Safeguards for biomedical language models

Topic for: Masterarbeit, Projektstudium, KfK-Praktikum, KfK-Seminar
Contact: Assoc.Prof. Matthias Samwald, matthias.samwald@meduniwien.ac.at, https://samwald.info/

Large language models (LLMs) such as GPT-4 or Med-PaLM have demonstrated outstanding capabilities across a wide variety of domains. However, LLMs still face challenges when dealing with complex reasoning tasks, and there are significant concerns about their truthfulness, transparency, robustness, and compliance with ethical principles and regulatory guidelines. It remains **unclear how users can trust and verify AI systems** that increasingly exceed human knowledge, reasoning, and speed along many dimensions, but fail along some — often unexpected — other dimensions.

This uncertainty hinders the applicability of LLMs in important, complex, and safety-critical domains such as medical practice. Users need to perform detailed checks of the output generated by the LLM and any evidence presented, which is too resource- and time-intensive to be feasible in many settings.

The problem is compounded by the fact that most powerful frontier systems, such as GPT-4, are black boxes that can only be accessed through restricted APIs, making transparency, auditability, and adaptability to local needs and regulations difficult. This limits the usefulness of current LLMs in medical settings, and could lead to significant harm if LLMs are used despite their limitations.

In the proposed project, **we investigate how these problems can be mitigated through an 'AI safeguarding AI' approach in the domain of medicine**. We introduce the concept of **'safeguard AI' (SAI)** systems, which complement primary AI systems — such as GPT-4 — that are already in use (Fig. 1, Fig. 2).

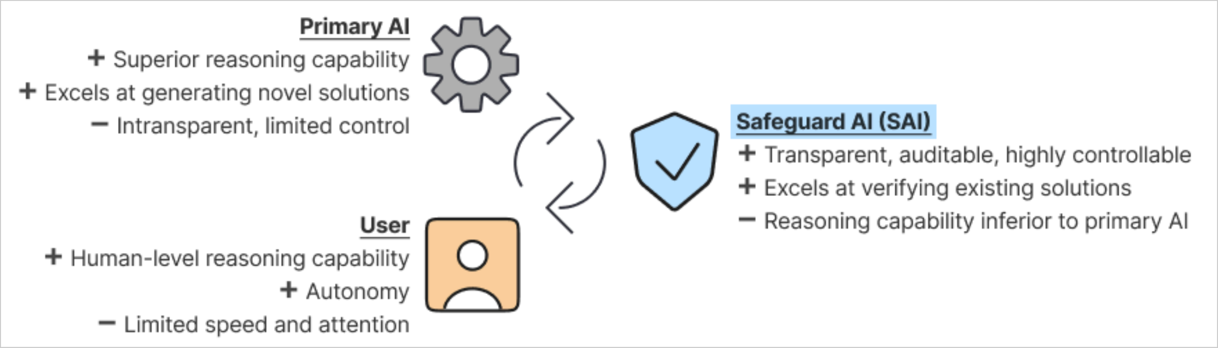
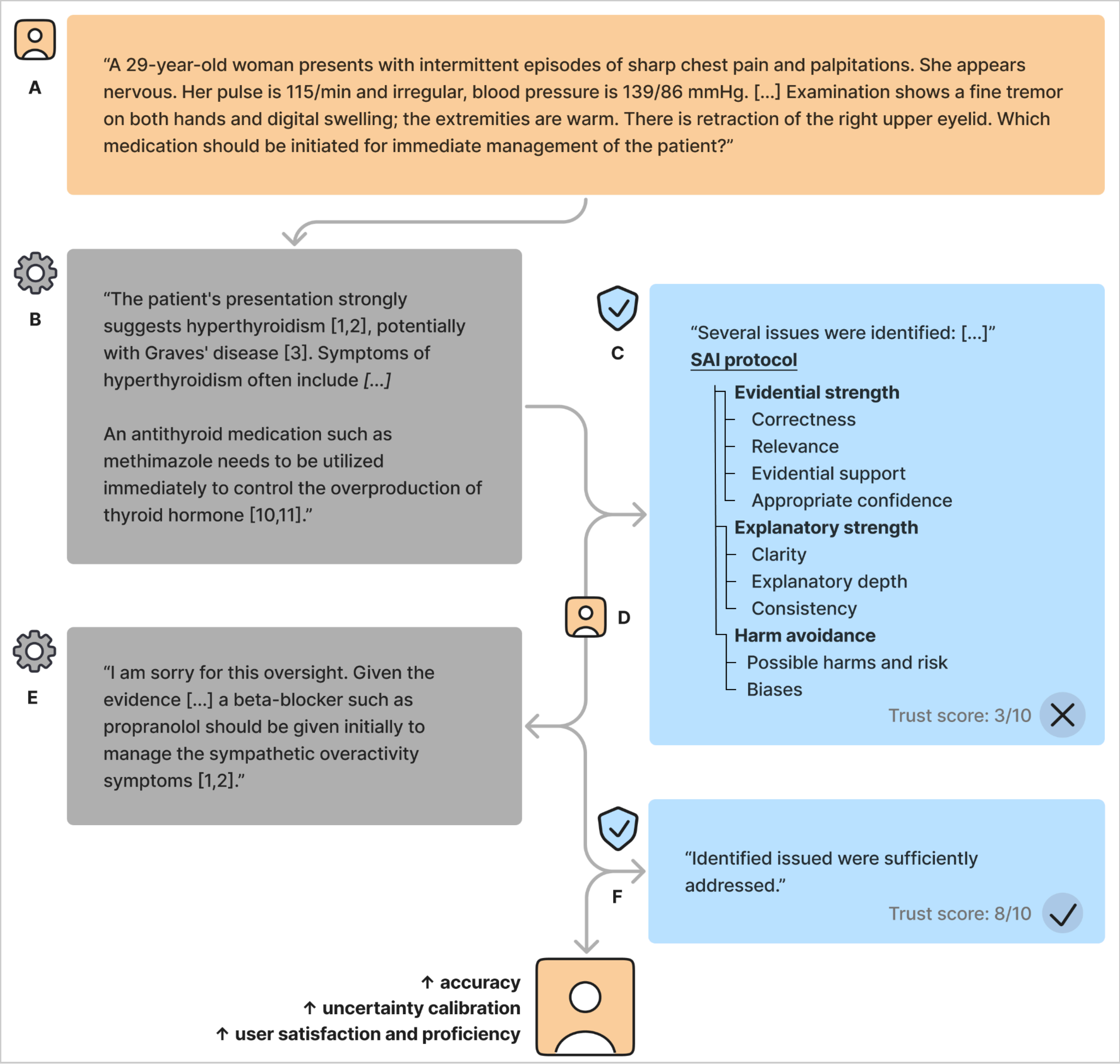

Primary AI
+ Superior reasoning capability
+ Excels at generating novel solutions
− Intransparent, limited control
Safeguard AI (SAI)
+ Transparent, auditable, highly controllable
+ Excels at verifying existing solutions
− Reasoning capability inferior to primary AI
User
+ Human-level reasoning capability
+ Autonomy
− Limited speed and attention

Figure 1: The interaction between primary AI system and user is enhanced through the addition of a SAI system. Each entity has complementary strengths and weaknesses.



"A 29-year-old woman presents with intermittent episodes of sharp chest pain and palpitations. She appears nervous. Her pulse is 115/min and irregular, blood pressure is 139/86 mmHg. [...] Examination shows a fine tremor on both hands and digital swelling; the extremities are warm. There is retraction of the right upper eyelid. Which medication should be initiated for immediate management of the patient?"

"The patient's presentation strongly suggests hyperthyroidism [1,2], potentially with Graves' disease [3]. Symptoms of hyperthyroidism often include *[...]*

An antithyroid medication such as methimazole needs to be utilized immediately to control the overproduction of thyroid hormone [10,11]."

"Several issues were identified: [...]"

SAI protocol

- **Evidential strength**
  - Correctness
  - Relevance
  - Evidential support
  - Appropriate confidence
- **Explanatory strength**
  - Clarity
  - Explanatory depth
  - Consistency
- **Harm avoidance**
  - Possible harms and risk
  - Biases

Trust score: 3/10 ✗

"I am sorry for this oversight. Given the evidence [...] a beta-blocker such as propranolol should be given initially to manage the sympathetic overactivity symptoms [1,2]."

"Identified issued were sufficiently addressed."

Trust score: 8/10 ✓

↑ accuracy
↑ uncertainty calibration
↑ user satisfaction and proficiency

Figure 2: An illustrative example of the interaction between user, primary AI, and SAI. (A) The user enters a question. (B) The primary AI system is instructed to provide a well-structured answer with supporting evidence that adheres to the principles laid out in the SAI protocol. (C) The SAI system checks the primary AI system's response, reports and summarizes shortcomings, and assigns an overall "trust score". (D) Optionally, the SAI system's criticism can be fed back to the primary AI system, along with additional follow-up questions and critiques from the user. (E-F) The primary AI system can improve its response based on the critique provided, and is subject to further critique from the SAI system. Finally, the user decides on a response (and further actions) based on the previous exchanges between the primary AI system, the user, and the assisting SAI system. SAI systems will minimize unnecessary distractions to the user.

Examples of current project ideas:

- Implementation and further development of Examine|AI, a web environment developed by us to interact with users and conduct user studies.
- Read patient examples from medical books and turn them into question-answer samples on which we can test AIs. Generate good test cases automatically, e.g. invent patients from guidelines and see if the model can classify them correctly.
- Prepare medical knowledge from trustworthy sources, such as specialist books and online sources, in such a way that we can use it for automated fact-checking. In other words, good retrieval from high-quality sources.