Title	法的言語モデルにおける幻覚の検出と理解：不確実性と内部シグナルに基づく複合的手法の研究
Author(s)	DANG HOANG ANH
Citation	
Issue Date	2025-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/20074
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

氏名	Dang Hoang Anh		
学位の種類	博士（情報科学）		
学位記番号	博情第 556 号		
学位授与年月日	令和 7 年 9 月 24 日		
論文題目	Detecting and Understanding Hallucinations in Legal Language Models: A Multi-Method Study of Uncertainty and Internal Signals		
論文審査委員	Nguyen Le Minh	JAIST	Professor
	Kiyoaki Shirai	JAIST	Professor
	Shinobu Hasegawa	JAIST	Professor
	Naoya Inoue	JAIST	Associate Professor
	Ken Satoh	NII	Director

論文の内容の要旨

Detecting and understanding hallucinations in large language models (LLMs) is crucial, especially within high-stakes domains such as law, where the reliability of information significantly impacts decision-making processes. This thesis addresses the critical challenge of hallucination detection in legal language models by leveraging uncertainty measures and internal model signals, developing a comprehensive multi-method framework specifically designed for legal question-answering (QA) tasks. Initially, this research investigates prominent existing hallucination detection techniques, focusing on semantic uncertainty, reference-based verification (RefChecker), and internal model checks (LLM-Check). It identifies limitations such as dependency on external references and susceptibility to confident yet incorrect assertions, motivating the development of a novel detection pipeline that combines these insights. The thesis introduces an internal scoring system tailored specifically for legal QA. This system exploits semantic entropy, attention patterns, and hidden-state embeddings from within the LLMs themselves, requiring no external reference and only a single forward pass for each query. Experimental evaluations demonstrate that this method significantly outperforms baseline approaches in accuracy, precision, recall, and computational efficiency, successfully identifying subtle and confidently stated hallucinations. Furthermore, the thesis contributes by developing and adapting the CUAD-QA dataset, converting real world commercial contracts into a robust evaluation benchmark that rigorously tests various hallucination detection methods. Extensive experiments conducted on this dataset highlight the effectiveness of the proposed internal scoring system, emphasizing practical applicability and superior performance compared to conventional methods. Finally, this research provides insights into the behavioral patterns of hallucinations in legal QA, pinpointing linguistic and structural cues associated with model uncertainty and errors. These findings inform guidelines for designing trustworthy LLM-based legal assistants, underscoring the importance of internal consistency checks, transparent confidence scores, and effective integration of detection signals. This thesis thus offers substantial advancements in the understanding and mitigation of hallucinations in legal LLM applications, laying the groundwork for more reliable and transparent AI tools in the legal domain.

Keywords: Hallucination Detection, Large Languge Models (LLMs), Legal Question Answering, Internal Scoring Signals, CUAD-QA dataset

論文審査の結果の要旨

This thesis addresses hallucination detection in legal LLMs with a reference-free, single-pass internal scoring approach that fuses semantic entropy, attention-pattern signals, and hidden-state embeddings. It begins with an excellent survey of state-of-the-art techniques—including recent work published in **Nature** and at **NeurIPS**—and motivates a method tailored to the legal domain. Buiding the data for the legal domain, the candidate can conduct complementary evaluations, the proposed approach consistently outperforms strong baselines in detection accuracy while reducing latency/compute and is particularly effective at identifying confident-but-incorrect answers. The work shows a practical, auditable detection pipeline and a contract-law benchmark adapted for legal hallucinations. In addition, the thesis introduces a legal-domain perplexity score and combines it with three existing internal metrics; the candidate fine-tuned a compact LLM to compute this score. Empirical results show further gains in legal hallucination detection when this score is integrated into the pipeline. The method is evaluated across multiple LLM families using the dataset developed in this study.

Alongside the thesis, the candidate has published in reputable international conferences and **journals** and achieved **top-ranked systems** in the **COLIEE** competitions.

Overall, this is an excellent dissertation, and we approve of awarding a doctoral degree to Mr. Dang Hoang Anh.