ADEA: An Argumentative Dialogue Dataset on Ethical Issues concerning Future A.I. Applications

Christian Hauptmann[1], Adrian Krenzer[1,] Antonia Krause[2], and Frank Puppe[1]

[1]Chair for Artificial Intelligence and Applied Computer Science, University of Würzburg

[2]Chair of Criminal Law, Criminal Procedural Law, Legal Theory, Information Law and Legal Informatics

Contact: christian.hauptmann@uni-wuerzburg.de

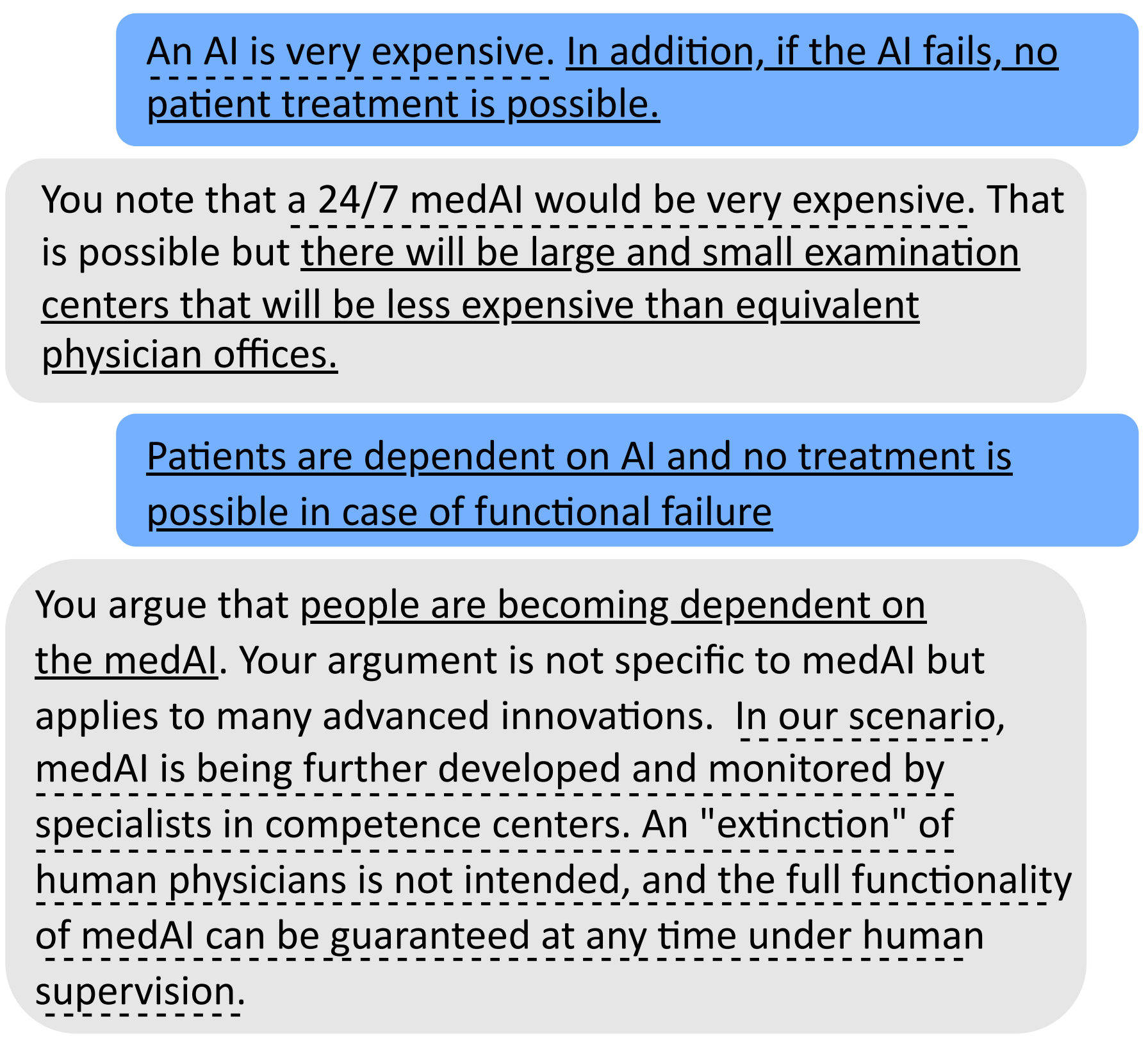
Introduction

Goal

- Develop a German dataset, aimed for training chatbots to engage in ethical AI conversation and help users understand more about that topic.
- Provide a benchmark for evaluating the dataset using GPT-4 via the OpenAI API for further use in AI research.

Material

- Over 2800 labeled dialogues across four topics.
- Features German argument graphs for knowledge and annotation.
- Dataset accessible at: https://github.com/HaupChris/ADEA-Dialogue-Dataset



An AI is very expensive. In addition, if the AI fails, no patient treatment is possible.

You note that a 24/7 medAI would be very expensive. That is possible but there will be large and small examination centers that will be less expensive than equivalent physician offices.

Patients are dependent on AI and no treatment is possible in case of functional failure

You argue that people are becoming dependent on the medAI. Your argument is not specific to medAI but applies to many advanced innovations. In our scenario, medAI is being further developed and monitored by specialists in competence centers. An "extinction" of human physicians is not intended, and the full functionality of medAI can be guaranteed at any time under human supervision.

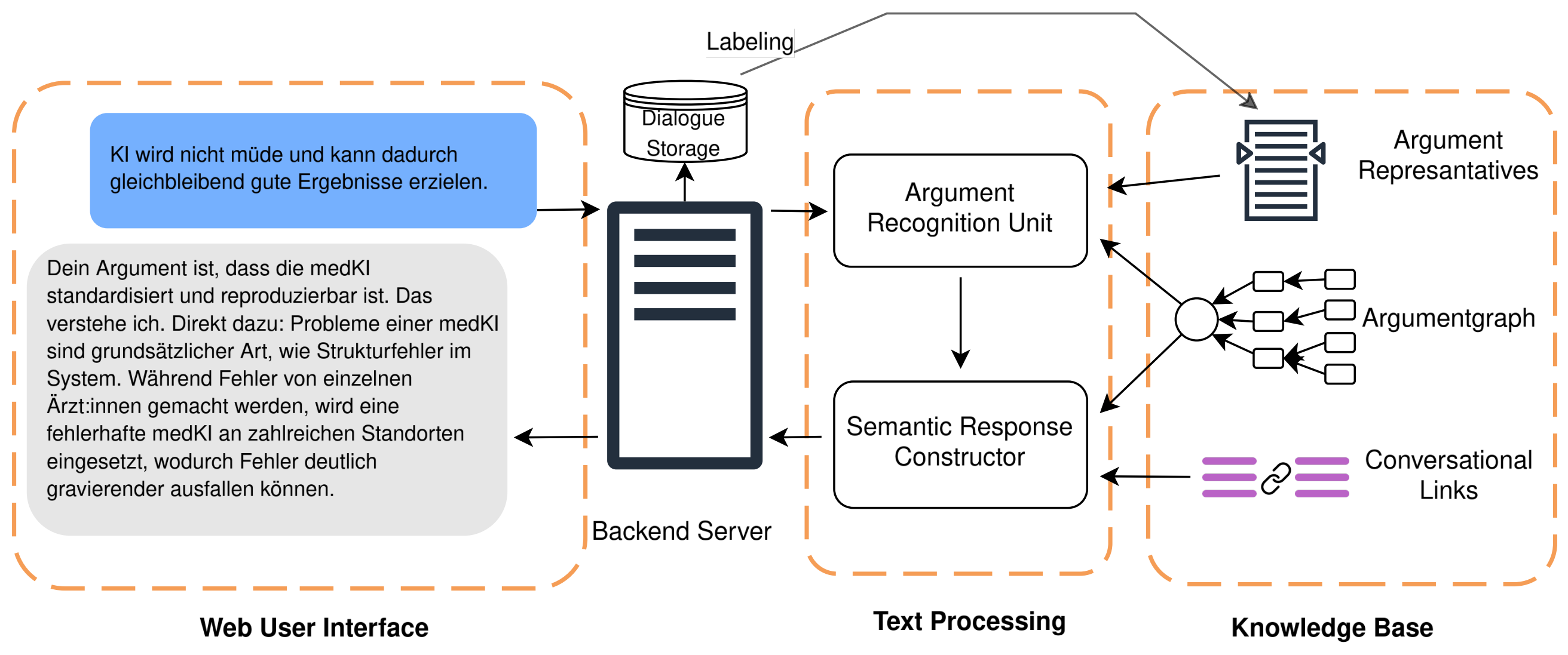
Translated dialogue excerpt with user utterances in blue, underlined arguments differentiated by line styles. Bot responses are templated and auto-annotated; non-underlined units are labeled in the corpus.

Data Collection

Two User Studies
University students participate remotely via smartphone.

Retrieval-based dialogue system
User arguments are identified, acknowledged and countered.

Knowledge Base
Includes scenario and question of discussion, FAQs and an Argument Graph.



KI wird nicht müde und kann dadurch gleichbleibend gute Ergebnisse erzielen.

Dein Argument ist, dass die medKI standardisiert und reproduzierbar ist. Das verstehe ich. Direkt dazu: Probleme einer medKI sind grundsätzlicher Art, wie Strukturfehler im System. Während Fehler von einzelnen Ärzt:innen gemacht werden, wird eine fehlerhafte medKI an zahlreichen Standorten eingesetzt, wodurch Fehler deutlich gravierender ausfallen können.

Overview of the retrieval-based dialogue system [1]

Argument Graphs

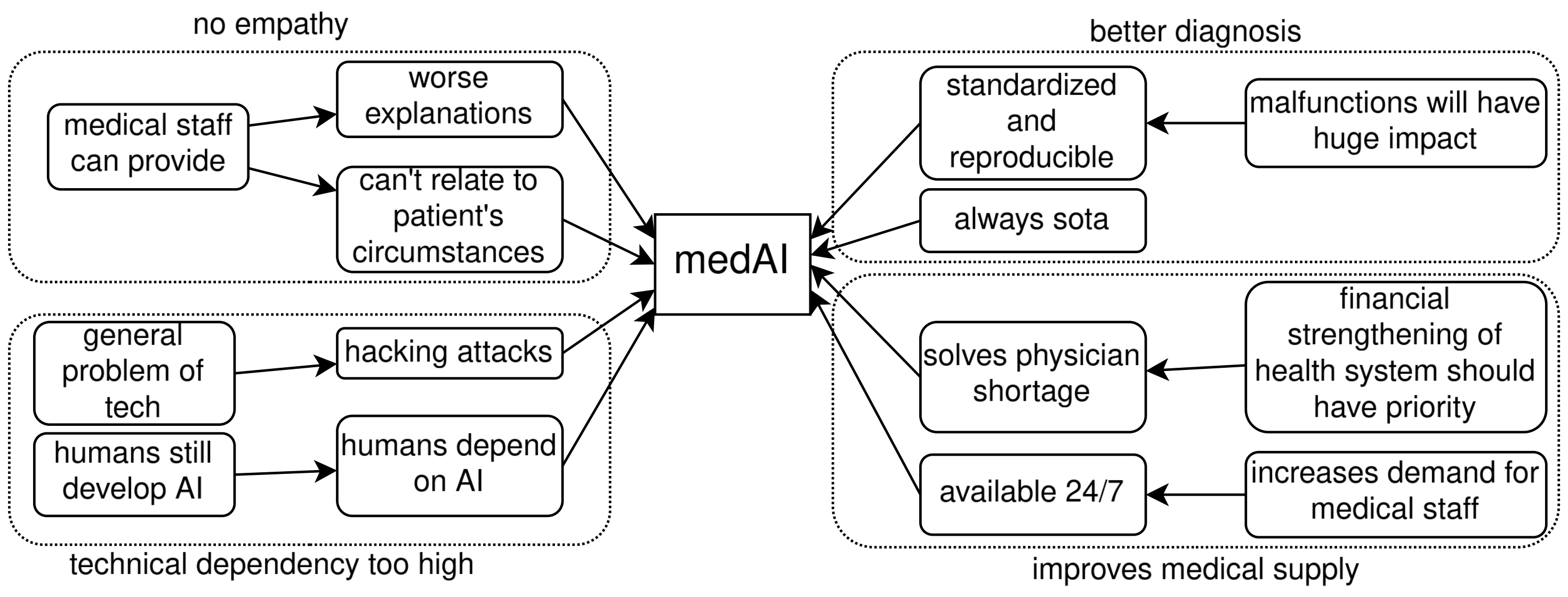
Each graph contains arguments and counter arguments to discuss hypothetical future scenarios in which an AI replaces the following human specialists: Physicians, Judges in civil law processes, car drivers, soccer referees.

Topic	Main	Counter
MedAI	25	33
LawAI	22	45
CarAI	29	50
RefAI	20	58

Number of main and counter arguments for each topic's graph.

Roles of the argument graph

- Maps user utterances to nodes for intent recognition.
- Provides the bot with arguments for responses.
- Serves as annotation scheme for user utterances.


no empathy
medical staff can provide
worse explanations
can't relate to patient's circumstances
general problem of tech
hacking attacks
humans still develop AI
humans depend on AI
technical dependency too high
medAI
better diagnosis
standardized and reproducible
malfunctions will have huge impact
always sota
solves physician shortage
financial strengthening of health system should have priority
available 24/7
increases demand for medical staff
improves medical supply

Dataset Annotation

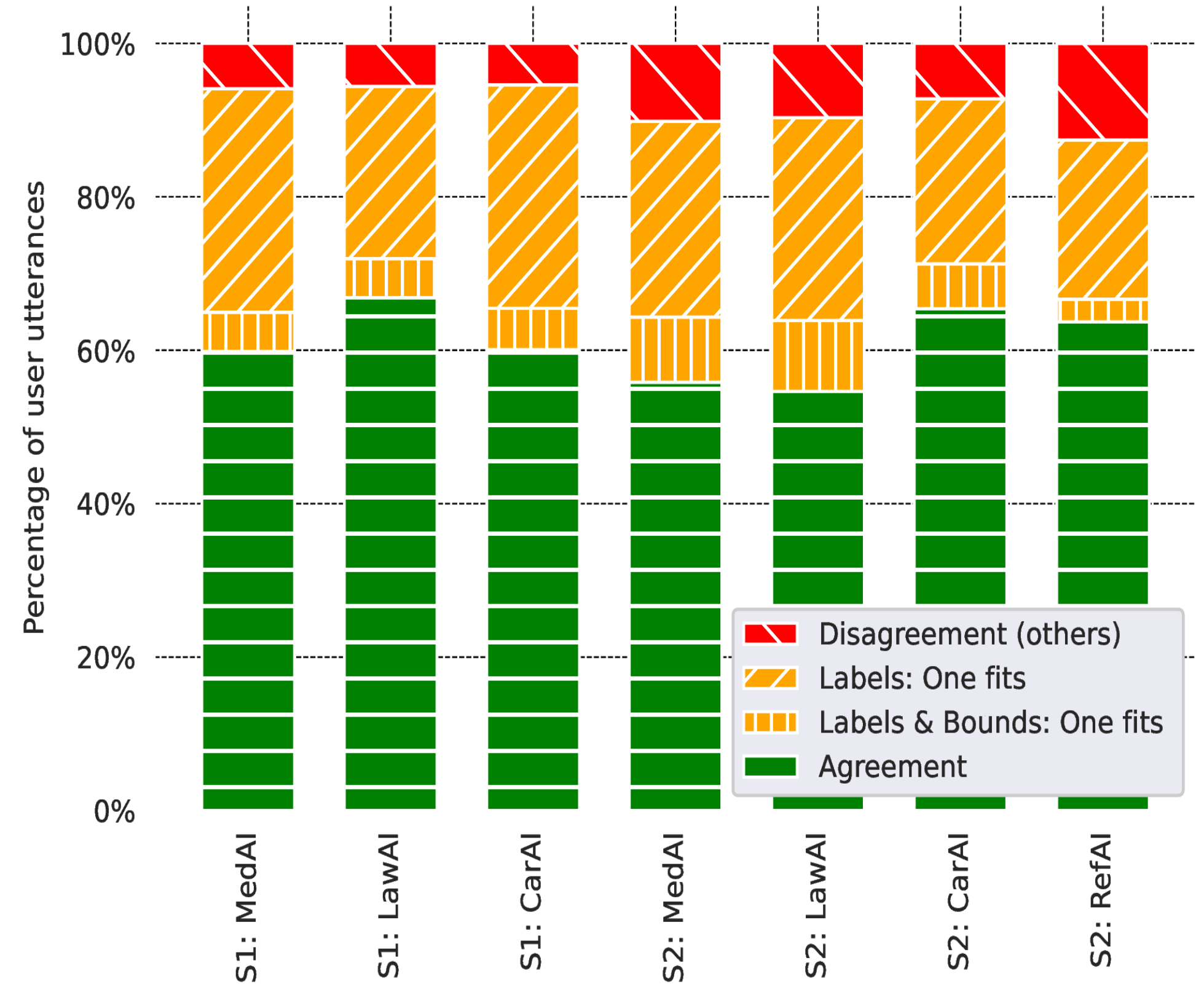
Two-Stage Annotation
Labelling of text segments and boundaries, similar to Stab et al. [2]

Annotation Types

- Well-founded Arguments
- Unfounded arguments
- Non argumentative units
- Miscellaneous

After 1st stage

- Cohens Kappa > 0.58 label agreement for all topics
- Observed boundary agreement > 92% for all topics



Inter Annotator Agreements by Topic: 'One Fits' indicates resolution by a third annotator agreeing with a prior annotation. Other disagreements occur when no prior annotations are chosen.

Dataset Statistics

Depth of Discussion
Dialogues average 7.8 turns with 12.3 words per utterance.

Argumentative Variety
Users present 4.2 unique arguments on average per dialogue.

Real-World Application
Dataset reflects real-world dialogues.

Study	Topic	Dialogue	User Utterances			Distinct Args. per Dia.			Arguments		Non-Arguments	
		Count	Count	Avg. Words	Avg. Count	User	Bot	Union	WF	UF	Q	Misc
1	MedAI	62	519	12	8.4	3.6	6.6	8.9	58.96	10.79	4.43	26.20
	LawAI	26	203	13.2	7.8	4.1	9.4	12.1	67.00	5.91	1.48	26.11
	CarAI	90	834	11.1	9.3	4.2	5.3	8.9	70.02	6.24	0.48	23.38
2	MedAI	82	534	14	6.5	4.1	10.3	11.9	61.05	13.86	5.99	20.41
	LawAI	33	227	13.7	6.9	4.6	8.2	9.6	78.41	3.96	2.20	15.42
	CarAI	58	428	12.6	7.4	5	9.6	11.3	76.17	3.97	0.70	19.16
	RefAI	27	135	9.2	5.0	3.1	4.8	6.3	64.44	2.22	3.70	30.37
1 + 2	Total	378	2880	12.3	7.8	4.2	7.6	10	-	-	-	-

Overview of dialogue and utterance statistics across topics, including types and percentages of well-founded (WF), unfounded (UF) arguments, questions (Q), and miscellaneous (Misc) responses.

Benchmark: User Utterance Classification

Objective
Classify user utterances into argument labels or as 'misc' (non-argumentative).

Method
Use OpenAI GPT-4 API for text classification with one-shot prompts.

Results

- Outperformed the majoritiy baseline but with modest accuracy
- Accuracy declines with longer or multi-label utterances
- Accuracy of 'misc' exceeds overall accuracy
- Dialogue context will probably improve performance

Study	Topic	Maj. Baseline	GPT-4
1	MedAI	0.32	0.54
	LawAI	0.28	0.46
	CarAI	0.3	0.52
2	MedAI	0.27	0.5
	LawAI	0.12	0.51
	CarAI	0.11	0.51
	RefAI	0.3	0.51

Conclusion

Argument Graphs
Introduced for German AI ethics discussions.

Annotated Corpus
Utilized two-stage annotation process.

Benchmark
Evaluation to measure dataset performance.

Dataset Utility
Identifying argumentative content, stance classification, segmentation of user utterances.

Future Work

Expand Topics
Include more topics about AI ethics to capter more parts of society.

References

[1] Hauptmann, Christian, et al. "Argumentation effect of a chatbot for ethical discussions about autonomous AI scenarios." Knowledge and Information Systems (2024): 1-31.

[2] Stab, Christian, and Iryna Gurevych. "Annotating argument components and relations in persuasive essays." Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers. 2014.