Identifying Documented Goals of Care Conversations Using Definition-informed Chain-of-Thought Prompting (DiCoT)

Kevin S. Li, MS[1], Trevor Cohen, MBChB, PhD [1,2], William B. Lober, MD, MS [1,2,4,5], Robert Y. Lee, MD, MS [1,3]

[1] Division of Biomedical and Health Informatics, Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA
[2] Division of Pulmonary, Critical Care, and Sleep Medicine, University of Washington, Seattle, USA
[3] Cambia Palliative Care Center of Excellence at UW Medicine, University of Washington, Seattle, USA
[4] Department of Biobehavioral Nursing and Health Informatics, University of Washington, Seattle, WA
[5] Department of Global Health, University of Washington, Seattle, WA

UW Medicine
UNIVERSITY of WASHINGTON

Background

Motivation

- Goals of care discussions (GoCD) facilitate alignment of care delivery with patients' priorities, yet their documentation varies
- Automated detection of documented GoCD can facilitate scalable measurement of this important process metric
- However, varied documentation creates challenges for Natural Language Processing (NLP) methods to apply to these clinical notes.
- BERT has been previously applied towards with notable performance, but requires extensive manually annotated training data—an expensive process that must be repeated as definitions evolve[1]

Research Objective

- Generative LLMs have been shown to perform well on a wide variety of tasks without the need for annotated training data, along side a considerably larger context window compared to BERT
- Capabilities of LLMs on the task of identifying GoCD remain to be determined

We propose a framework for a novel GoCD-focused version of Chain-of-Thought prompting, ***Definition-informed CoT (DiCOT)****, injecting definitions of key concepts into a Chain-of-Thought prompt that encourages the LLM to generate reasoning steps before reaching a solution.*

Method

- *Data Sources*: Datasets from three trials in the Project to Improve Communication about Serious Illness (PICSI) series:
  - PICSI Pilot study: 4,642 EHR notes from a sample of 150-patient
  - PICSI Hospital Trial 1: 2,974 EHR notes from a sample of 160 patients
  - PICSI Hospital Trial 2: 11,574 EHR notes from 617 patients.
- *LLM model*: Llama 3.3 70b
- *NLP Development and Prompt Engineering*:
  - Iteratively refined our approach from zero-shot prompting by incorporating medical domain knowledge, enforcing structured output formatting, and adding explicit reasoning steps

Prompt Design

Baseline (Zero-shot) Prompt:

1. You are a medical expert. Using the information provided in the INPUT, does the INPUT contain discussions on [GOALS OF CARE, PALLIATIVE CARE, AND END OF LIFE CARE] content? Please answer YES/NO.
Use the following example format:
ANSWER: [YES/NO]

CoT Prompt:

2. Before we begin, let's define a few key concepts:
- GOALS-OF-CARE DISCUSSIONS are defined as discussions between clinicians and patients/families about the patient's aims of medical care, which are ideally informed by their underlying values and priorities, established within the current clinical context, and used to guide decisions about the use of or limitations on life-sustaining treatments.
- CODE STATUS refers to inpatient orders for a patient that specify what is to be done in the event of cardiac or respiratory arrest--i.e., whether to perform CPR or not ("DNR"), whether to intubate or not ("DNI"). Code status discussions are one type of goals-of-care discussion, albeit one of narrow scope.
- "Advance care planning documents" are structured documents such as advance directives and POLST forms that may provide insight into a patient's goals of care or treatment preferences for future health states, as well as durable power of attorney (DPOA) forms. One component of advance care planning, the designation of a "surrogate decision-maker" or "durable power of attorney", refers to designation by a patient of a proxy who can make healthcare-related decisions on their behalf if they are incapacitated.

Our research outcome, COMPOSITE_JS, is defined as ANY of three constructs: (1) goals-of-care discussions that extend -beyond- decisions relating to code status alone; (2) incident designation or review of surrogate decision-maker or durable power of attorney (not just citations of past documentation); and, (3) incident completion or review of advance care planning documents such as advance directives or POLSTs (not just citation of prior documents). COMPOSITE_JS should -not- include text that only references statements aspiring to future goals-of-care discussions, treatment-specific therapeutic goals, or documentation focused on discharge or disposition planning (e.g., plans for post-discharge placement).

Using the medical record excerpts provided in INPUT, do the excerpts in INPUT reflect documentation of a discussion meeting criteria for COMPOSITE_JS? Consider the evidence and assign an impact score adding up to 10, where higher scores indicate stronger influence on the final decision. If key elements are missing (e.g., no documented goals-of-care discussion, no incident designation/review of a surrogate, or no review/completion of advance care planning documents), assign higher scores to reflect their critical absence.
Please only answer YES or NO.
Use the following example format:
ANSWER: [YES/NO]

Results

- CoT prompting improves LLM performance (Table 1)
- CoT prompting performs comparably to, or better than BERT (Table 1)

Trial Dataset	Strategy	Sensitivity	Specificity	PPV	NPV	F1	AUC	AUPRC
Note level								
PISCI-H1	BERT	**0.956**	0.794	0.339	0.994	0.500	**0.962**	0.793
	Zero-Shot	0.747	0.934	0.554	**0.971**	0.637	0.940	0.733
	CoT	0.677	**0.976**	**0.757**	0.965	**0.715**	0.961	**0.795**
PICSI-H2	BERT	**0.961**	0.906	0.629	**0.993**	0.760	**0.981**	**0.879**
	Zero-Shot	0.780	0.936	0.668	0.962	0.719	0.941	0.782
	CoT	0.806	**0.972**	**0.825**	0.968	**0.815**	0.979	0.873
Average of datasets	BERT	**0.958**	0.850	0.484	**0.993**	0.630	**0.971**	**0.836**
	Zero-Shot	0.764	0.935	0.611	0.967	0.678	0.941	0.758
	CoT	0.741	**0.974**	**0.791**	0.966	**0.765**	0.970	0.834
Patient Level								
PICSI-H1	BERT	**1.000**	0.804	0.378	**1.000**	0.548	**0.959**	0.719
	Zero-Shot	0.741	0.863	0.754	0.854	0.748	0.911	0.867
	CoT	0.655	**0.961**	**0.905**	0.831	**0.760**	0.915	**0.886**
PICSI-H2	BERT	**0.976**	0.830	0.674	0.990	0.797	**0.981**	**0.959**
	Zero-Shot	0.822	0.916	0.779	0.935	0.800	0.943	0.907
	CoT	0.841	**0.976**	**0.926**	**0.945**	**0.881**	0.978	0.956
Average of datasets	BERT	**0.988**	0.817	0.526	**0.995**	0.673	**0.970**	0.839
	Zero-Shot	0.782	0.890	0.767	0.895	0.774	0.927	0.887
	CoT	0.748	**0.968**	**0.915**	0.888	**0.821**	0.947	**0.921**

Table 1: Performance metrics including sensitivity (recall), specificity, positive predictive value (PPV), negative predictive value (NPV), F1-score, area under the receiver operating characteristic curve (AUC-ROC), and area under the precision-recall curve (AUPRC) stratified by trial dataset and strategy.

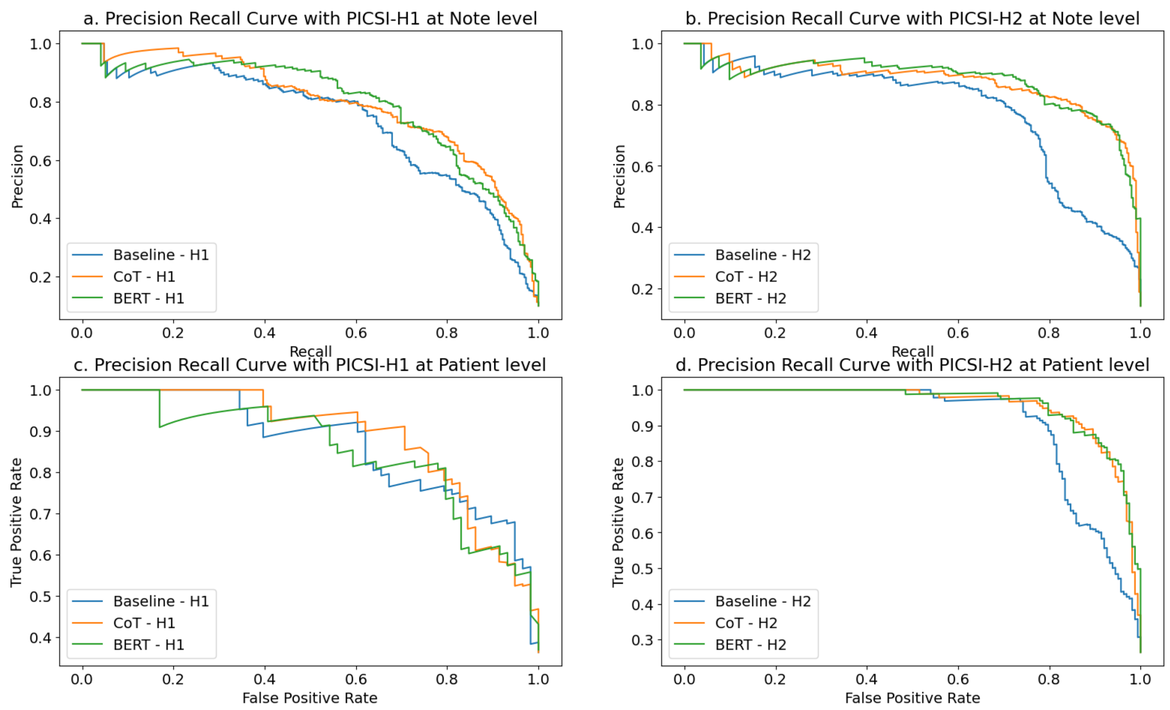

Figure 1 Precision Recall Curve plots of a) PISCI-H1 data at Note Level, b) PISCI-H2 data at Note Level, c) PISCI-H1 data at Patient Level, d) PISCI-H2 data at Patient Level

Conclusion

Discussions

- LLMs using DiCoT prompting strategy achieved comparable performance to state-of-the-art fine-tuned models for identifying documented GoCD in clinical narratives, despite requiring no expensive human annotation of thousands of training notes.
- These results suggest that with appropriate prompting, LLMs offer a practical approach for automated detection of complex linguistic constructs in clinical text, enabling more nuanced measurement of clinical and process metrics to improve patient-centered care quality.

Limitations/Future Work

- Conducted experiments with one LLM model
  - Raises concerns about generalizability
- Comparative analysis was limited to a single fine-tuned BERT-based model as the primary baseline
- Explore whether smaller, more efficient LLMs can achieve comparable performance

References

1. Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, et al. Assessment of Natural Language Processing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. JAMA Netw Open. 2023 Mar 2;6(3):e231204.