Learning to Guide Human Experts via Personalized Large Language Models

Debodeep Banerjee[1,2], Stefano Teso[3,1], and Andrea Passerini[1]

[1] DISI, University of Trento, Italy
[2] DI, University of Pisa, Italy
[3] CIMeC, University of Trento, Italy

Abstract. In *learning to defer*, a predictor identifies risky decisions and defers them to a human expert. One key issue with this setup is that the expert may end up over-relying on the machine's decisions, due to anchoring bias. At the same time, whenever the machine chooses the deferral option the expert has to take decisions entirely unassisted. As a remedy, we propose *learning to guide* (LTG), an alternative framework in which – rather than suggesting ready-made decisions – the machine provides *guidance* useful to guide decision making, and the human is entirely responsible for coming up with a decision. We also introduce SLOG, an LTG implementation that leverages (a small amount of) human supervision to convert a generic large language model into a module capable of generating *textual* guidance, and present preliminary but promising results on a medical diagnosis task.

Keywords: Hybrid Decision Making · Learning to Defer · Interactive Machine Learning · Large Language Models · Medical Diagnosis.

1 Introduction

High-stakes applications in healthcare, criminal justice and policy making can substantially benefit from the introduction of AI technology. However, full automation in these scenarios is not desirable, for ethical, safety and legal concerns, if not explicitly forbidden by law [Government of Canada, 2019, European Commission, 2021]. For these reasons, human-AI or *Hybrid decision making* (HDM) is becoming increasingly popular to tackle high-stakes tasks. HDM algorithms pair a human decision maker with an AI agent – often implemented as a machine learning model – capable of providing support, with the goals of improving *decision quality* and lowering *cognitive effort*.

Most current approaches to HDM follow a principle of *separation of responsibilities*, in the sense that they work by routing novel inputs to exactly one of the two agents – *either* the human *or* the AI – who is then responsible for coming up with a decision. Specifically, in existing approaches [Madras et al., 2018, Mozannar and Sontag, 2020, Keswani et al., 2022, Verma and Nalisnick, 2022, Liu et al., 2022, Wilder et al., 2021, De et al., 2020, Raghu et al., 2019, Okati

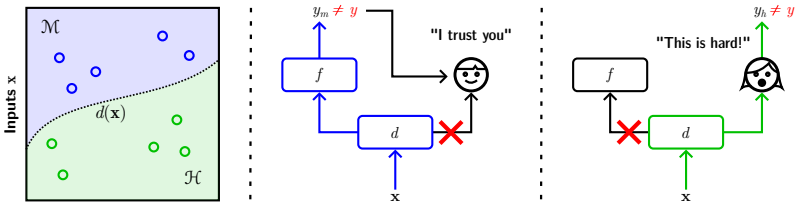

Fig. 1. **Left**: Existing HDM approaches employ a deferral function $d(\mathbf{x})$ to *partition* the input space $\mathcal{X}$ into $\mathcal{H}$ and $\mathcal{M}$. **Middle**: A predictor $f(\mathbf{x})$ handles those inputs falling in $\mathcal{M}$ (**blue** arrow). Because of *anchoring bias*, the human expert may end up blindly trusting their (possibly poor) decisions y_m. **Right**: The human, on the other hand, is left completely unassisted for those (possibly hard) decisions falling in $\mathcal{H}$, increasing the chance of mistakes (**green** arrow).

et al., 2021], the AI first assesses whether an input can be handled in autonomy – e.g., it is either low-risk or can be handled with confidence – and defers it to a human partner otherwise. These algorithms are beneficial in that they enable the human to focus on those cases that (according to the machine) require their attention. We argue that, as shown in Fig. 1, this setup is *suboptimal* and potentially *unsafe*. It is suboptimal because, whenever the machine opts for deferral, the human is left resolving hard cases completely unassisted, thus conflicting with the goals of HDM. At the same time, it is unsafe, because humans are affected by *anchoring bias* [Rastogi et al., 2022], a phenomenon whereby human decision makers tend to blindly rely on an initial impression (the anchor) and refrain from exploring alternative hypotheses. As a result, they will tend to over-trust the machine's decisions when available and ignore their own opinions, a well-studied phenomenon called *automation bias* [Cummings, 2012]. This effectively undermines the *human oversight* over algorithmic decisions that is increasingly being required by governments around the world to regulate the use of AI in high-stakes applications [Green, 2022].

As a remedy, we propose *learning to guide* (LTG), an alternative algorithmic setup that is aimed at mitigating these issues. In LTG the machine is trained to supply its human partner with interpretable *guidance* highlighting those aspects of the input that are useful for coming up with a sensible decision. For instance, in pathology prediction, the guidance could highlight those symptoms present in the input X-ray scan that are indicative of possible diagnoses. In LTG, *by construction*, all decisions are taken by the human expert – thus preventing automation bias – but facilitated by accompanying machine guidance.

We showcase LTG on a high-stakes medical decision making task, focusing on guidance formulated in *natural language*. Specifically, we propose SLOG, and algorithm for turning vision-language large language models (VLMs) [Radford et al., 2021, Yan and Pei, 2022, Sharma et al., 2021] into a high-quality guidance generator. In a nutshell, SLOG fine-tunes an VLM pre-trained for caption generation using human feedback in the form of numerical scores representing

the informativeness of the VLM generated guidance for the downstream decision making task. Since feedback is expensive to acquire and therefore available in modest amounts, SLOG uses it to train a *surrogate model* that predicts the human's judgments, and uses the latter to fine-tune the VLM in an end-to-end fashion. Our experiments on a challenging medical diagnosis task indicate that VLMs fine-tuned with SLOG output interpretable task-specific guidance that can be used to infer high-quality decisions.

2 Learning to Guide with SLOG

Learning to guide for medical diagnosis. We consider the problem of diagnosing lung pathologies y from X-ray scans $\mathbf{x}$. Rather than learning a classifier for inferring y directly, as in LTD, in LTG the goal is to learn a *guidance generator* that, given $\mathbf{x}$, extracts *textual guidance* $\mathcal{G}$ in the form of a short caption capturing salient properties of the scan that are useful for supporting human decision making. Naturally, guidance should be both *interpretable* and *informative*, so that the human decision maker can make a reasonable decision based on it.

The SLOG algorithm. To address this problem, we propose SLOG. It requires access to the following elements: (*i*) A *caption generator* $g : \mathbf{x} \mapsto (\mathcal{G}, \mathbf{z})$, implemented with an LLM, that extracts textual guidance $\mathcal{G}$ as well as the latent representation $\mathbf{z}$ of $\mathcal{G}$, and (*ii*) A *decision maker* (DM) that, given $\mathbf{x}$ and $\mathcal{G}$, comes up with a decision y, say `healthy` vs. `pneumonia`, and – whenever explicitly requested – also with a *score* $q \in \mathbb{R}$ summarizing how good the guidance $\mathcal{G}$ is for inferring a decision.

The LLM g is pre-trained and as such it does not generate captions specifically tailored for the decision making task at hand. The only party capable of determining whether the textual guidance is good enough is the human partner, so in principle, we can use their judgment to fine-tune g so as to produce more useful guidance. The limiting factor here is annotation cost: an annotator can only produce so much feedback, making it difficult to fine-tune the LLM with it.

SLOG tackles these challenges in an iterative fashion. Let $\mathcal{D} = \{\mathbf{x}_i\}$ be a data set of X-ray images. In each step t, SLOG takes an (initially pre-trained) LLM g and uses it to generate guidance $\mathcal{G}$ and corresponding embeddings $\mathbf{z}$ for a small set of images $\mathbf{x} \in \mathcal{D}$. These are shown to the decision maker, who scores all of them. This way, we obtain a *fine-tuning set* $\mathcal{F} = \{(\mathbf{x}, \mathbf{G}, \mathbf{z}, q)\}$ exemplifying the human's opinion of generated guidance. These annotations are then used to fit a *surrogate model* $s : \mathbf{z} \mapsto \widehat{q}$, implemented using an appropriate regression architecture. The surrogate is responsible for generalizing the (scarce) human feedback, and can be used to score *any* guidance generated by the LLM. Once the surrogate is trained, we freeze it and use it to fine-tune g by minimizing the following augmented loss for a handful of epochs:

$$\mathcal{L}(g, \mathcal{D}) + \lambda \cdot \mathbb{E}_{(\mathbf{x},\mathcal{G},\mathbf{z},q)\sim\mathcal{F}}[-s(\mathbf{z})] \tag{1}$$

The first term is the regular LLM loss – for instance, the negative log-likelihood of the generated text – on the data set $\mathcal{D}$, while the second one is a novel penalty

term that encourages the model to generate captions obtaining a high score according to the surrogate. Here, $\lambda > 0$ is a hyperparameter. This step increases the overall quality of the generated guidance while making sure that the LLM still outputs sensible captions. The SLOG loop then repeats. Since the LLM's embedding space changes during fine-tuning, the surrogate is fit anew in each iteration. This operation is very cheap in comparison to fine-tuning the LLM itself. If the surrogate manages to properly generalize the human's feedback, the LLM gradually learns to output image captions that work well as textual guidance and that are tailored for the specific task and human expert at hand.

Related Work. Using human guidance to fine-tune large language models has recently become a popular topic. Bazi et al. [2023], Yunxiang et al. [2023], Wang et al. [2023] proposed medical chat models. While Bazi et al. [2023] introduced a specially designed vision transformer, the other two opted to fine-tune already available language models. Seo et al. [2020] presented a method for improving the performance of an image caption generator with offline human feedback. Hou et al. [2021], Chen et al. [2020] focused on machine-driven pathological report generation from chest X-ray images. They experimented with their models using the Mimic-CXR-IV [Johnson et al., 2019] and Indiana University chest X-ray data [Demner-Fushman et al., 2016]. None of these approaches are concerned with fine-tuning LLMs for the purpose of generating textual guidance useful for supporting human decision-making.

3 Empirical Analysis

Data set. We evaluate SLOG on the Mimic-CXR-IV data set [Johnson et al., 2019]. The data consists of 377,100 chest X-ray images and 227,827 corresponding radiology reports. We filtered retained only examples whose reports have information relevant for decision making (specifically, a *findings* or *impression* field). Thereafter, we split the data into train, validation, fine-tuning, and test. In this paper, we use a only subset of this obtained data set for computational ease. Ground-truth human judgments are derived as follows. For each of the labels presented in Irvin et al. [2019], we assign scores of 1, -1, and 0 depending on the presence, absence, and ambiguous mention or missing information of that particular label in the report. Then, for each report, we sum over the labels to obtain an aggregate, numerical information score.

Architectures and metrics. We used the offline LLM architecture developed by Chen et al. [2020]. for generating pathological reports from chest X-ray images. The LLM is a memory-driven transformer with a relational memory layer recording additional information and then used during decoding. The surrogate model itself is a non-linear ridge regression model. Since the latent representation $\mathbf{z}$ of the reports in the training set is not available, we generate it using the pre-trained LLM and weight each example using the BLEU4 score [Papineni et al., 2002] of the corresponding generated text. The surrogate model is then fit on this data using a weighted loss.

Split	#Examples
Train	13,753
Validation	2,889
Test	3073

Table 1. Train, validation, and test split for training the surrogate model

Fig. 2. Training and validation loss of our non-linear ridge regression surrogate model.

Results. The RMSE on the training and test sets are reported in **??**. We observe that the surrogate model quickly fits the training examples (in **red**) while performing well on the validation data (in **blue**). The split of the data is reported in Table 1 When tested on a separate set of 3073 test inputs, the test RMSE is 0.1744, which is well within the variance of the ground-truth human judgments (0.1701). This allows us to conclude that even simple non-linear models can in fact be used to generalize human judgments of textual guidance in this setting.

4 Conclusion

We introduced *learning to guide* as an alternative setup for hybrid decision making that ensures the human is always in the loop, as well as SLOG, an end-to-end approach for fine-tuning an LLM to produce high-quality textual guidance. Our preliminary results suggest that it is in fact possible to generalize human judgments using a surrogate model, supporting the feasibility of our approach. In future work, we plan to properly evaluate the quality of textual guidance that can be obtained using SLOG.

5 Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. The research of ST and AP was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. AP acknowledges support by the project AI@Trento (FBK-Unitn).

Bibliography

Government of Canada. Directive on automated decision-making. 2019.

European Commission. Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act). *eur-lex.europa.eu*, 2021.

David Madras et al. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. *NeurIPS*, 2018.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, 2020.

Vijay Keswani et al. Designing closed human-in-the-loop deferral pipelines. *arXiv:2202.04718*, 2022.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *ICML*, 2022.

Jessie Liu et al. Incorporating uncertainty in learning to defer algorithms for safe computer-aided diagnosis. *Scientific Reports*, 2022.

Bryan Wilder et al. Learning to complement humans. In *IJCAI*, 2021.

Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *AAAI*, 2020.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

Nastaran Okati et al. Differentiable learning under triage. *NeurIPS*, 2021.

Charvi Rastogi et al. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.*, 2022.

Mary Cummings. *Automation Bias in Intelligent Time Critical Decision Support Systems*. 2012. https://doi.org/10.2514/6.2004-6313. URL https://arc.aiaa.org/doi/abs/10.2514/6.2004-6313.

Ben Green. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990, 2022.

Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826, 2021.

Yakoub Bazi et al. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 2023.

Li Yunxiang et al. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv:2303.14070*, 2023.

Sheng Wang et al. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv:2302.07257*, 2023.

Paul Hongsuck Seo et al. Reinforcing an image caption generator using off-line human feedback. In *AAAI*, 2020.

Benjamin Hou et al. Ratchet: Medical transformer for chest x-ray diagnosis and reporting. In *MICCAI*, 2021.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, pages 1439–1449, 2020.

Alistair Johnson et al. MIMIC-CXR-JPG-chest Radiographs with Structured Labels (version 2.0.0). *PhysioNet*, 2019.

Dina Demner-Fushman et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc*, 2016.

Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

Kishore Papineni et al. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.