Towards Human-AI Collaboration in Healthcare: Guided Deferral Systems with Large Language Models

Joshua Strong [1] **Qianhui Men** [1] **J. Alison Noble** [1]

Abstract

Large language models (LLMs) present a valuable technology for various applications in healthcare, but their tendency to hallucinate introduces unacceptable uncertainty in critical decision-making situations. Human-AI collaboration (HAIC) can mitigate this uncertainty by combining human and AI strengths for better outcomes. This paper presents a novel *guided deferral* system that provides intelligent guidance when AI defers cases to human decision-makers. We leverage LLMs' verbalisation capabilities and internal states to create this system, demonstrating that fine-tuning small-scale LLMs with data from large-scale LLMs greatly enhances performance while maintaining computational efficiency and data privacy. A pilot study showcases the effectiveness of our proposed deferral system.

1. Introduction

Implementing artificial intelligence (AI) in decision-sensitive fields such as healthcare involves balancing the benefits of autonomy with the risks and costs of errors. Human-AI Collaboration (HAIC) aims to find this balance, by combining human and AI efforts. One approach to HAIC is using deferral systems which allow AI to handle straightforward cases while deferring complex ones to humans.

Clinicians often make decisions using their expertise in addition to the *intelligent guidance* of colleagues, which we define as task-based recommendations and informed reasoning rooted in logic. Current deferral systems lack this guidance, isolating human decision-makers. We propose that effective deferral systems should additionally simulate providing this intelligent guidance to decision-makers. This paper explores using LLMs to achieve this.

[1]Department of Engineering Science, University of Oxford, Oxford, United Kingdom. Correspondence to: Joshua Strong <joshua.strong@kellogg.ox.ac.uk>.

Workshop on Large Language Models and Cognition at International Conference on Machine Learning (ICML), Vienna, Austria.

The main challenge in building such systems is the computational expense of LLMs. Proprietary LLMs offer advantages such as state-of-the-art performance and easy implementation without needing high-performance hardware. However, they are impractical for data-sensitive applications due to lack of internal state access and privacy concerns. Open-source LLMs can perform well with relatively large amounts of parameters, but are slow and require high-performance hardware. Smaller LLMs are faster but less effective. These issues affect all real-world LLM applications, not just deferral systems. In this paper, we propose methodology for developing efficient and accurate LLMs of which are capable of guided deferral and suitable for healthcare applications.

In summary, the contributions of this paper are:

- We propose a novel deferral system, *guided deferral*, for large language models (LLMs) in computer-aided clinical diagnosis. This system not only defers cases to human decision-makers, but also provides *intelligent guidance*. We detail its practical application in healthcare and evaluate its efficacy through a pilot study.
- We evaluate the classification and deferral performance of two distinct sources of predictions; the *verbalised* and *hidden-state* predictions. Additionally, we demonstrate how a combination of these sources leads to a significant prediction in terms of classification and deferral performance.
- We demonstrate that instruction-tuning an open-source, efficient, and small-scale LLM on the *guard-railed* generation of a large-scale version of the same LLM leads to significantly improved classification and deferral performance in this task, surpassing even that of the larger model through the use of *guardrails*. This methodology ensures data privacy, which is critical for high-stakes environments such as healthcare.

2. Related Work

Human-AI Collaboration with Large Language Models. Few studies explore the use of LLMs in HAIC. Wiegreffe et al. (2022) first examined LLMs for explaining classification decisions using a human-in-the-loop approach to train a filter assessing explanation quality. Rastogi et al. (2023)

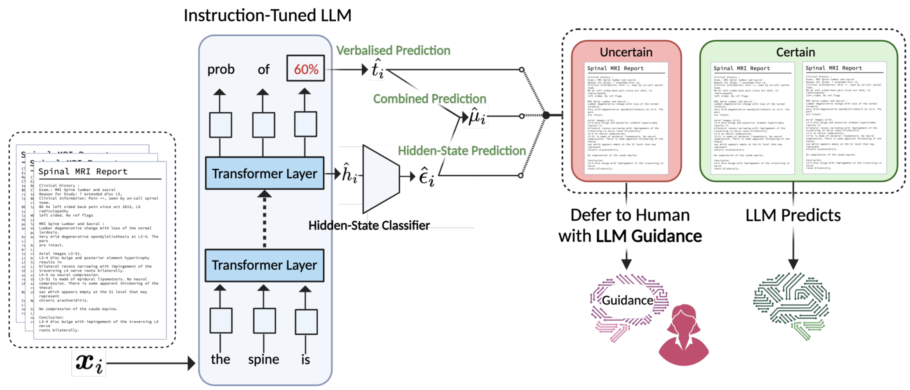

Figure 1. Our *guided deferral* system. Reports are parsed by an instruction-tuned LLM for clinical disorders. From the text output, we extract a *verbalised prediction* $\hat{t}$. We calculate a *hidden-state* $\hat{\epsilon}$ prediction from the final hidden-layer of the LLM, and its combination with $\hat{t}$ through their mean $\hat{\mu}$. Uncertain predictions, determined by either $\hat{t}$, $\hat{\epsilon}$, or $\hat{\mu}$, are deferred to humans with guidance. Certain predictions are autonomously handled by the LLM.

used HAIC to audit error-prone LLMs with other LLMs. Dvijotham et al. (2023) proposed CoDoC, a deferral system with a black-box classifier and learned deferral AI for healthcare. Our system differs by using LLMs for guided deferrals. Banerjee et al. (2023) used LLMs to provide textual guidance for clinicians in decision-making on clinical imaging tasks and argued that deferral systems are sub-optimal due to anchoring bias and lack of decision-maker support. They proposed learning-to-guide, which trains LLMs to provide decision-making guidance without deferring cases. We contend that this approach still burdens decision-makers with the time and fatigue issues deferral systems address. Our work combines LLMs in deferral systems with valuable guidance for decision-makers on deferred cases, laying the foundation for advanced deferral algorithms like Learning-to-Defer (Madras et al., 2018) or Learning with Rejection (Cortes et al., 2016). Few studies (Mozannar & Sontag, 2021; Liu et al., 2021) have evaluated the efficacy of HAIC on text classification tasks but are not LLM focused.

Selective Prediction of Large Language Models. Existing research on LLMs in selective prediction includes methods to measure uncertainty in model responses after generation (Varshney & Baral, 2023). Chen et al. (2023) proposed improving selective prediction by incorporating self-assessment. Ren et al. (2023) explored detecting out-of-distribution instances in summarisation and translation tasks. Our paper applies selective prediction to deferral systems in clinical parsing. We show that combining the model's internal state with its generated prediction enhances selective prediction without post-generation methods. We provide a comprehensive evaluation of selective prediction performance in clinical classification using real-world data, applied in deferral systems with in-distribution data.

Instruction-Tuning of Large Language Models. Finally, whilst there exists literature demonstrating improved zero-shot performance of instruction-tuned (IT) LLMs on unseen tasks (Wei et al., 2022), our work studies this improvement specifically on an in-domain task through the use of *guardrails*. Additionally, there exists works on the use of IT in the medical domain (Liu et al., 2023), but these focus on report summarisation. No works have researched the applications of IT LLMs for the use of deferral systems. Zhang et al. (2023) provide a recent survey for additional insight in this area.

3. Methods

3.1. Sources of Predictions

Utilising LLMs in clinical parsing for disorder classification presents a unique challenge in determining the classification approach. We focus our study on the top-performing methods of two distinct sources of classifications; one from the internal-states of the LLM and another from the generated textual output. Additionally, we experiment with a third through combining these sources. Specifically, the *verbalised*, *hidden-state* and *combined* sources. Next, we formally define these sources.

Verbalised Prediction Source. The *verbalised probability* is the probability of the positive class extracted from the generated text of the LLM. We denote these probabilities for input $\boldsymbol{x}_i$ as $\hat{t}_i$.

Hidden-State Prediction Source. The second prediction probability is defined based on the hidden representations of LLM to implicitly detect classifications. This is inspired by Ren et al. (2023), which utilises the final-layer hidden embedding of LLMs for out-of-distribution detection. The output embedding $\hat{h}_i \in \mathcal{H}$ for input $\boldsymbol{x}_i$ is computed as the average of the decoder's final-layer hidden-state vectors $\boldsymbol{g}_{ik} \in \mathbb{R}^d$ over all K output tokens with a hidden dimension of $d = 5120$ for the small-scale LLM:

$$\hat{h}_i := \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{g}_{ik}^{n}.$$

Then, an MLP $g : \mathcal{H} \mapsto \mathbb{R}$ is trained as a hidden-state classifier to learn the probability of the disorder from the LLM hidden representation, the *hidden-state prediction probabilities*, denoted $\hat{\epsilon}_i$, i.e. $\hat{\epsilon}_i = g(\hat{h}_i)$. We experimented with different models for the hidden-state classifier in ambitions of fully utilising the information embedded in $\hat{h}$ (see Appendix D.1 for details).

Combined Prediction. Additionally, we combine the verbalised prediction and the hidden state prediction through their mean to form the *combined prediction*, of which we denote $\hat{\mu}_i$.

3.2. Instruction-Tuning Methodology

In generating well-calibrated verbalised probabilities, we use the "*verb. 1S top-k*" prompting strategy (Tian et al., 2023), prompting the LLM to provide the top k guesses and their probabilities in a single response. Adapting this to $k = 1$ for a binary setting, we prompt the LLM to return the probability of the positive class. Additionally, we prompt for the top reason the disorder *might* and *might not* be present, using *dialectic reasoning* (Hegel, 2018) to provide intelligent guidance for decision-makers. This technique has proven effective in decision support (Jarupathirun & Zahedi, 2007). An example output is in Figure 2.

Example Generated Guidance

TOP REASON FOR: The MRI report indicates that there is narrowing of the right exit foramen at the L5-S1 level, which is causing compression of the exiting right L5 nerve root. This suggests that there is foraminal stenosis (FS) at this level.

TOP REASON AGAINST: There is no mention of foraminal stenosis specifically at the L5-S1 level.

CONCLUSION: Based on the information provided, there is a possibility of foraminal stenosis at the L5-S1 level due to the narrowing of the right exit foramen and compression of the exiting right L5 nerve root. However, the report does not explicitly mention foraminal stenosis at this level.

PROBABILITY OF FS PRESENT AT L5-S1: 60%

Figure 2. Example guidance based on a spinal MRI radiological report. The instruction-tuned LLM is able to intelligently infer a diagnosis with sound logic without explicit textual diagnosis.

3.3. Deferral Mechanism

Our deferral strategy is based on the confidence of predictions that is determined by its distance to the chosen decision boundary of 0.5. Formally, predictions $\hat{p} \in \{\hat{t}, \hat{\epsilon}, \hat{\mu}\}$ are transformed to sorted relative confidence probabilities $\tilde{p} = \text{sorted}\,(2\,|\hat{p} - 0.5|)$ for equal comparison between the positive and negative classes. The resulting $\tilde{p}$ determines the hierarchy of cases to be deferred.

The deferral performance is measured through recursively iterating through all elements $\tilde{p}$, deferring this prediction (without replacement) and measuring the classification performance of the LLM on the remaining cases. This procedure describes the AUARC metric (Nadeem et al., 2009) when measuring accuracy. Intuitively, good deferral behaviour should demonstrate a monotonically increasing accuracy with increasing deferral rate.

3.4. Pilot Study in Investigating the Effectiveness of Generated Guidance

We conduct a pilot study with 20 participants to evaluate the effectiveness of our guidance in the scenario of deferring 30 ($\approx$5%) of the most uncertain test predictions. Participants received background information, including clinical details, examples of prediction outcomes with associated MRI reports and the LLM's performance on a validation set to help participants develop a sufficient mental model of the LLM, of which has been shown to be important in HAIC systems (Kulesza et al., 2012; Bansal et al., 2019). The ordering of the 30 questions were randomised for each participant to reduce order bias.

Participants moved to the next question if their prediction matched the LLM's. If it differed, they received guidance and could either change their prediction or keep it based on the guidance and their understanding of the LLM. This process allowed us to assess human performance with and without guidance. Effective guidance should help participants recognise both the accuracy and inaccuracy of their judgements. Full details are in Appendix C.

4. Experiments

Data. We use the OSCLMRIC (Oxford Secondary Care Lumbar MRI Cohort) dataset, containing professionally annotated lumbar MRIs and radiological reports for various types of stenosis at different spinal levels, of which serve as our ground-truth labels. An example report is shown in Listing 2. The dataset is highly imbalanced, with $\approx$95% of labels negative. Each report is parsed to detect the binary presence of three types of spinal stenosis (foraminal stenosis [FS], spinal canal stenosis [SCS], and lateral recess stenosis [LRS]) at six lumbar spine levels, resulting in 1,800 examples. The data is randomly split into 30% for generating an instruction-tuning dataset, 20% for training the hidden-state classifier, 20% for validation, and 30% for testing.

SOTA Baseline. We use Tulu V2 70B as the SOTA baseline model for our experiments, of which is the highest-performing open-source LLM against several benchmarks at the time these experiments were conducted (Beeching et al., 2023). Proprietary LLMs are omitted as baselines due to the aforementioned data privacy issues.

Table 1. Classification, calibration, deferral performance and LLM efficiency against test split (N=540) of 9 setups: 13B base and instruction-tuned models on their verbalised ($\hat{t}_{\text{BASE-13B}}$ and $\hat{t}_{\text{INSTRUCT-13B}}$ respectively), hidden-state ($\hat{\epsilon}_{\text{BASE-13B}}$ and $\hat{\epsilon}_{\text{INSTRUCT-13B}}$ respectively) and combined probability predictions ($\hat{\mu}_{\text{BASE-13B}}$ and $\hat{\mu}_{\text{INSTRUCT-13B}}$ respectively). We include Tulu-70B experiments as the SOTA baselines highlighted in grey. Statistically significant best results in **bold**.

SETUP		CLASSIFICATION PERF.			CALIBRATION PERF.		DEFERRAL PERF.	LLM EFFICIENCY		
		RECALL↑	PRECISION↑	F1-SCORE↑	ECE↓	ACE↓	AUARC↑	REL. S./GEN.↓	MEM.↓	E.R.↓
(1)	$\hat{t}_{\text{BASE-13B}}$	0.85±0.02	0.61±0.01	0.61±0.01	0.26±0.01	0.32±0.01	0.897460±0.0063	|	|	|
(2)	$\hat{\epsilon}_{\text{BASE-13B}}$	0.73±0.02	0.94±0.03	0.80±0.02	0.03±0.01	**0.04±0.01**	0.994715±0.0006	**0.08±0.02**	**6.02**	0.29
(3)	$\hat{\mu}_{\text{BASE-13B}}$	0.78±0.03	0.87±0.04	0.82±0.03	0.13±0.01	0.18±0.01	0.993318±0.0010	|	|	|
(4)	$\hat{t}_{\text{BASE-70B}}$	0.90±0.02	0.78±0.01	0.83±0.01	0.27±0.01	0.33±0.01	0.973529±0.0052	|	|	|
(5)	$\hat{\epsilon}_{\text{BASE-70B}}$	0.87±0.01	0.97±0.01	0.91±0.01	0.02±0.01	0.05±0.01	0.997191±0.0002	1.00	32.08	0.04
(6)	$\hat{\mu}_{\text{BASE-70B}}$	0.90±0.01	0.95±0.02	0.92±0.01	0.14±0.01	0.19±0.01	0.997260±0.0003	|	|	|
(7)	$\hat{t}_{\text{INSTRUCT-13B}}$	0.85±0.02	0.92±0.02	0.88±0.02	0.27±0.01	0.32±0.01	0.987501±0.0024	|	|	|
(8)	$\hat{\epsilon}_{\text{INSTRUCT-13B}}$	0.89±0.01	**0.98±0.01**	0.93±0.01	**0.01±0.01**	0.05±0.01	**0.997409±0.0002**	**0.08±0.02**	**6.02**	**0.0002**
(9)	$\hat{\mu}_{\text{INSTRUCT-13B}}$	**0.91±0.01**	**0.98±0.01**	**0.94±0.01**	0.14±0.01	0.19±0.01	0.996978±0.0013	|	|	|

Abbreviations. We abbreviate IT and base (non-IT) models as "INSTRUCT" and "BASE", respectively. Tests for statistical significance (SS) are conducted using a one-sided non-parametric Mann-Whitney U test (Mann & Whitney, 1947), unless stated otherwise.

Evaluation Metrics. The Area Under Accuracy-Rejection Curve (AUARC) and F1-score are used to measure the deferral and classification performance, respectively. We further evaluate the calibration of our setups with ECE (Guo et al., 2017) and ACE (Nixon et al., 2020). LLM efficiency is evaluated based on *Rel. s/Gen.* (seconds/generation relative to the baseline), *Mem.* (GPU VRAM required in GB), and *E.R.* the error rate (the number of unsuccessful generations for each successful generation).

4.1. Computational Results

Instruction-Tuned Models Are More Accurate, Equally Calibrated, Better Deferral Systems, and More Efficient. We find a SS improvement in F1-Score of verbalised predictions of $\hat{t}_{\text{INSTRUCT-13B}}$ in comparison to the baseline $\hat{t}_{\text{BASE-70B}}$, demonstrating instruction-tuning methodology to be successful. The greatest deferral performance results from the hidden-state predictions of the 13B setup, surpassing that of the 70B baseline. The IT model is seen to significantly improve in all LLM efficiency metrics.

Access to Internal States of LLMs Improves Deferral Performance. Utilising open-source LLMs allows for hidden-state predictions which are shown to be the best calibrated and are utilised in the greatest deferral performance (Table 1, Row 8). This finding suggests deferral systems built utilising proprietary LLMs without access to internal states and relying solely on verbalised predictions are inadequate.

The Combined Prediction Leads to Greatest Classification Performance. $\hat{\mu}_{\text{INSTRUCT-13B}}$ is the greatest classification source (Table 1, Row 9), surpassing that of $\hat{\epsilon}_{\text{INSTRUCT-13B}}$ and $\hat{t}_{\text{INSTRUCT-13B}}$ ($p<0.01$). The improvement in $\hat{\mu}$ over $\hat{t}$ and $\hat{\epsilon}$ is seen in all setups. This implies an interesting finding; in that the verbalised and hidden-state sources contain valuable and distinct pieces of information contributing to classification. There exists a positive correlation of $\rho = 0.53 \pm 0.07$ between these predictions. When focusing our analysis when $\hat{\mu}$ is correct and the predictions of $\hat{t}$ and $\hat{\epsilon}$ differ, a correct $\hat{\epsilon}$ prediction is combining with an incorrect $\hat{t}$ prediction in the majority of cases, at an average of (75 ± 0.14)% of these cases. When $\hat{\mu}$ is incorrect, a correct $\hat{t}$ prediction is combining with an incorrect $\hat{\epsilon}$ prediction in the majority of cases, at an average of (75 ± 0.25)% of these cases. See Figure 3 for confusion matrices.

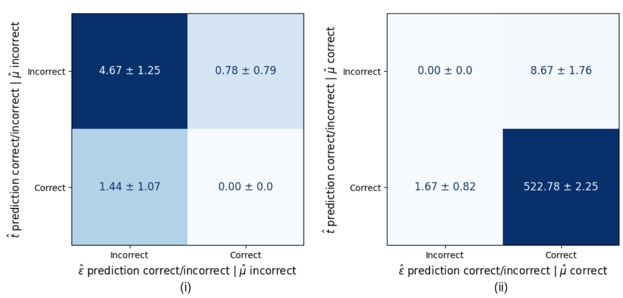

Figure 3. Confusion matrix illustrating the number of test cases for $\hat{t}$ and $\hat{\epsilon}$ precision, given (i) the incorrect prediction of their combination $\hat{\mu}$, and (ii) correct prediction.

The Combined Prediction Has The Greatest Deferral Performance in Mostly-Autonomous Systems. We find a SS ($p<0.1$) improvement in the deferral strategy of deferring on $\hat{\epsilon}_{\text{INSTRUCT-13B}}$ and classifying on $\hat{\mu}_{\text{INSTRUCT-13B}}$, in comparison to the next-best strategy of deferring and classifying solely on $\hat{\epsilon}_{\text{INSTRUCT-13B}}$, in situations where $<5\%$ of cases are deferred. In practice, this would be cases where users would require the system to be mostly autonomous. This implies that the combined prediction is a greater predictor of disease in high-uncertainty situations. In deferral rates $\geq 5\%$, there exists no SS difference between the two strategies, but a SS over all other setups.

4.2. Pilot Study Results

Following the computational deferral results in high uncertainty, we defer 5% of the test cases based on the hidden-state prediction and provide the classification prediction of the combined prediction.

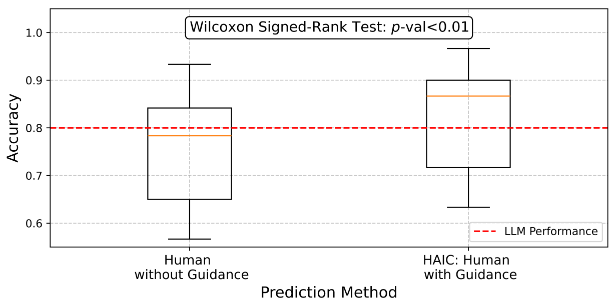

Figure 4. **Human accuracy without guidance is lower and more variable than with guidance (HAIC).** Accuracy of prediction methods from pilot study. On average, guided humans outperform both unguided humans and the LLM.

Guidance is Effective in Aiding Human Decision-Making. Figure 4 displays the accuracy of participant decision-making with and without guidance, compared to LLM performance. All participants' accuracy improved with guidance. We rejected the null hypothesis that performance without guidance equals performance with guidance (paired one-sided Wilcoxon signed-rank test, $p<0.01$). When confronted with AI disagreement, the provided guidance proved effective in assisting participants to arrive at the correct final decision. This was true not only when the participant was incorrect, but importantly also when the LLM was incorrect. We rejected the null hypothesis that the proportion of humans changing their prediction is independent of the AI's prediction correctness (χ^2-test, $p<0.01$).

5. Conclusion

We develop a guided deferral framework using LLMs for clinical decisions, with HAIC as a key focus. Our study shows that IT small-scale LLMs can achieve significant deferral and classification performance while maintaining computational efficiency through instruction-tuning on supervised data from a large-scale LLM. We utilise two prediction sources—verbalised and hidden-state, yielding valuable insights for classification and deferral. The combination of these sources results in a significant classification and high-uncertainty deferral prediction. Finally, we validate the efficacy of our proposed guided deferral system through a pilot study, with the results showing a significant improvement in human decision-making performance under the LLM guidance.

Limitations. In this paper, we prioritise healthcare in our analysis due to its greater ethical significance in research. However, there exists a severe lack of high-quality report data with labels in medical domain, making it difficult for LLM evaluation. When implementing our system, extra efforts are required to educate clinicians about its capabilities to prevent unexpected errors. This involves addressing human biases such as anchoring and confirmation bias, as well as briefing them on the system's training data distribution, including concepts such as data drift and out-of-distribution instances. Such understanding is vital, particularly in scenarios where variations in report style or format might affect the system's performance.

Acknowledgements. JS is supported by the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1). JS gratefully acknowledges the Kellogg College Research Support Grant of which supported this research. The authors acknowledge UKRI grant reference EP/X040186/1 (Turing AI Fellowship: Ultra Sound Multi-Modal Video-based Human-Machine Collaboration). The OSCLMRIC dataset was collected by Professor Jeremy Fairbank with HRA approval (IRAS Project ID: 207858, Protocol Number:12139). The authors extend their gratitude to Professor. Andrew Zisserman and Dr. Amir Jamaludin from the Visual Geometry Group, University of Oxford, for the provision of this data. A special thank-you to all participants of the pilot study.

References

Banerjee, D., Teso, S., and Passerini, A. Learning to guide human experts via personalized large language models, 2023.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pp. 2–11, 2019.

Beeching, E., Fourrier, C., Habib, N., Han, S., Lambert, N., Rajani, N., Sanseviero, O., Tunstall, L., and Wolf, T. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Chen, J., Yoon, J., Ebrahimi, S., Arik, S. O., Pfister, T., and Jha, S. Adaptation with self-evaluation to improve selective prediction in llms, 2023.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. 2016. URL https://cs.nyu.edu/~mohri/pub/rej.pdf.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms, 2023.

Dvijotham, K. D., Winkens, J., Barsbey, M., Ghaisas, S., Stanforth, R., Pawlowski, N., Strachan, P., Ahmed, Z., Azizi, S., Bachrach, Y., Culp, L., Daswani, M., Freyberg, J., Kelly, C., Kiraly, A., Kohlberger, T., McKinney, S., Mustafa, B., Natarajan, V., Geras, K., Witowski, J., Qin, Z. Z., Creswell, J., Shetty, S., Sieniek, M., Spitz, T., Corrado, G., Kohli, P., Cemgil, T., and Karthikesalingam, A. Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians. *Nature Medicine*, 29(7):1814–1820, Jul 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02437-x. URL https://doi.org/10.1038/s41591-023-02437-x.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks, 2017.

Hegel, G. W. F. *Georg Wilhelm Friedrich Hegel: The Phenomenology of Spirit*. Cambridge Hegel Translations. Cambridge University Press, 2018.

Ivison, H., Wang, Y., Pyatkin, V., Lambert, N., Peters, M., Dasigi, P., Jang, J., Wadden, D., Smith, N. A., Beltagy, I., and Hajishirzi, H. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023.

Jarupathirun, S. and Zahedi, F. M. Dialectic decision support systems: System design and empirical evaluation. *Decision Support Systems*, 43(4):1553–1570, August 2007. ISSN 0167-9236. doi: 10.1016/j.dss.2006.03.002. URL http://dx.doi.org/10.1016/j.dss.2006.03.002.

Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*, pp. 1–10, 2012.

Liu, J., Gallego, B., and Barbieri, S. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis, 2021.

Liu, Z., Li, Y., Shu, P., Zhong, A., Yang, L., Ju, C., Wu, Z., Ma, C., Luo, J., Chen, C., et al. Radiology-llama2: Best-in-class large language model for radiology. *arXiv preprint arXiv:2309.06419*, 2023.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: Improving fairness and accuracy by learning to defer, 2018.

Mann, H. B. and Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947. doi: 10.1214/aoms/1177730491. URL https://doi.org/10.1214/aoms/1177730491.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert, 2021.

Nadeem, M. S. A., Zucker, J.-D., and Hanczar, B. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In Džeroski, S., Guerts, P., and Rousu, J. (eds.), *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pp. 65–81, Ljubljana, Slovenia, 05–06 Sep 2009. PMLR. URL https://proceedings.mlr.press/v8/nadeem10a.html.

Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., and Tran, D. Measuring calibration in deep learning, 2020.

Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H., and Amershi, S. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23. ACM, August 2023. doi: 10.1145/3600211.3604712. URL http://dx.doi.org/10.1145/3600211.3604712.

Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., and Liu, P. J. Out-of-distribution detection and selective generation for conditional language models, 2023.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023.

Varshney, N. and Baral, C. Post-abstention: Towards reliably re-attempting the abstained instances in qa, 2023.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. Reframing human-ai collaboration for generating free-text explanations, 2022.

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

A. Additional Listings

```
### QUESTION: Is there <disorder> at the <ivd> level?

### INSTRUCTIONS: You are an expert spinal surgeon, giving guidance to novices in diagnosing <disorder> at the <ivd> level based on the
     contents of a report. Firstly, start by giving the single top reason both for and against, considering information ONLY
    regarding the <ivd> level. Finally, at the end of your response, give a final detailed conclusion including a definitive answer
    to the question, based solely on your top reasons, and importantly including a probability that <disorder short> is present at
    the <ivd> level in this format: 'CONCLUSION: <your conclusion>. PROBABILITY OF <disorder short> PRESENT AT <ivd>: <your
    probability>% </s>'. You must end your response after this.

### RULES:
1. Your probability must not be 50%.
2. If you think it is present, your probability must be greater than 50%.
3. If you think it is not present, your probability must be less than 50%.
4. Even if the question is not answerable, you must still give a definitive answer and probability.
5. Give your answer only in English.

### ASSUMPTIONS:
1. Assume that any severity or of <disorder> indicates that it is present and your probability should be greater than 50%.
2. The absence of information about <disorder> at the <ivd> level indicates that it is not present and your probability should be less
    than 50%.
3. Assume that you can only use information at the <ivd> level to diagnose <disorder>.
4. Foraminal Stenosis is the narrowing of the foramen, Lateral Recess Stenosis is the narrowing of the lateral recess, and Spinal Canal
     Stenosis is the narrowing of the central spinal canal.
5. The presence of <disorder1> or <disorder2> does not imply the presence of <disorder>.

### REPORT: <report>
```

Listing 1. Base-prompt used in generating data to instruction-tune and in parsing of reports. The relevant level, disorder, question, and report were inserted to make the final prompt.

```
Clinical History :
Exam.: MRI Spine lumbar and sacral
Reason for Study: ? Right Si Nerve Root Pain suitable for
injection/surgery
Clinical Information: 7/12 severe and worsening right leg
pain S1. Normal neurology. High disability, 2
crutches/wheelchir outdoors. IIieitis/crohns.

MRI Spine Lumbar and Sacral :
L5 vertebra plana in addition to diffuse infiltration of L4,
with minor collapse. There is also patchy involvement of the
S1 vertebral body, and of the right ilium and to a lesser
extent the left ilium. Small lesion in L1. Satisfactory
vertebral alignment.

Axial images L3-S1.
L3-4 no neural compression.
L4-5 narrowing of the right lateral recess due to tumour
with impingement of the traversing right L5 nerve root.
L5-S1 narrowing of the right exit foramen due to tumour with
compression of the exiting right L5 nerve root. Minor
displacement of the right S1 nerve root.

Conclusion:
Features of metastatic disease. Right-sided nerve root
compression as described.
```

Listing 2. An example spinal MRI radiological report.

B. Experimental Details

B.1. Inference

Inference hyperparameters are as follows:

- `max_new_tokens`: 1000
- `temperature`: 0.8
- `top_k`: 50
- `top_p`: 0.95

B.2. Training

We instruction-tune the 13 billion parameter LLM *Tulu V2 DPO* (Ivison et al., 2023) using QLoRA (Dettmers et al., 2023) on generations solely. The base model can be downloaded from `https://huggingface.co/allenai/`

tulu-2-dpo-13b. Training time on a single GeForce RTX 4090 varied by number of generated sets, but for no longer than 6 hours, with CUDA v12.4.

The hyperparameters used for training are listed below:

- **Model Preparation**:
  - `rank (r)`: 128
  - `target_modules`: {q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj}
  - `lora_alpha`: 16
  - `lora_dropout`: 0
  - `bias`: none
  - `use_gradient_checkpointing`: True
  - `random_state`: 3407
  - `max_seq_length`: 4096
- **Training Configuration**:
  - `Training Batch Size`: 10
  - `warmup_steps`: 10
  - `logging_steps`: 1
  - `learning_rate`: 2e-4
  - `Optimiser`: adamw_8bit
  - `bf16`: True
  - `weight_decay`: 0.1
  - `warmup_ratio`: 0.01
  - `lr_scheduler_type`: linear

B.3. Guard-rail Implementation

Algorithm 1 details the implementation of the guard-rail used in generating our instruction-tuning data.

Algorithm 1 Guard-rail Algorithm in Generating Instruction-Tuning Data

```
Set failure_count = 0
repeat
    Generate a response
    if response contains non-English text or does not end with an EOS token or contains multiple probability predictions
    or contains no firm answer then
        Retry generation
    else
        if prediction matches the label then
            Generation is successful
            End Algorithm
        else
            Increment failure_count by 1
        end if
    end if
until failure_count = 10
Ignore label check and save the prediction.
```

C. Pilot Study

We detail the instructions given to pilot study participants in this section.

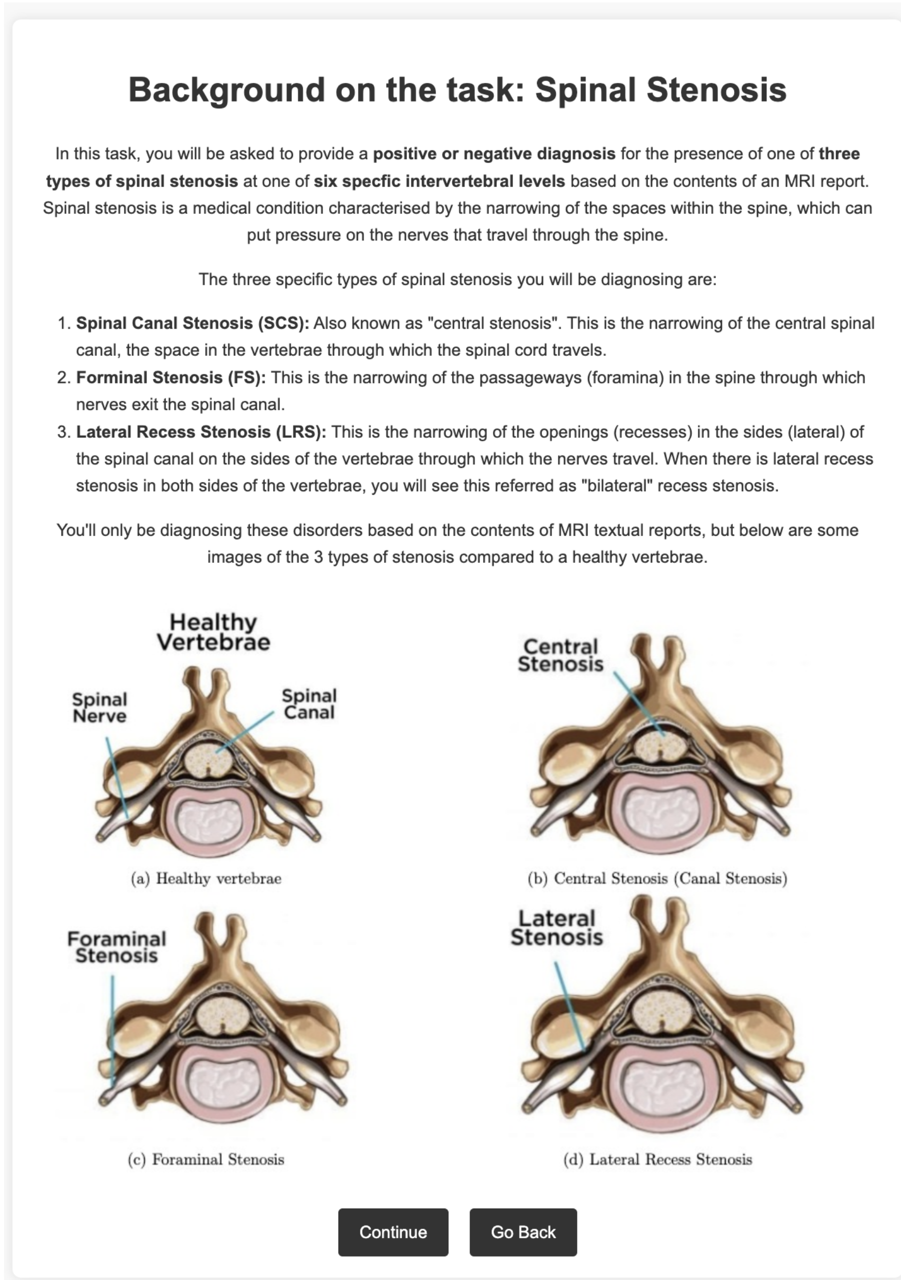

Figure 5. Page 1: Participants are first briefed on the clinical diagnosis task at hand, including the three types of spinal stenosis.

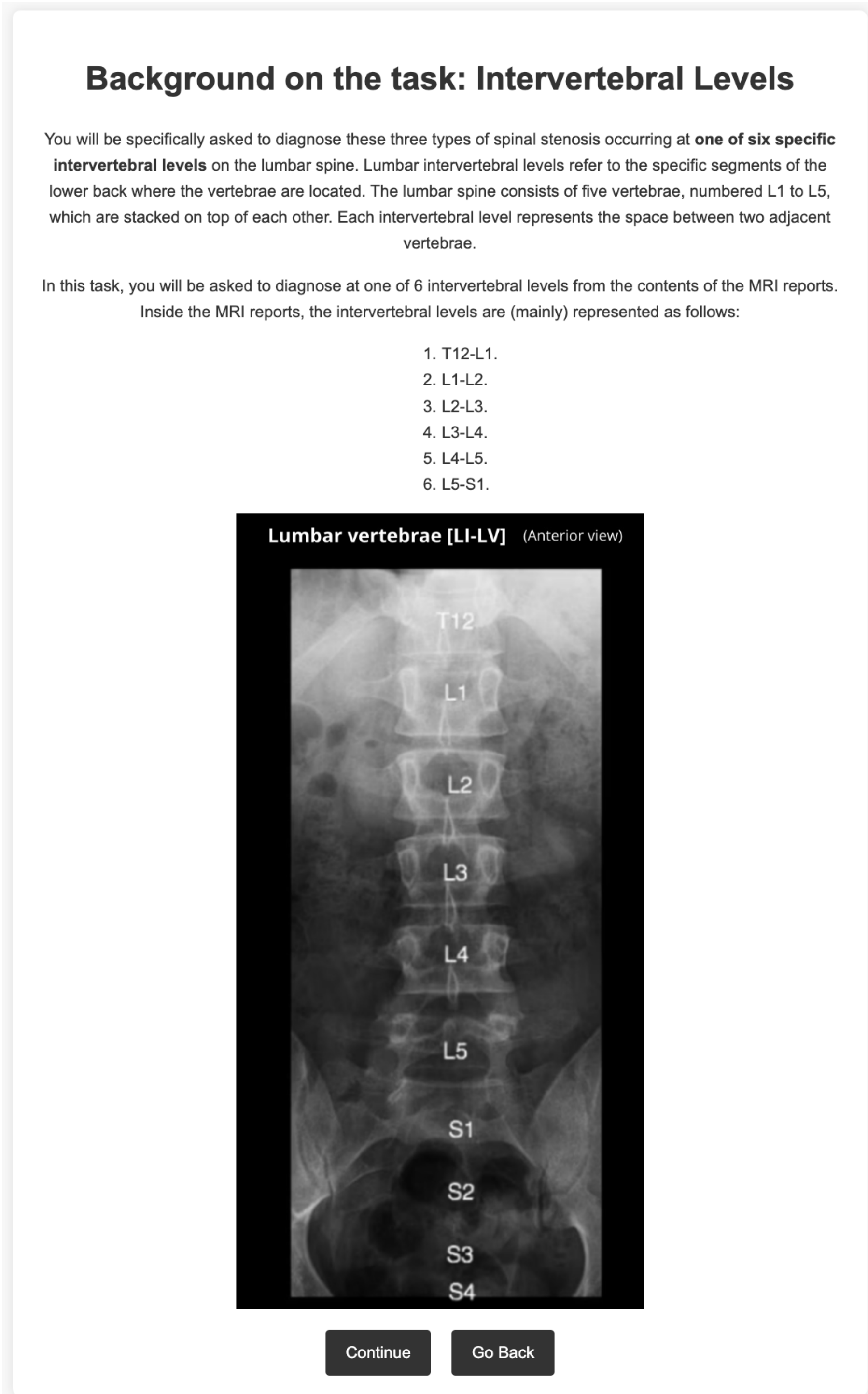

Figure 6. Page 2: Next, participants are first briefed on the notion of intervertebral levels.

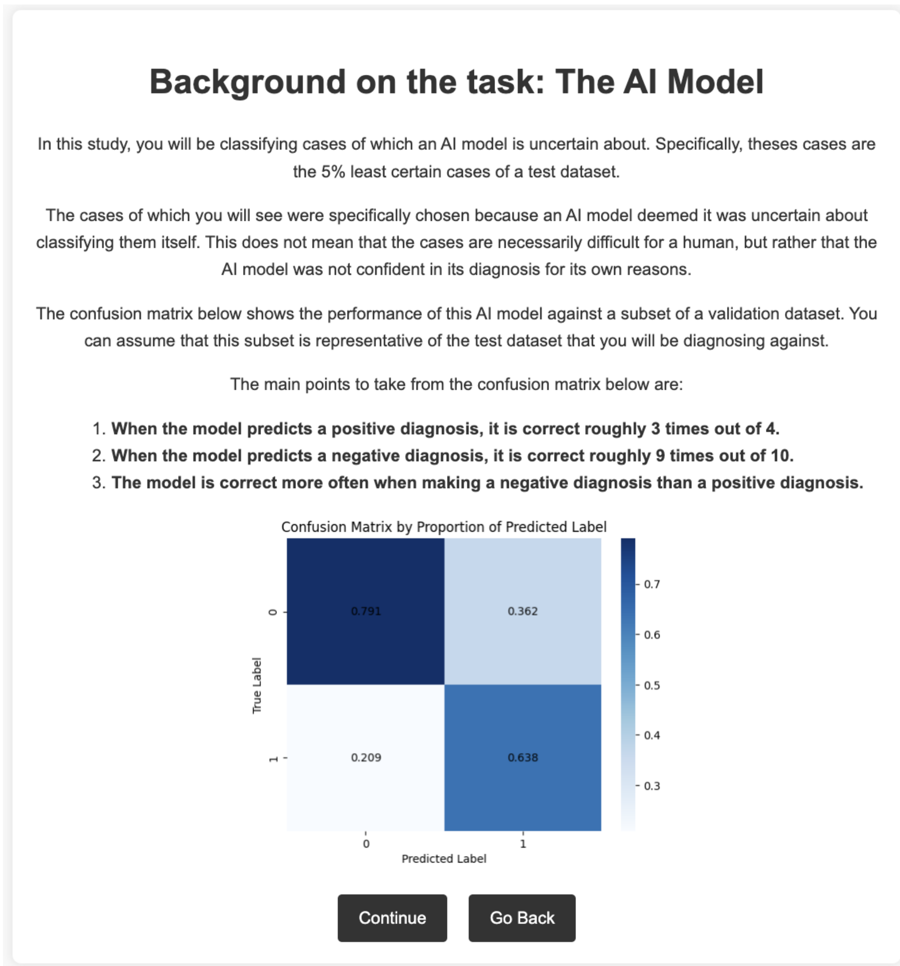

Figure 7. Page 3: We give participants a mental model of the classification performance of the LLM of which they should expect.

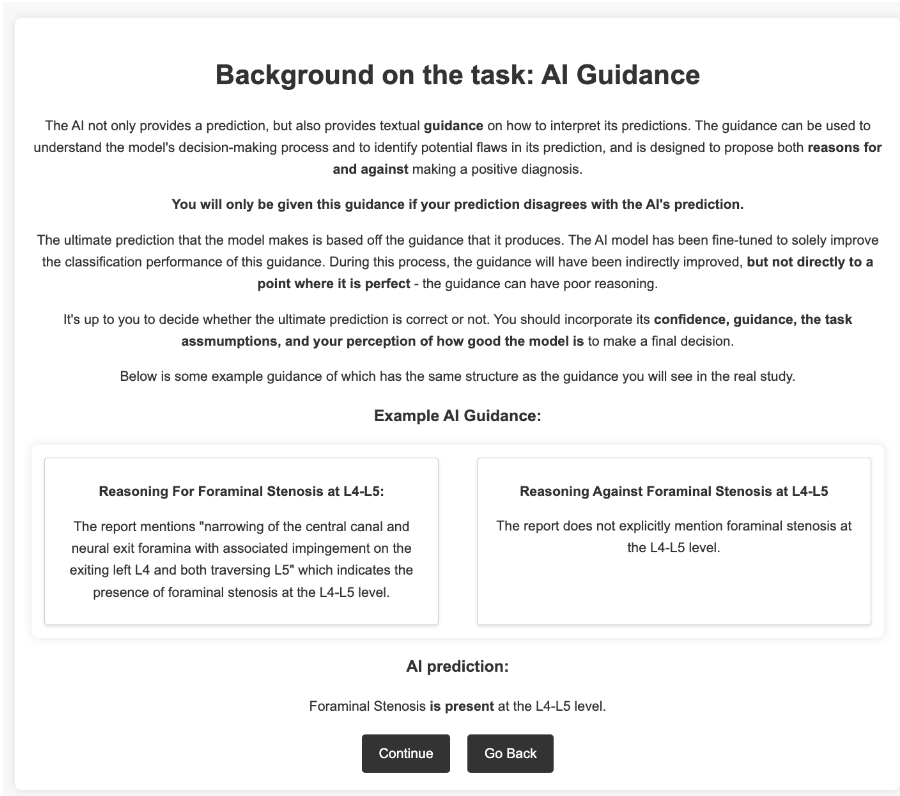

Figure 8. Page 4: We brief participants on the type of guidance they should expect during the study.

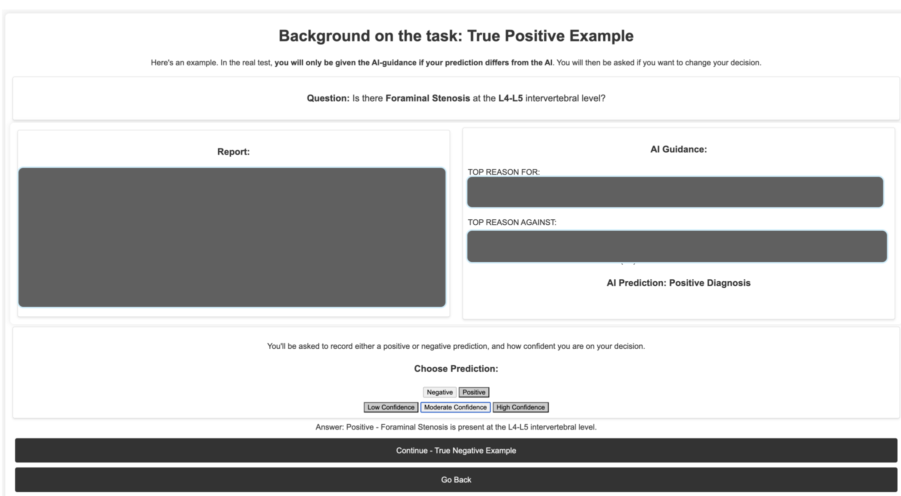

Figure 9. Pages 5-9: Participants are given 4 examples; a true a positive, a false positive, a true negative and a true positive. Sensitive data is censored behind grey boxes.

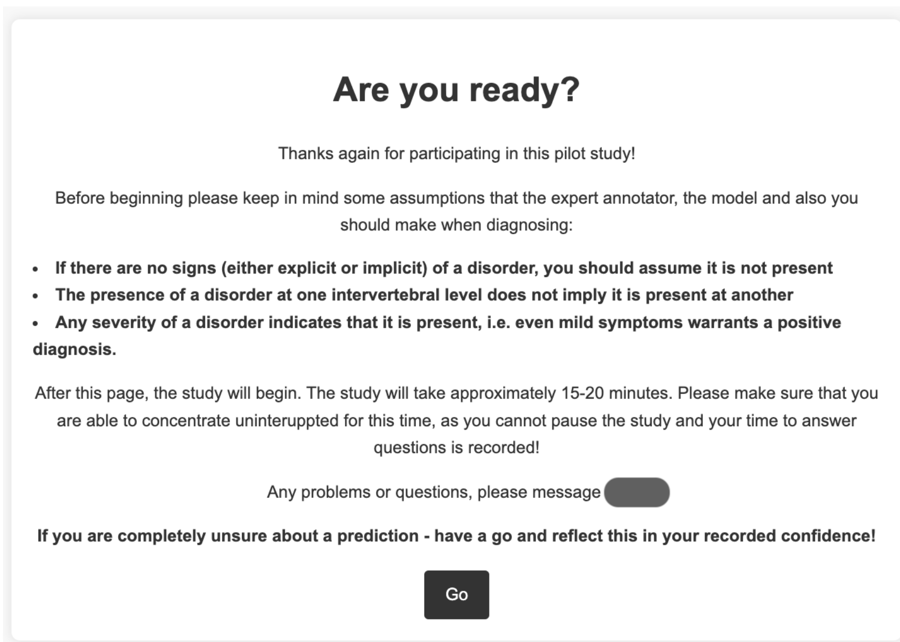

Figure 10. Page 10: Before starting the study, users are briefed on the assumptions they should make when predicting.

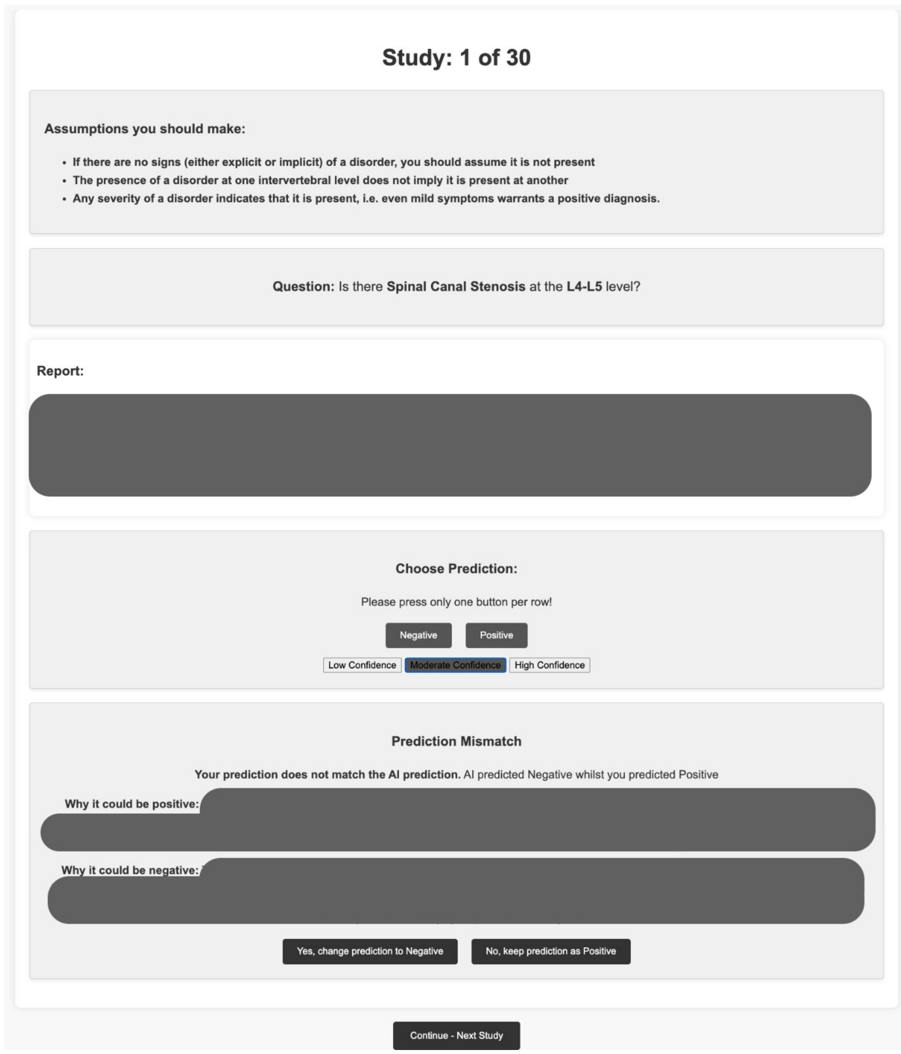

Figure 11. Page 11-41: The study consists of 30 of the most uncertain predictions from the LLM. Participants are first asked to make their own prediction. If their prediction matches with the LLM, they continue to the next question. Otherwise, they are told they are in disagreement with the AI and are provided guidance. They are then given the opportunity to change their decision. Additionally, user's time to completion and self-recorded confidence are recorded during the study. Confidential data is censored behind grey boxes.

D. Additional Experiments

D.1. Hidden-State Classifier Experiments

In order to maximise the predictive capabilities of the final-hidden layer of the LLM, we experiment with different hidden-state MLP classifier architectures. We choose the architecture of which maximises the deferral performance (AUARC) on the validation split. Table 2 details the results of these experiments. We choose a three fully-connected layer as our hidden-state classifier due to joint statistically significant deferral performance and F1-Score.

Setup	**F1-Score**	**AURAC**
One fully-connected layer	0.88 ± 0.03	0.995734 ± 0.0004
Three fully-connected layers	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.995976 \pm 0.0003}$
Five fully-connected layers	0.90 ± 0.01	$\mathbf{0.996068 \pm 0.0003}$

Table 2. Comparison of classification performance (F1-Score) and deferral performance (AURAC) of two setups: one fully-connected layer and three fully-connected layers.