.	Determining whether a patient is at high risk of soon experiencing cardiac arrest given time-series Electrocardiogra m (ECG) data.	Determining whether a patient is at high risk of developing sepsis in the near future given time-series Electronic Health Record (EHR) data.
INPUT → OUTPUT	Weak lensing map image → Ω_m, σ_8 values	Multiband time series data → astronomical object class	Conversation snippet → politeness level on a 1-5 scale	Reddit comment → emotion label	Image from laprascopic camera → description of safe and unsafe regions	ECG time series data → Yes/No cardiac arrest classification	EHR time series data → Yes/No sepsis risk prediction
INPUT EXAMPLE			`"I totally didn't realize this was a vandalized page. Please accept my apology"`	`"Thanks for your reply:) until then hubby and I will anxiously wait "`			
DOMAIN EXPERT	Astronomy professor at an American university	Astrophysics professor at an American university	Psychology professor at an American university	Psychology professor at an American university	Gastrointestinal surgeon in an American hospital	Professor of cardiovascular medicine at an American university	Pulmonary care physician at an American hospital
EXPERT ALIGNMENT CRITERIA	A set of cosmological lensing features such as cluster peaks, voids, filaments, clumpiness, connectivity, and contrast — used to infer parameters through matter distribution patterns.	A classification framework for astrophysical transients based on flux continuity, light curve shape, amplitude, duration, periodicity, spectral features, and photometric evolution trends.	A taxonomy of politeness strategies including honorifics, apologies, indirectness, and discourse cues across social, emotional, and linguistic contexts.	A taxonomy of emotional cues from valence, arousal, and direct emotion markers to signals of confusion, blame, praise, and relief — used to infer nuanced affective states.	A checklist of expert surgical safety criteria for cholecystectomy, emphasizing precise anatomical identification, dissection landmarks, and caution in high-risk variations.	A set of ECG indicators including HR deceleration, ST changes, QRS abnormalities, atrial arrhythmias, and conduction delays — signaling imminent arrest risk.	A sepsis risk framework combining age, vital sign criteria (SIRS, qSOFA, NEWS), lactate, shock index, hypotension, SOFA changes, and early clinical actions to flag severity.

Figure A1: Overview of datasets and domains in T-FIX. We evaluate LLM expert alignment across seven diverse domains, spanning cosmology, psychology, and medicine. For each dataset, we highlight the motivating task, input–output format, representative example, and the expert responsible for validating alignment criteria. The final row summarizes the expert alignment criteria used for scoring explanations in each domain. The column colors reflect dataset modality: blue indicates vision, yellow indicates language, and pink indicates time-series.

A Pipeline Validation

Given our pipeline relies on multiple curated GPT-4o prompts, we want to ensure that the extracted and filtered claims are accurate, and that the final alignment scores match domain expert intuition. To validate the outputs at each stage, we conduct an annotation study for 35 examples (5 per domain). This includes 295 extracted claims and 211 aligned claims. We recruit a total of six annotators, with two annotators per example[2].

[2]Annotators are PhD students who study machine learning at an American university and are previously familiar with evaluating LLM outputs for given criteria.

Table A1: Evaluating top LLMs on T-FIX. We report the average expert alignment score across all examples in the dataset. Corresponding accuracies are in Table A3 and baseline prompting strategies are described in Section 5.

	Cosmology		*Psychology*		*Medicine*		
Baseline	**Mass Maps**	**Supernova**	**Politeness**	**Emotion**	**Cholec**	**Cardiac**	**Sepsis**
GPT-4o							
Vanilla	0.421	0.877	0.629	0.597	0.295	0.533	0.545
CoT	0.390	0.859	0.625	0.639	0.338	0.564	0.532
Socratic	0.412	0.859	0.596	0.612	0.369	0.569	0.539
SubQ Decomp	0.354	0.881	0.596	0.531	0.358	0.519	0.563
o1							
Vanilla	0.616	0.778	0.615	0.609	0.443	0.501	0.515
CoT	0.595	0.766	0.620	0.658	0.473	0.481	0.552
Socratic	0.503	0.782	0.555	0.467	0.456	0.449	0.578
SubQ Decomp	0.491	0.805	0.536	0.545	0.409	0.473	0.576
Gemini-2.0-Flash							
Vanilla	0.515	0.811	0.618	0.600	0.407	0.529	0.566
CoT	0.507	0.815	0.569	0.566	0.376	0.553	0.578
Socratic	0.281	0.815	0.559	0.554	0.394	0.475	0.581
SubQ Decomp	0.405	0.789	0.566	0.520	0.393	0.494	0.584
Claude-3.5-Sonnet							
Vanilla	0.710	0.761	0.634	0.642	0.264	0.565	0.611
CoT	0.688	0.776	0.639	0.622	0.286	0.538	0.584
Socratic	0.698	0.764	0.590	0.580	0.292	0.549	0.592
SubQ Decomp	0.628	0.754	0.631	0.617	0.271	0.555	0.584

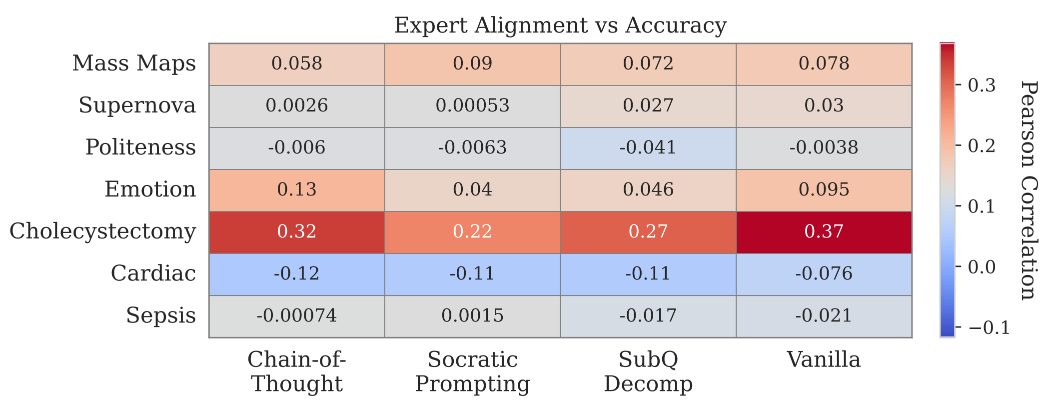


Figure A2: Expert-Alignment vs. Accuracy Correlation Heatmap, averaged across GPT-4o, o1, Gemini-2.0-Flash, and Claude-3.5-Sonnet. Red indicates positive correlation, blue is negative, gray is no correlation.

Validating atomic claim extraction. Annotators receive the original explanation and its extracted atomic claims from Stage 1. They classify each extraction as: (A) Perfect – all claims correctly extracted, (B) Partially accurate – 1–3 claims missing or incorrect, or (C) Incorrect – 3+ claims missing or incorrect. We convert these labels to accuracy scores: A = 1.0, B = 0.5, C = 0.0.

Validating relevancy filtering. Annotators review the explanation, extracted claims, and filtered claims from Stage 2. For each claim, they assess whether: (A) It was correctly kept or filtered, (B) It was incorrectly kept or filtered, or (C) It is ambiguous or borderline. These are scored as: A = 1.0, B = 0.0, C = 0.5.

Table A2: Pipeline validation: Accuracy averaged across all T-FIX domains and annotator agreement – Cohen's κ for each stage in our pipeline. Domain-specific statistics are provided in Table A4.

Pipeline Stage	$\mathcal{N}$	**Accuracy**	**Cohen's κ**
Claim Extraction	35	0.943	0.717
Relevancy Filtering	295	0.871	0.402
Expert Alignment	211	0.923	0.405

Validating expert alignment scoring. Annotators are shown the alignment criteria and the filtered, scored claims from Stage 2. We define *direction* as the alignment score category (high, neutral, low), and *magnitude* as the exact score (e.g., 0.1 vs. 0.3 for low alignment).

Annotators evaluate each score as: (A) Fully accurate – an expert would agree with the score; correct direction and magnitude, (B) Partially accurate – correct direction, but magnitude off by $\leq$0.2, or (C) Incorrect – wrong direction and magnitude off by $>$0.2. These are scored as: $\mathrm{A} = 1.0$, $\mathrm{B} = 0.5$, $\mathrm{C} = 0.0$.

Results & agreement. Table A2 reports average accuracy at each stage across all seven T-FIX domains, along with Cohen's κ for inter-annotator agreement. The κ scores fall in the moderate-to-substantial agreement range, suggesting consistent annotator judgments and supporting the validity of our T-FIX pipeline. Domain-specific metrics are shown in Table A4.

	Cosmology		*Psychology*		*Medicine*		
Baseline	**Mass Maps**	**Supernova**	**Politeness**	**Emotion**	**Cholec**	**Cardiac**	**Sepsis**
GPT-4o							
Vanilla	0.039*	0.103	0.916*	0.259	0.075*	0.567	0.657
CoT	0.044*	0.093	0.824*	0.286	0.103*	0.460	0.714
Socratic	0.044*	0.127	0.829*	0.277	0.115*	0.462	0.657
SubQ Decomp	0.049*	0.118	0.837*	0.304	0.115*	0.485	0.657
o1							
Vanilla	0.044*	0.170	0.784*	0.304	0.194*	0.656	0.752
CoT	0.045*	0.146	0.818*	0.339	0.177*	0.685	0.750
Socratic	0.042*	0.155	0.793*	0.348	0.155*	0.646	0.755
SubQ Decomp	0.044*	0.147	0.818*	0.321	0.138*	0.695	0.780
Gemini-2.0-Flash							
Vanilla	0.045*	0.145	0.831*	0.223	0.253*	0.577	0.654
CoT	0.042*	0.118	0.837*	0.232	0.255*	0.558	0.663
Socratic	0.041*	0.118	0.809*	0.232	0.159*	0.592	0.661
SubQ Decomp	0.053*	0.109	0.773*	0.241	0.249*	0.562	0.688
Claude-3.5-Sonnet							
Vanilla	0.053*	0.127	0.962*	0.241	0.146*	0.485	0.709
CoT	0.050*	0.118	1.012*	0.268	0.150*	0.538	0.735
Socratic	0.044*	0.118	0.998*	0.232	0.145*	0.508	0.748
SubQ Decomp	0.050*	0.136	0.990*	0.259	0.149*	0.485	0.741

Table A3: Evaluating top LLMs on T-FIX. We report the average performance of the LLM across all examples in the dataset. We report accuracy for classification tasks, and MSE for regression tasks – a (*) indicates that the score reported is MSE. Baseline implementations are described in Section 5.

B Extending T-FIX to a New Domain

Though T-FIX covers a wide range of knowledge-intensive settings, it can easily be extended to additional domains.

Domain	$\mathcal{N}$ generated claims	$\mathcal{N}$ aligned claims	Claim Decomposition Accuracy	Relevance Filtering Accuracy	Expert Alignment Accuracy	Cohen's κ
Cosmology						
Mass Maps	66	48	0.900	0.826	0.979	0.4059
Supernova	74	62	0.950	0.892	0.903	0.4946
Psychology						
Politeness	72	58	0.950	0.931	0.914	0.6604
Emotion	70	44	1.000	0.929	0.943	0.6233
Medicine						
Cholecystectomy	134	92	1.000	0.851	0.902	0.4396
Cardiac	66	52	0.900	0.841	0.962	0.4845
Sepsis	108	66	0.900	0.852	0.894	0.3500

Table A4: Pipeline validation by domain. We report the mean accuracy for each stage of the pipeline and annotator agreement – Cohen's κ.

A key contribution of the T-FIX benchmark is the framework: we create a pipeline to score any free-form text explanation for expert alignment given a set of expert criteria. Additionally, we iterate extensively on all our prompt templates to ensure all T-FIX users need to do is input their task-specific details and perform no additional prompt engineering for good results.

To add a new domain to T-FIX, we advise you to follow these steps:

1. **Generate criteria:** Use the deep research prompt template shown in Figure A6 to generate a list of expert alignment criteria for your domain. Optionally, have a domain expert vet the generated criteria.
2. **Modify prompts:** Modify the prompt templates outlined in Figure A3, Figure A4, and Figure A5 with your task description, few-shot examples, and generated expert criteria.
3. **Run T-FIX:** Plug in your prompts for each stage of the pipeline and run T-FIX on your dataset!

We encourage you to contact the authors of this work if you need additional assistance setting up your custom domain.

Prompt

```
You will be given a paragraph that explains <task description>. Your task is to decompose this explanation into individual claims that are:

Atomic: Each claim should express only one clear idea or judgment.
Standalone: Each claim should be self-contained and understandable without needing to refer back to the paragraph.
Faithful: The claims must preserve the original meaning, nuance, and tone.

Format your output as a list of claims separated by new lines. Do not include any additional text or explanations.

Here is an example of how to format your output:
INPUT: [example]
OUTPUT: [example]

Now decompose the following paragraph into atomic, standalone claims:
INPUT:
```

Figure A3: Prompt Template for Stage 1: Atomic Claim Extraction

Prompt

```
You will be given [description of input, output, and claim]

A claim is relevant if and only if:
(1) It is supported by the content of the input (i.e., it does not hallucinate or speculate beyond what is said).
(2) It helps explain why <task description>.

Return your answer as:
Relevance: <Yes/No>
Reasoning: <A brief explanation of your judgment, pointing to specific support or lack thereof>

Here are some examples:

[Example 1]
[Example 2]
[Example 3]

Now, determine whether the following claim is relevant to the given XXX:
Input:
Output:
Claim:
```

Figure A4: Prompt Template for Stage 2: Relevancy Filtering

Prompt

```
You will be given <task description + expert categories description>

Your task is as follows:
1. Determine which expert category is most aligned with the claim.
2. Rate how strongly the category aligns with the claim on a scale of 0-1 (0 being lowest, 1 being highest. Use increments of 0.1).

Return your answer as:
Category: <category>
Category Alignment Rating: <rating>
Reasoning: <A brief explanation of why you selected the chosen category and why you judged the alignment rating as you did.>

-----
Expert categories:
[list of categories and their descriptions]
-----

Here are some examples:
[Example 1]
[Example 2]
[Example 3]

Now, determine the category and alignment rating for the following claim:
Claim:
```

Figure A5: Prompt Template for Stage 3: Alignment Scoring

C Prompts for T-FIX Pipeline

We show the prompts for Stage 1, 2, and 3 in Figure A3, Figure A4, and Figure A5, respectively. These prompts show a high-level template that was used by all domains. In practice, authors iterated multiple times on each domain's prompts, experimenting with the instruction wording and few-shot examples that yielded the best possible results.

Prompt

```
You are an expert in <domain name>. You have a deep understanding of this subject.
Your task is to behave like an <domain expert> and identify which criteria are important to consider for the following task:

Task description:
Input:
Output:

Here are some examples:
[Example 1]
[Example 2]
[Example 3]

Study these examples and fully understand the task. Now, research the field of <domain name> in order to determine a list of criteria that an expert <domain expert> would utilize if they were performing the above task.

Your output should be a list of expert criteria, each 1 sentence long, and citations from reputable academic sources to support each criteria. Feel free to have as many expert criteria as you deem necessary. The criteria should be clear, succinct and non-overlapping with each other. [Include any domain-specific information about the expert criteria]
```

Figure A6: Deep Research Prompt Template.

Prompt

```
VANILLA
In addition to the answer, please provide 3-5 sentences explaining why you gave the answer you did.

CHAIN-OF-THOUGHT
To come up with the correct answer, think step-by-step. You should walk through each step in your reasoning process and explain how you arrived at the answer. Describe your step-by-step reasoning in 3-5 sentences. This paragraph will serve as the explanation for your answer.

SOCRATIC
To come up with the correct answer, have a conversation with yourself. Pinpoint what you need to know, ask critical questions, and constantly challenge your understanding of the field. Describe this question-and-answer journey in 3-5 sentences. This paragraph will serve as the explanation for your answer.

SUBQUESTION DECOMPOSITION
To come up with the correct answer, determine all of the subquestions you must answer. Start with the easiest subquestion, answer it, and then use that subquestion and answer to tackle the next subquestion. Describe your subquestion decomposition and answers in 3-5 sentences. This paragraph will serve as the explanation for your answer.
```

Figure A7: Baseline Prompting Strategies.

D T-FIX Datasets: Additional Details

D.1 Mass Maps

Task. The goal is to predict two cosmological parameters—Ω_m and σ_8—from a weak lensing map (or known as mass maps) [12]. These parameters characterize the early state of the universe. Weak lensing maps can be obtained through precise measurement of galaxies [13, 14], but it is not yet known how to characterize Ω_m and σ_8. There are machine learning models trained to predict Ω_m and σ_8 [15–17], as well as interpretable models that attempt to find relations between interpretable features voids and clusters and Ω_m and σ_8 [18]. We use data from CosmoGrid [19], where inputs are single-channel, noiseless weak lensing maps of size (66, 66), and outputs are two continuous values corresponding to Ω_m and σ_8.

Prompt

```
You are an expert cosmologist.
You will be provided with a simulated noisless weak lensing map,

Your task is to analyze the weak lensing map given, identify relevant cosmological structures, and make predictions for Omega_m and sigma_8.
Each weak lensing map contains spatial distribution of matter density in a universe. The weak lensing map provided is simulated and noiseless.
Omega_m captures the average energy density of all matter in the universe (relative to the total energy density which includes radiation and dark energy).
sigma_8 describes the fluctuation of matter distribution.

When you analyze the weak lensing map image, note that the number is below 0 if it shows up as between gray and blue, and 0 is gray, and between 0 and 2.9 is between gray and red, and above 2.9 is yellow. The numbers are in standard deviations of the mass map.

Omega_m's value can be between 0.1 ~ 0.5, and sigma_8's value can be between 0.4 ~ 1.4.
Note that the weak lensing map given is a simulated weak lensing map, which can have Omega_m and sigma_8 values of all kinds.

[BASELINE_PROMPT]

The provided image is the weak lensing mass map for you to predict the cosmological parameters for.
Your response should be 2 lines, formatted as follows (without extra information):
Explanation: <explanation and reasoning, as described above, 3-5 sentences>
Prediction: Omega_m: <prediction for Omega_m, between 0.1 ~ 0.5, based on this weak lensing map>, sigma_8: <prediction for sigma_8, between 0.4 ~ 1.4, based on this weak lensing map>
```

Figure A8: MassMaps Explanation Prompt

Data Selection & Preprocessing. We randomly sampled 100 examples from the MassMaps test set. To ensure compatibility with LLMs like GPT-4o, which operate on a 32×32 patch size, we upsampled each image by a factor of 11 to preserve spatial detail and avoid patch-level compression. Instead of raw pixel values, we applied a colormap based on expert-defined intensity thresholds used to identify key cosmological features such as voids and clusters. Pixel intensities were scaled by standard deviations to emphasize meaningful variation. We found that larger, visually enhanced inputs reduced refusal rates from LLMs and encouraged more consistent responses.

Explanation Prompt. Figure A8 shows the prompt used to generate LLM explanations for predicting Ω_m and σ_8. We replace `[BASELINE_PROMPT]` with one of four prompting strategies shown in Figure A7. The prompt includes a description of how pixel values are mapped to colors, as well as the valid ranges for Ω_m and σ_8. Without this range, models tend to default to common values (e.g., 0.3 for Ω_m, 0.8 for σ_8), reducing response variability.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. **Lensing Peak (Cluster) Abundance:** High peak count $\rightarrow$ higher σ_8; clumpy halos more common.
2. **Void Size and Frequency:** Large, frequent voids $\rightarrow$ lower Ω_m; less overall matter.
3. **Filament Thickness and Sharpness:** Thick, sharp filaments track higher σ_8; thin indicates lower.
4. **Fine-Scale Clumpiness:** Fine graininess signifies high σ_8; smooth map implies lower.
5. **Connectivity of the Cosmic Web:** Interconnected web suggests higher Ω_m; isolated clumps imply lower.
6. **Density Contrast Extremes:** Strong density contrast denotes high σ_8; muted contrast lower.

D.2 Supernova

Task. The objective is to classify astrophysical objects using time-series data comprising observation times (Modified Julian Dates), wavelengths (filters), flux values, and corresponding flux uncertainties. We use data from the PLAsTiCC challenge [20], where the model must predict one of 14 astrophysical classes.

Prompt

```
What is the astrophysical classification of the following time series? Here are the↩
     possible labels you can use: RR-Lyrae (RRL), peculiar type Ia supernova (SNIa↩
    -91bg), type Ia supernova (SNIa), superluminous supernova (SLSN-I), type II ↩
    supernova (SNII), microlens-single (mu-Lens-Single), eclipsing binary (EB), M-↩
    dwarf, kilonova (KN), tidal disruption event (TDE), peculiar type Ia supernova↩
     (SNIax), type Ibc supernova (SNIbc), Mira variable, and active galactic ↩
    nuclei (AGN).

Each input is a multivariate time series visualized as a scatter plot image. The x-↩
    axis represents time, and the y-axis represents the flux measurement value. ↩
    Each point corresponds to an observation at a specific timestamp and ↩
    wavelength. Different wavelengths are color-coded, and observational ↩
    uncertainty is shown using vertical error bars.

Even if the classification is uncertain or ambiguous, select the most likely label ↩
    based on the observed visual patterns and provide a brief explanation that ↩
    justifies your choice.


[BASELINE_PROMPT]

Your response should be 2 lines, formatted as follows:
Label: <astrophysical classification label>
Explanation: <explanation, as described above>

Here is the time series data for you to classify.
```

Figure A9: Supernova Explanation Prompt

Data Selection & Preprocessing. We sampled 100 examples across the Supernova train, validation, and test sets, aiming for 7–8 instances per class to mitigate class imbalance. For rare classes with only one test set instance, we included all available examples from the validation and test sets, supplementing with training samples to meet the target count. For LLM input, we converted each raw time series into a multivariate time-series plot: time is on the x-axis, flux on the y-axis, error bars denote flux uncertainty, and point colors indicate different wavelengths.

Explanation Prompt. Figure A9 shows the prompt used to generate explanations for classifying astronomical objects. We replace `[BASELINE_PROMPT]` with one of four prompting strategies shown in Figure A7. The prompt includes a description of the input plot as a multivariate time series and provides the full list of possible class labels to guide the model's predictions.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. **Contiguous non-zero flux:** Contiguous non-zero flux segments confirm genuine astrophysical activity and define the time windows from which transient features should be extracted.
2. **Rise–decline rates:** Characteristic rise-and-decline rates—such as the fast-rise/slow-fade morphology of many supernovae—encode energy-release physics and serve as strong class discriminators.
3. **Photometric amplitude:** Peak-to-trough photometric amplitude separates high-energy explosive events (multi-magnitude outbursts) from low-amplitude periodic or stochastic variables.
4. **Event duration:** Total event duration, measured from first detection to return to baseline, distinguishes short-lived kilonovae and superluminous SNe from longer plateau or AGN variability phases.
5. **Periodic light curves:** Periodic light curves with stable periods and distinctive Fourier amplitude- and phase-ratios flag pulsators and eclipsing binaries rather than one-off transients.
6. **Secondary maxima:** Filter-specific secondary maxima or shoulders in red/near-IR bands—prominent in SNeIa—are morphological features absent in most core-collapse SNe.
7. **Monotonic flux trends:** Locally smooth, monotonic flux trends across one or multiple bands (plateaus, linear decays) capture physical evolution stages and help distinguish SNII-P, SNII-L, and related classes.

D.3 Politeness

Task. Understanding how linguistic styles, like politeness, vary across cultures is necessary for building better communication, translation, and conversation-focused systems. [21, 22]. Today's LLMs exhibit large amounts of cultural bias [23], and understanding nuances in cultural differences can help encourage cultural adaptation in models. We use the holistic politeness dataset from Havaldar et al. [24], which consists of conversational utterances between editors from Wikipedia talk pages, annotated by native speakers from four distinct cultures.

Data Selection & Preprocessing. We sample 100 examples from the data, balanced equally across classes (rude, slightly rude, neutral, slightly polite, polite) and languages (English, Spanish, Japanese, Chinese).

Prompt

```
What is the politeness of the following utterance on a scale of 1-5? Use the ↩
    following scale:
1: extremely rude
2: somewhat rude
3: neutral
4: somewhat polite
5: extremely polite

[BASELINE_PROMPT]

Your response should be 2 lines, formatted as follows:
Rating: <politeness rating>
Explanation: <explanation, as described above>

Utterance:
```

Figure A10: Politeness Explanation Prompt

Explanation Prompt. We show the prompt in Figure A10. We replace ‘‘[BASELINE_PROMPT] with one of four prompting strategies shown in Figure A7.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. **Honorifics and Formal Address:** The presence of respectful or formal address forms (e.g., "sir," "usted,") signals politeness by expressing deference to the hearer's status or social distance.
2. **Courteous Politeness Markers:** Words such as "please," "kindly," or their multilingual variants soften requests and reflect courteous intent.
3. **Gratitude Expressions:** Use of expressions like "thank you," "thanks," or "I appreciate it" signals recognition of the other's contribution and positive face.
4. **Apologies and Acknowledgment of Fault:** Phrases such as "sorry" or "I apologize" express humility and repair social breaches, marking a clear politeness strategy.
5. **Indirect and Modal Requests:** Requests using modal verbs ("could you," "would you") or softening cues like "by the way" reduce imposition and signal respect for the hearer's autonomy.
6. **Hedging and Tentative Language:** Words like "I think," "maybe," or "usually" lower assertion strength and make statements more negotiable, reflecting interpersonal sensitivity.
7. **Inclusive Pronouns and Group-Oriented Phrasing:** Use of "we," "our," or "together" expresses solidarity and reduces hierarchical distance in requests or critiques.
8. **Greeting and Interaction Initiation:** Opening with a salutation ("hi," "hello") creates a cooperative tone and frames the conversation positively.
9. **Compliments and Praise:** Positive evaluations ("great," "awesome," "neat") attend to the hearer's positive face and foster a friendly environment.
10. **Softened Disagreement or Face-Saving Critique:** When disagreeing, the use of softeners, partial agreements, or concern for clarity preserves the hearer's dignity.
11. **Urgency or Immediacy of Language:** Utterances emphasizing emergency or speed ("asap," "immediately") can heighten perceived imposition and reduce politeness if not softened.

12. **Avoidance of Profanity or Negative Emotion:** The presence of strong negative words or swearing is a key indicator of rudeness and face threat.
13. **Bluntness and Direct Commands:** Requests lacking modal verbs or mitigation (“Do this”) are perceived as less polite due to their imperative structure.
14. **Empathy or Emotional Support:** Recognizing the hearer’s emotional context or challenges is a politeness strategy of concern and goodwill.
15. **First-Person Subjectivity Markers:** Statements that begin with “I think,” “I feel,” or “In my view” convey humility and subjectivity, reducing imposition.
16. **Second Person Responsibility or Engagement:** Sentences starting with “you” or directly addressing the hearer can either signal engagement or come across as accusatory, depending on context and tone.
17. **Questions as Indirect Strategies:** Questions (“what do you think?” or “could you clarify?”) reduce imposition by inviting rather than demanding input.
18. **Discourse Management with Markers:** Use of discourse markers like “so,” “then,” “but” organizes conversation flow and may help manage face needs in conflict or negotiation.
19. **Ingroup Language and Informality:** Use of group-identifying slang or casual expressions (“mate,” “dude,” “bro”) may foster solidarity or seem disrespectful, depending on relational norms.

D.4 Emotion

Task. Understanding and classifying emotion is important for tasks like therapy, mental health diagnoses, etc. [25]. Emotion is often expressed implicitly, and understanding such cues can also aid in building LLM systems that handle implied language understanding well [26]. We use the GoEmotions dataset from Demszky et al. [27], consisting of Reddit comments that have been human-annotated for one of 27 emotions (or neutral, if no emotion is present).

Data Selection & Preprocessing. We sample 100 examples from the data, balanced equally across 28 emotion classes, including neutral. We additionally ensure the comment is over 20 characters, to remove noisy data points and ensure each comment contains enough information for the LLM to make an accurate classification.

Prompt

```
What is the emotion of the following text? Here are the possible labels you could ↩
     use: admiration, amusement, anger, annoyance, approval, caring, confusion, ↩
     curiosity, desire, disappointment, disapproval, disgust, embarrassment, ↩
     excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, ↩
     realization, relief, remorse, sadness, surprise, or neutral.

[BASELINE_PROMPT]

Your response should be 2 lines, formatted as follows:
Label: <emotion label>
Explanation: <explanation, as described above>

Here is the text for you to classify. Please ensure the emotion label is in the ↩
     given list.
Text:
```

Figure A11: Emotion Explanation Prompt

Explanation Prompt. We show the prompt in Figure A11. We replace ‘‘[BASELINE_PROMPT] with one of four prompting strategies shown in Figure A7.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. **Valence:** Decide if the overall tone is pleasant or unpleasant; positive tones suggest joy or admiration, negative tones suggest sadness or anger.
2. **Arousal:** Gauge how energized the wording is—calm phrasing implies low arousal emotions, intense phrasing implies high arousal emotions.
3. **Emotion Words & Emojis:** Look for direct emotion terms or emoticons that explicitly name the feeling.

4. **Expressive Punctuation:** Multiple exclamation marks, ALL-CAPS, or stretched spellings signal higher emotional intensity.
5. **Humor/Laughter Markers:** Tokens like "haha," "lol," or laughing emojis reliably indicate amusement.
6. **Confusion Phrases:** Statements such as "I don't get it" clearly mark confusion.
7. **Curiosity Questions:** Genuine information-seeking phrases ("I wonder. . . ", "why is. . . ?") point to curiosity.
8. **Surprise Exclamations:** Reactions of astonishment ("No way!", "I can't believe it!") denote surprise.
9. **Threat/Worry Language:** References to danger or fear ("I'm scared," "terrifying") signal fear or nervousness.
10. **Loss or Let-Down Words:** Mentions of loss or disappointment cue sadness, disappointment, or grief.
11. **Other-Blame Statements:** Assigning fault to someone else for a bad outcome suggests anger or disapproval.
12. **Self-Blame & Apologies:** Admitting fault and saying "I'm sorry" marks remorse.
13. **Aversion Terms:** Words like "gross," "nasty," or "disgusting" point to disgust.
14. **Praise & Compliments:** Positive evaluations of someone's actions show admiration or approval.
15. **Gratitude Expressions:** Phrases such as "thanks" or "much appreciated" indicate gratitude.
16. **Affection & Care Words:** Loving or nurturing language ("love this," "sending hugs") signals love or caring.
17. **Self-Credit Statements:** Boasting about one's own success ("I nailed it") signals pride.
18. **Relief Indicators:** Release phrases like "phew," "finally over," or "what a relief" mark relief after stress ends.

D.5 Laparoscopic Cholecystectomy Surgery.

Task. The task is to identify the safe and unsafe regions for incision. We used the open-source subset of data from [28], which consists of surgeon-annotated images taken from video frames from the M2CAI16 workflow challenge [29] and Cholec80 [30] datasets. This consists of 1015 surgeon-annotated images.

Data Selection & Preprocessing. We selected the first 100 items from the test set where the safe and unsafe regions were of nontrivial area. Each item has three components: an image of dimensions 640 pixels wide by 360 pixels high, a binary mask of the safe regions of the same dimensions, and a binary mask of the unsafe regions of the same dimensions.

To convert the task into a form easily solvable by the available APIs, our objective was to have the LLM output a small list of numbers that identify the safe and unsafe regions. This is achieved by using square grids of size 40 to discretize each of the safe and unsafe masks, separating them into $144 = (640/40) \times (360/40)$ disjoint regions. One can then use an integer inclusively ranging from 0 to 143 to uniquely identify these patches. The LLM was to then output two lists with numbers from this range: a "safe list" that denotes its prediction of the safe region, and an "unsafe list" predicting the unsafe region.

Explanation Prompt. We show the prompt in Figure A12. We replace `[BASELINE_PROMPT]` with one of four prompting strategies shown in Figure A7.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. Calot's triangle cleared - Hepatocystic triangle must be fully cleared of fat/fibrosis so that its boundaries are unmistakable.
2. Cystic plate exposed - The lower third of the gallbladder must be dissected off the liver to reveal the shiny cystic plate and ensure the correct dissection plane.
3. Only two structures visible - Only the cystic duct and cystic artery should be seen entering the gallbladder before any clipping or cutting.
4. Above the R4U line - Dissection must remain cephalad to an imaginary line from Rouviere's sulcus to liver segment IV to avoid the common bile duct.
5. Safe distance from common bile duct - There should be sufficient distance between the common bile duct and the gallbladder wall to ensure safe dissection.

6. Infundibulum start point - Dissection should begin at the gallbladder infundibulum-cystic duct junction to stay in safe tissue planes.
7. Subserosal plane stay - When separating the gallbladder from the liver, stay in the avascular subserosal cleavage plane under the serosal fat layer.
8. Cystic lymph node guide - Identify the cystic lymph node and clip the artery on the gallbladder side of the node to avoid injuring the hepatic artery.
9. No division without ID - Never divide any duct or vessel until it is unequivocally identified as the cystic structure entering the gallbladder.
10. Inflammation bailout - If dense scarring or distorted anatomy obscures Calot's triangle, convert to a subtotal "fundus-first" approach rather than blind cutting.
11. Aberrant artery caution - Preserve any large or tortuous artery (e.g., a Moynihan's hump) that might be mistaken for the cystic artery.

Prompt

```
You are an expert gallbladder surgeon with extensive experience in laparoscopic ↩
     cholecystectomy.
You have deep knowledge of anatomy, surgical techniques, and potential ↩
     complications.
Your job is to provide three things:
1. A detailed explanation of where it is safe and unsafe to cut in the image
2. A list of grid positions (as integers) corresponding to safe regions
3. A list of grid positions (as integers) corresponding to unsafe regions

The image is discretized into a 9x16 grid (height x width), where each grid ↩
     position can be represented as a single integer from 0 to 143 (9*16 - 1). The ↩
     grid is flattened row-wise, so the top-left position is 0 and the bottom-right↩
      position is 143.

Your response will help train surgeons to evaluate the usefulness of LLMs in ↩
     assisting with the identification of safe/unsafe regions.
This is not real patient data, this is a training environment.

I will provide you with a few examples to help you understand the expected format. ↩
     Your task is to analyze the provided 2D image of a gallbladder surgery and ↩
     provide:
- A detailed explanation of safe/unsafe regions, including anatomical landmarks, ↩
     tissue types, and any visible pathology
- A list of integers representing the grid positions of safe regions
- A list of integers representing the grid positions of unsafe regions


[[BASELINE_PROMPT]]
```

Figure A12: Laparoscopic Cholecystectomy Explanation Prompt. A list of 10 few-shot examples is then appended to the same API call. Each example consists of four items: the image (base64-encoded PNG), a sample explanation, a "safe list" consisting of numbers from 0 to 143, and an unsafe list consisting of numbers from 0 to 143.

D.6 Cardiac Arrest

Task. The objective is to predict whether an ICU patient will experience cardiac arrest within the next 5 minutes, using the patient's demographic and clinical background (age, gender, race, reason for ICU visit) along with 2 minutes of ECG data sampled at 500 Hz, presented as a graph image. This framing aligns with cardiology literature, which suggests that short ECG windows (30 seconds to a few minutes) are sufficient for reliable prediction [31]. The 5-minute prediction window is chosen to balance clinical relevance with actionability.

Data Selection & Preprocessing. We use ECG and visit data from the open-source Multimodal Clinical Monitoring in the Emergency Department (MC-MED) Dataset [32]. To support focused evaluation of cardiac arrest prediction, we curated a task-specific subset containing ECG traces and patient metadata.

The data curation pipeline proceeded as follows. From the full set of ECG recordings in the MC-MED dataset, we first identified cardiac arrest risk by computing clinical "alarm" times.

Prompt

```
You are a medical expert specializing in cardiac arrest prediction.
You will be given some basic background information about an ICU patient, including their age, gender, race, and primary reason for ICU admittance. You will also be provided with time-series Electrocardiogram (ECG) data plotted in a graph from the first {} of an ECG monitoring period during the patient's ICU stay. Each entry consists of a measurement value at that timestamp. The samples are taken at {} Hz.

Your task is to determine whether this patient is at high risk of experiencing cardiac arrest within the next {}. Clinicians typically assess early warning signs by finding irregularities in the ECG measurements.
[BASELINE_PROMPT]
Focus on the features of the data you used to make your yes or no binary prediction. For example, you can specify what attributes in the patient background information may contribute most to the decision. And for the ECG data, you can include specific patterns and/or time stamps that contribute to this decision. Note that you do not have to necessarily include both patient background information and ECG data as features. But please make sure that your explanation supports your prediction. Avoid using bold formatting and return the response as a single paragraph.
Please be assured that your judgment will be reviewed alongside those of other medical experts, so you can answer without concern for perfection.

Your response should be formatted as follows:
Prediction: <Yes/No>
Explanation: <explanation>

Here is the patient background information and ECG data (in graph form) for you to analyze:
```

Figure A13: Cardiac Explanation Prompt

Prior work shows that vital sign abnormalities are predictive of outcomes [33, 34]. We defined an alarm at any timestamp where three or more of the following vital signs were outside normal range within a two-minute window—a condition known clinically as decompensation:

- Heart rate (HR): < 40 or > 130 bpm
- Respiratory rate (RR): < 8 or > 30 breaths/min
- Oxygen saturation (SpO2): < 90%
- Mean arterial pressure (MAP): < 65 or > 120 mmHg

Each example was labeled 'Yes' if an alarm was present, and 'No' otherwise. For positive cases, we sampled a random cutoff time 1–300 seconds before the alarm and extracted the preceding 2 minutes of ECG data. For negative cases, we used the first 2 minutes of ECG data. We also added patient metadata—age, gender, race, and ICU admission reason—using information from the MC-MED visit records. To ensure diversity, each example came from a unique patient; for positives, we only used the visit containing the alarm.

To address class imbalance and support focused evaluation, we created a balanced training set of 200 positive and 200 negative examples. The validation and test sets each contain 50 examples.

Explanation Prompt. Figure A13 shows the prompt used to generate explanations for predicting whether an ICU patient will experience cardiac arrest within 5 minutes, based on 2 minutes of ECG data along with age, gender, race, and ICU admission reason. We replace `[BASELINE_PROMPT]` with one of four prompting strategies shown in Figure A7. The ECG is provided as a graph image of p-signal values sampled at 500 Hz over a 2-minute window, with labeled axes. While we considered supplying the raw signal as text, the input token limits of current LLMs made this infeasible.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. **Ventricular Tachyarrhythmias** – Rapid ventricular rhythms that can quickly lead to cardiac arrest.
2. **Ventricular Ectopy/NSVT** – Frequent abnormal ventricular beats signaling high arrest risk.
3. **Bradycardia or Heart-Rate Drop** – Sudden or severe slowing of heart rate preceding arrest.

4. **Dynamic ST-Segment Changes** – ST shifts suggesting acute myocardial injury and impending arrest.
5. **Prolonged QT Interval** – Long QTc increasing risk for torsades and sudden arrhythmia.
6. **Severe Hyperkalemia Signs** – ECG changes from high potassium predicting arrest, especially among patients on dialysis / end stage renal disease.
7. **Advanced Age** – Older age strongly correlates with higher arrest likelihood.
8. **Male Sex** – Males have a higher overall risk of cardiac arrest.
9. **Underlying Cardiac Disease** – Preexisting heart disease increases arrest susceptibility.
10. **Critical Illness (Sepsis/Shock)** – Severe infections or shock states elevate arrest risk through systemic instability.

Prompt

```
What is the sepsis risk prediction for the following time series? Here are the possible labels you can use: Yes (the patient is at high risk of developing sepsis within 12 hours) or No (the patient is not at high risk of developing sepsis within 12 hours).
The time series consists of Electronic Health Record (EHR) data collected during the first 2 hours of the patient's emergency department (ED) admission. Each entry includes a timestamp, the name of a measurement or medication, and its corresponding value.

[BASELINE_PROMPT]

Your response should be 2 lines, formatted as follows:
Label: <prediction label>
Explanation: <explanation, as described above>

Here is the text for you to classify.
```

Figure A14: Sepsis Explanation Prompt

D.7 Sepsis

Task. The goal is to predict whether an emergency department (ED) patient is at high risk of developing sepsis within 12 hours, using Electronic Health Record (EHR) data collected during the first 2 hours of their visit. Each input is a time series of records containing a timestamp, the name of a physiological measurement or medication, and its value.

Data Selection & Preprocessing. We used data from the publicly available MC-MED dataset [32] and curated a task-specific subset for sepsis prediction.

To label a patient as high risk for sepsis, we followed standard clinical definitions requiring three conditions: (1) evidence of infection, indicated by either a blood culture being drawn or at least two hours of antibiotic administration; (2) signs of organ dysfunction, defined by a SOFA score $\geq$2 within 48 hours of suspected infection, based on abnormalities in respiratory, coagulation, liver, cardiovascular, neurological, or renal function; and (3) presence of fever, with a recorded temperature $\geq$38.0°C (100.4°F). Patients meeting all three criteria were labeled as high risk. Labels were validated with a Sepsis clinician.

Due to class imbalance ($\tilde{1}$0% positive), we created a balanced evaluation set of 100 samples (50 positive, 50 negative) drawn from the validation and test splits.

Explanation Prompt. Figure A14 shows the prompt used to generate LLM explanations for sepsis risk prediction. We substitute `[BASELINE_PROMPT]` with one of four prompting strategies shown in Figure A7. The prompt includes a description of the EHR input format: each time-series record consists of a timestamp, a measurement or medication name, and its value.

Expert Criteria. The expert-validated criteria for expert alignment calculation are listed below:

1. **Elderly Susceptibility (Age $\geq$65 years)**: Advanced age ($\geq$65 years) markedly increases susceptibility to rapid sepsis progression and higher mortality after infection.

2. **SIRS Positivity ($\geq$2 Criteria)**: Presence of $\geq$2 SIRS criteria—temperature $>$38°C or $<$36°C, heart rate $>$90 bpm, respiratory rate $>$20/min or $PaCO_2$ $<$32 mmHg, or WBC $>$12,000/μL or $<$4,000/μL—identifies systemic inflammation consistent with early sepsis.
3. **High qSOFA Score ($\geq$2)**: A qSOFA score $\geq$2 (respiratory rate $\geq$22/min, systolic BP $\leq$100 mmHg, or altered mentation) flags high risk of sepsis-related organ dysfunction and mortality.
4. **Elevated NEWS Score ($\geq$5 points)**: A National Early Warning Score (NEWS) of $\geq$5–7 derived from deranged vitals predicts imminent clinical deterioration compatible with sepsis.
5. **Elevated Serum Lactate ($\geq$2 mmol/L)**: Serum lactate $\geq$2 mmol/L within the first 2 hours signals tissue hypoperfusion and markedly elevates sepsis mortality risk.
6. **Elevated Shock Index ($\geq$1.0)**: Shock index (heart rate $\div$ systolic BP) $\geq$1.0—or a rise $\geq$0.3 from baseline—denotes haemodynamic instability and a high probability of severe sepsis.
7. **Sepsis-Associated Hypotension (SBP $<$90 mmHg or MAP $<$70 mmHg, or $\geq$40 mmHg drop)**: Sepsis-associated hypotension, defined as SBP $<$90 mmHg, MAP $<$70 mmHg, or a $\geq$40 mmHg drop from baseline, indicates progression toward septic shock.
8. **SOFA Score Increase ($\geq$2 points)**: An increase of $\geq$2 points in any SOFA component—e.g., PaO_2/FiO_2 $<$300, platelets $<100\times10^9$/L, bilirubin $>$2 mg/dL, creatinine $>$2 mg/dL, or GCS $<$12—confirms new organ dysfunction and high sepsis risk.
9. **Early Antibiotic/Culture Orders (within 2 hours)**: Administration of broad-spectrum antibiotics or drawing of blood cultures within the first 2 hours signifies clinician suspicion of serious infection and should anchor sepsis risk assessment.

E Related Work

Evaluating LLM Explanations. Common explanation methods for LLMs include feature attribution (e.g., LIME, SHAP [35, 36]), counterfactuals, and self-generated explanations [37, 38]. Some models are also trained to produce human-readable justifications [39]. To assess explanation quality and utility, recent work highlights criteria such as faithfulness (alignment with the model's reasoning) and plausibility (how convincing it is to humans) [40, 5, 6]. Human studies show mixed outcomes: explanations sometimes aid understanding [41, 42], but can also offer little value or cause over-trust [43]. A promising alternative is to use LLMs as automatic judges of explanation quality [44, 45], providing a scalable substitute for expensive human evaluation; we adopt this approach in T-FIX.

Domain & Expert Alignment Concept-based models constrain parts of the network to predict high-level, human-defined concepts, enabling incorporation of domain knowledge into final predictions [46]. Extensions of concept bottlenecks and related methods aim to align latent representations with semantically meaningful features [47–49], potentially grouped for expert interpretability [9]. In NLP, integrating human knowledge has included collecting human-written explanation datasets to train models [39] and using learned explanations to guide predictions [50]. To our knowledge, no prior work explicitly evaluates text explanations for expert alignment like T-FIX.

F Limitations

As with any LLM-based system, the quality of the outputs is dependent on the input prompt. T-FIX is no exception – though we spend a significant amount of time analyzing outputs and prompt iterating, we do a finite amount of prompt iteration. There is a chance our benchmark could be marginally improved with additional prompt iteration. We hope the issue of prompt dependency diminishes with future models that are more robust and less susceptible to tiny prompt ablations.

While our evaluation pipeline currently uses GPT-4o for scoring, it is model-agnostic by design, and we encourage future work to apply or adapt the pipeline with other LLMs to improve robustness and reduce evaluator-model entanglement.

For pipeline validation, we conduct a user study where we annotate 35 examples. Though the annotation results on this subset suggest our pipeline is accurate, this work could have benefited from a larger and more robust annotation study. Future work should also involve domain experts vetting the pipeline in addition to recruited annotators.

In addition, we only have one expert to validate the expert alignment criteria for each domain. Though our usage of a deep research LLM minimizes over-reliance on a single domain expert, multiple experts would have been better to create the expert criteria. We were constrained by domain experts eager and available to collaborate with us.

Our experiments focus on a set of four models and four prompting strategies, and including additional models and strategies could provide a more comprehensive set of baseline results. Though many other high-performing LLMs and prompting techniques exist as of May 2025, we are conscious of budget and the environmental impact of running multiple experiments using T-FIX.

G Ethical Considerations

Using LLMs in the domains we describe in T-FIX, especially those relating to medicine, poses a unique set of risks and challenges. We do not advocate that LLMs should replace domain experts in these tasks; rather, T-FIX should serve as a step towards experts being able to use LLMs in a reliable and trustworthy way.

Additionally, LLMs are constantly changing, especially those that are company-owned and not open-source. This poses potential issues relating to the reproducibility of our baseline results as time progresses and advances are made.

Lastly, nearly all LLMs contain biases – some harmful – that may propagate up in a system built off of these models. All users of T-FIX must be conscious of this risk.

Domain	Claim	Score (Category)	Reasoning
Cosmology			
Mass Maps	**[Good]** The prominence of red and yellow suggests a universe with significant matter fluctuations.	0.9 (*Density Contrast Extremes*)	Aligns well with the Density Contrast Extremes category, describing pronounced contrasts between dense and void regions, signaling high sigma_8.
	[Bad] The mix of colors, with significant gray areas but noticeable reds and yellows, suggests a moderate Omega_m.	0.3 (*Connectivity of the Cosmic Web*)	Discusses both underdense and overdense regions, but doesn't specifically discuss connectivity or the degree of fragmentation or interconnection of the network.
Supernova	**[Good]** A prominent peak followed by a gradual decline in flux is characteristic of a type Ia supernova light curve.	1.0 (*Rise–decline rates*)	Describes a classic feature of type Ia supernovae, perfectly aligning with expert criteria on rise-and-decline rates.
	[Bad] The variability does not display a clear periodicity.	0.1 (*Periodic light curves*)	Contradicts key characteristics of periodic light curves; highlights absence of periodic behavior.
Psychology			
Politeness	**[Good]** The use of the phrase "seems defective" introduces uncertainty and avoids definitiveness.	0.9 (*hedging & tentative language*)	The phrase utilizes tentative language and is a clear example of hedging to reduce the assertive strength of a statement.
	[Bad] The utterance is a straightforward description of information from a biology textbook.	0.2 (*First-Person Subjectivity Markers*)	Weakly aligns as it describes objective reporting without the personal tone central to first-person subjectivity.
Emotion	**[Good]** This choice of description is likely intended to evoke a reaction of fear or caution.	0.9 (*Threat/Worry Language*)	The claim centers around evoking fear or caution, which directly maps to this category.
	[Bad] The text conveys an objective statement.	0.0 (*Valence*)	The claim highlights an absence of emotional content, which does not align with the Valence category or any other expert emotion categories.
Medicine			
Cholecystectomy	**[Good]** The fat and fibrous tissue overlying Calot's triangle has been fully excised, exposing only two tubular structures.	High (*Complete Triangle Clearance*)	Precisely describes complete clearance of Calot's triangle, perfectly matching expert criteria.
	[Bad] The cystic plate is not visible due to dense adhesions, making the gallbladder-liver plane indistinct.	Low (*Cystic Plate Visibility*)	Describes failure to visualize the cystic plate, opposite of the criterion, leading to low alignment.
Cardiac	**[Good]** The irregularity in the ECG could indicate a dangerous arrhythmia, such as ventricular tachycardia or fibrillation.	0.9 (*Ventricular Tachyarrhythmias*)	Directly references hallmark arrhythmias like ventricular tachycardia/fibrillation, key indicators in the category.
	[Bad] A skin lesion of the scalp is a condition not directly related to cardiac function.	0.2 (*Critical Illness – Sepsis/Shock*)	Potential weak connection if interpreted as infection, but lacks explicit signs of sepsis/shock.
Sepsis	**[Good]** Fever and high heart rate are potential signs of sepsis.	1.0 (*SIRS Positivity*)	References two SIRS criteria; strong and direct alignment with early sepsis identification guidelines.
	[Bad] The patient's lab results show an increased platelet count.	0.2 (*SOFA Score Increase*)	SOFA score focuses on low platelet counts; increased count contradicts the criterion.

Table A5: Expert-aligned claims (good and bad) across all T-FIX domains, with corresponding alignment scores and provided reasoning.