

## Development of techniques for reasoning with knowledge in language models

**Principal Investigator**  
James Thorne

**Department**  
Kim Jaechul Graduate School of AI

**Co-Investigators**  
Na Min An, Max Glockner, Iryna Gurevych, Noah Lee, Philhoon Oh, Yejoon Lee, Ieva Staliūnaitė, Gisela Vallejo, Andreas Vlachos

**Homepage**  
<https://xfact.net>

It is critical to validate whether large language models (LLMs) are consistent with reasoning and understand how they reason when presented with facts and external knowledge. This project scrutinizes factual reasoning inside LLMs, particularly how they utilize different forms of knowledge for tasks such as question answering and fact-checking. Prof. James Thorne's research group studied how LLMs handle ambiguity in tasks such as fact-checking, where evidence may be conflicting, and released AmbiFC, a reference benchmark of fact-checked claims which require contextual understanding and reasoning under uncertainty. The research group developed efficient methods for integration, demonstrating that the factual accuracy and run-time efficiency of models can be improved when using fewer search results by filtering out detrimental information and summarizing the retrieved information.

### Background

The ability of large language models (LLMs) to store and recall information enables coherent and fluent responses to be generated by models. However, this information, once embedded within the parameters of the model, becomes static and difficult to scrutinize or modify. Consequently, this leads to a risk where LLMs might perpetuate outdated facts or, worse, produce erroneous statements called 'hallucinations.' For tasks that demand a high degree of knowledge fidelity, such as fact-checking or answering questions, it is necessary to retrieve information from a corpus of trusted facts which provides the necessary evidence to ensure the responses are factual. While providing models with evidence can improve answer accuracy, the presence of irrelevant or excessive information in search results can be detrimental, as it may lead to confusion within the model and result in wrong answers too. For application to challenging real-world settings such as fact-checking, it is necessary to reason in cases where this information in the search results can self-conflict.

### Description

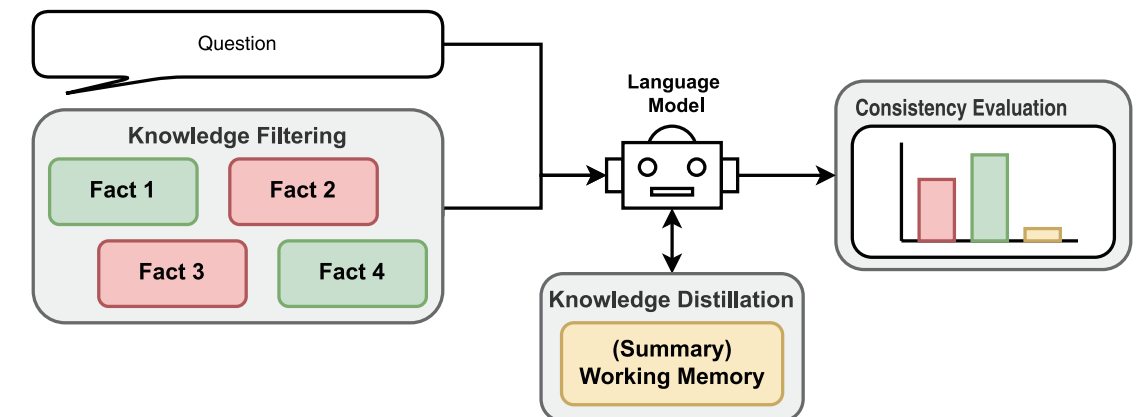
Prof. Thorne's research group developed and introduced a specialized dataset, AmbiFC, to train and evaluate models for fact-checking claims against evidence. This dataset consists of 10,000 textual claims which can be validated from 50,000 passages from Wikipedia. The core differentiating feature of this dataset is circumstantial reasoning where the circumstances and evidence selected can change the outcomes of a fact-check. The research group additionally studied the stability and self-consistency of language models when performing factual reasoning tasks, demonstrating even in non-ambiguous cases, the language models capture a range of reasoning skills that differ from human reviewers. Finally, the research group studied novel methods for

the integration of evidence in language models for knowledge-intensive reasoning tasks, such as question answering and fact-checking. The research demonstrated that by adaptively filtering information that causes inconsistent predictions in models and by performing summarization of search results, it is possible to achieve more accurate question answering in language models with a substantially smaller amount of input evidence, resulting in faster, more stable, and more efficient inference.

### Implications

While methods such as retrieval augmented generation (RAG) are gaining popularity in building adaptable LLMs that reason with changing world-knowledge, it is critical that the evidence is effectively used to generate factual answers. The investigations conducted by the research group revealed the challenging interactions between evidence and the responses generated by large language models. The research demonstrated that models could achieve efficient and accurate question answering with a fraction of the evidence typically used, illustrating the efficiency gains of precision over volume. The AmbiFC benchmark will encourage the development of more sophisticated models capable of nuanced understanding and interpretation. Combined with the group's evaluation of LLM consistency and integration of knowledge for factual reasoning tasks, Prof. Thorne anticipates improvements in model reliability and trustworthiness.

Diagram



### Research outcomes

M Glockner, I Staliūnaitė, J Thorne, G Vallejo, A Vlachos, I Gurevych AmbiFC: Fact-Checking Ambiguous Claims with Evidence, Transactions of the Association for Computational Linguistics, 2024.  
N Lee, NM An, J Thorne, Can Large Language Models Capture Dissenting Human Voices?, Empirical Methods for Natural Language Processing 2023.  
P Oh, James Thorne, Detrimental Contexts in Open-Domain Question Answering, Findings of Empirical Methods for Natural Language Processing 2023.  
Y, Lee, P Oh, James Thorne, Knowledge Corpus Error in Question Answering, Findings of Empirical Methods for Natural Language Processing 2023.

### Research funding

KAIST New Faculty Startup Finding (G04220030)  
The AmbiFC dataset was collected with the support of Google