NLP&AI 연구실 세미나 (08/22, Thu)

LLMs' Long-chain Hallucination

구선민

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

Papers

Deceptive Semantic Shortcuts on Reasoning Chains: How Far Can Models Go without Hallucination?

Bangzheng Li[1] **Ben Zhou**[2] **Fei Wang**[3] **Xingyu Fu**[2] **Dan Roth**[2] **Muhao Chen**[1]
[1]University of California, Davis [2]University of Pennsylvania
[3]University of Southern California
bzhli@ucdavis.edu

DFA-RAG: Conversational Semantic Router for Large Language Model with Definite Finite Automaton

---

Yiyou Sun [1 2] **Junjie Hu** [1] **Wei Cheng** [2] **Haifeng Chen** [2]

Deceptive Semantic Shortcuts on Reasoning Chains: How Far Can Models Go without Hallucination?

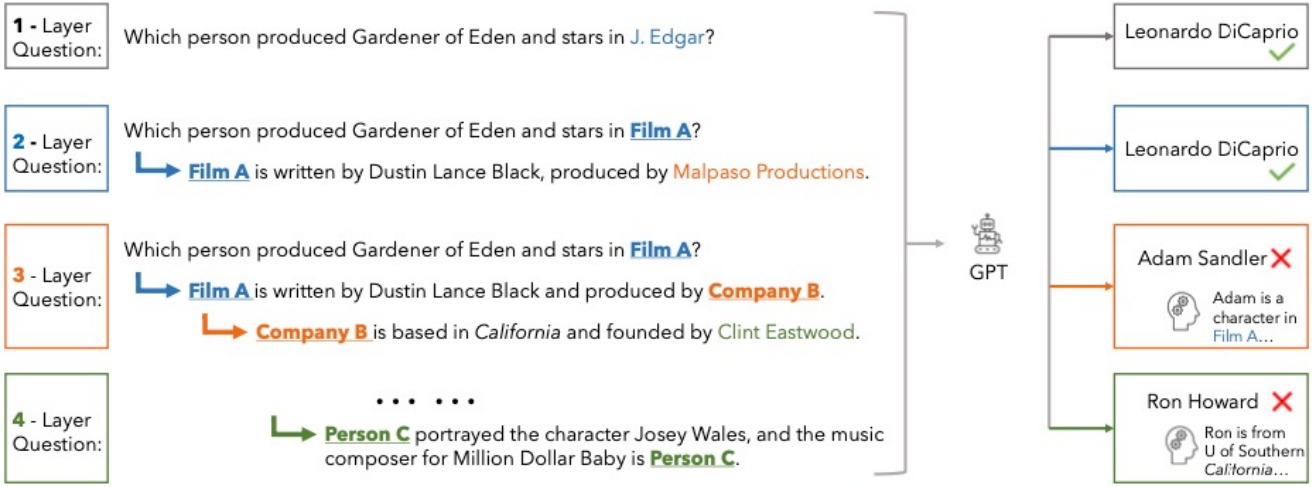
Bangzheng Li[1] **Ben Zhou**[2] **Fei Wang**[3] **Xingyu Fu**[2] **Dan Roth**[2] **Muhao Chen**[1]
[1]University of California, Davis [2]University of Pennsylvania
[3]University of Southern California
bzhli@ucdavis.edu

NAACL 2024

Motivation

LLMs 은 reasoning chains 의 깊이가 깊어짐에 따라 hallucination 정도가 어떻게 변해갈까?

합리적인 reasoning paths 에 기반을 두고 있는지 vs. 모델이 semantic associations (즉, 사전 학습 데이터의 단어 분포)에 의존해서 그럴듯한 답을 생성하는 것은 아닌지?



→ EUREQA (Extending Underlying reasoning Chains in QA) 제안

EUREQA

Reasoning Chain

- EUREQA 은 implicit reasoning chain 을 통해 구성

- chain은 layer로 구성되며, 각 layer는 3가지 구성요소로 구성

  entity e_i : 엔티티

  각 엔터티 ei는 KB에 "엔터티"로 존재해야 함

  relation r_i : 현재 layer의 e_i 와 체인의 다음 layer의 e_{i+1} 간의 관계

  각 relation ri는 ei를 2개 이상의 잠재적 후보 엔터티 ei+1과 연결해야

  이렇게 하면 ei와 ei+1 사이에 직접적인 semantic shortcut 이 없음

  e.g. , "Tiger Woods" 엔티티에 대해 "award" rel. O / "college" rel. X

  an associated **fact** f_i about e_i : 엔티티에 연관된 fact

  fact selection 에 easy and hard 2가지 기준 적용

  easy: f_i만으로 e_i 를 식별하기에 충분함

  e.g. , 자식 이름

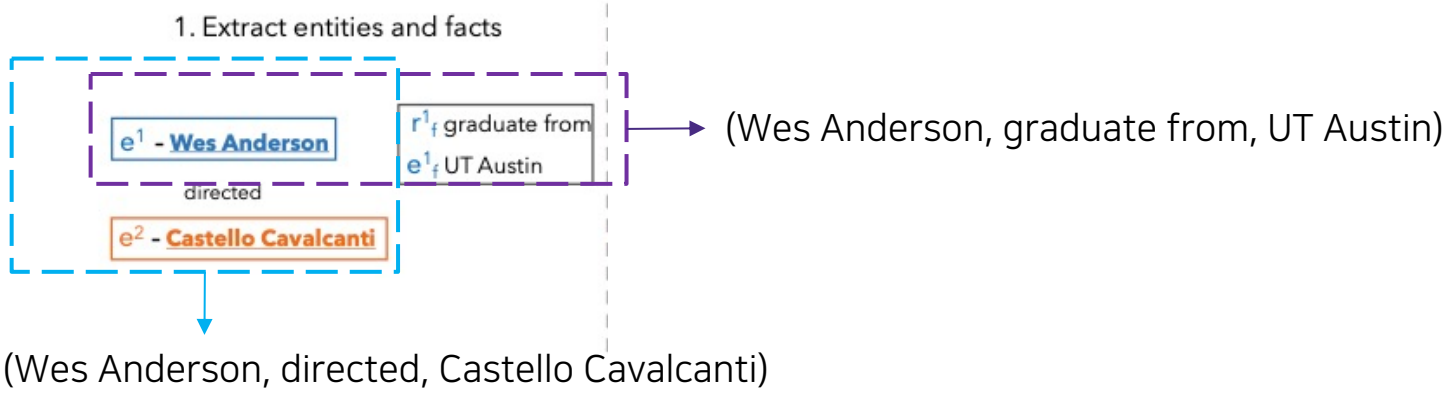
  hard: f_i만으로 e_i 를 식별할 수 없음

  e.g. , 동일한 이름 가진 여러 회사

EUREQA

Chain Construction

- random-walk 알고리즘을 사용하여 리즈닝 체인 생성



EUREQA

Chain Construction

- Reasoning-dependent
  모든 중간 레이어는 해결을 위해 체인의 후속 계층의 정보에 의존
  결과적으로 모델은 최종 레이어에서 시작하여 체인 전체에서 순차적인 추론에 참여해야 함

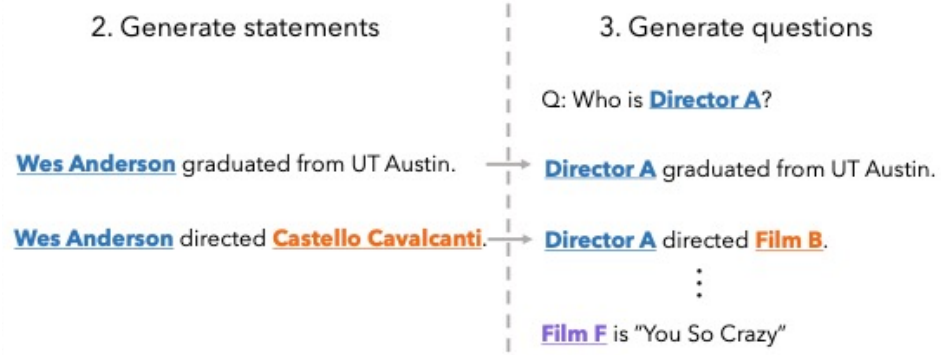
- Length-flexible
  리즈닝 체인의 범위는 레이어를 추가하거나 제거하여 쉽게 조정할 수 있으므로
  리즈닝 깊이를 평가 가능함

- Determinism-adjustable
  Determinism of the reasoning chain 은 레이어에서 fact f_i를 생략하여 변경할 수 있음
  여러 잠재적 답변이 있는 질문을 다루는 모델 평가를 용이하게 함

EUREQA

Question Generation

- QG 과정
  1) 구조화된 체인을 사람이 읽을 수 있는 텍스트로 변환
  2) 모든 문장의 각 엔터티 e를 placeholder로 대체
  3) q_0로 표현되는 의문문과 q_n에 관련된 entity information statement (m_{n+1}) 생성



EUREQA

Data Statistics

- Hard 의 5 미만 layer는 5 layer question 기반으로 layer 제거한 것

Difficulty	easy	hard	hard	hard	hard	hard
#Layers	5	5	4	3	2	1
Count	428	1363	300	300	300	300

Table 1: Statistics of EUREQA.

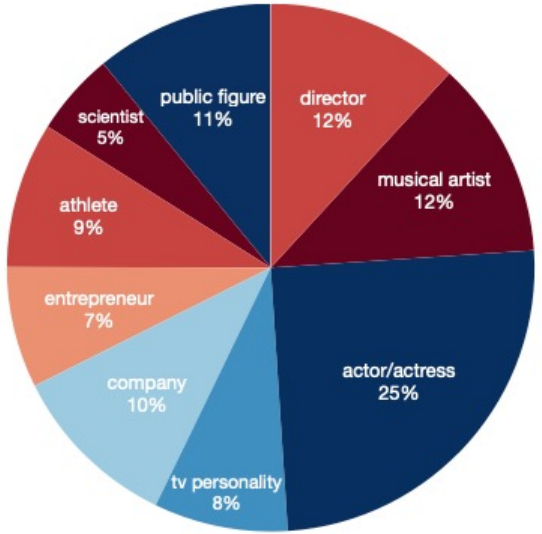


Figure 4: Categorical distribution of seed entities in questions of EUREQA.

Experiment Setup

- ChatGPT (gpt-3.5-turbo-0301), Gemini- Pro (Gemini 1.0 Pro) and GPT-4 (gpt-4-0314) 대상

- Prompting Methods
  direct : 추가적인 context 없이 raw questions만 모델에게 제공
  icl: 쿼리 앞에 2개의 context-specific examples 제공
  humans 이 작성한 solution

- accuracy 로 평가

Results

- 일반적으로 엔티티가 추론 체인 레이어의 설명으로 재귀적으로 대체되어 표면 수준의 의미적 단서가 제거되면 이러한 모델은 더 많은 잘못된 답변을 생성

- semantic shortcuts 응답의 정확도에 상당한 영향을 미친다는 사실을 강조하며, GPT-4가 이러한 shortcuts 를 식별하고 활용하는 능력이 훨씬 더 뛰어나다는 사실도 나타냄

- LLM이 depths of reasoning 가 깊어질 수록 항상 성능이 더 낮아지는 것은 아님
  → reasoning path 가 아닌 surface-level semantic shortcuts 에 따라 질문에 답할 수도 있음

depth	hard										easy	
	d=1		d=2		d=3		d=4		d=5		d=5	
	direct	*icl*	*direct*	*icl*	*direct*	*icl*	*direct*	*icl*	*direct*	*icl*	*direct*	*icl*
ChatGPT	22.3	53.3	7.0	40.0	5.0	39.2	3.7	39.3	7.2	39.0	13.1	47.0
Gemini-Pro	45.0	49.3	29.5	23.5	27.3	28.6	25.7	24.3	17.2	21.5	30.6	38.9
GPT-4	60.3	76.0	50.0	63.7	51.3	61.7	52.7	63.7	46.9	61.9	66.4	81.8

Table 2: Accuracy of ChatGPT, Gemini-Pro and GPT-4 across different depths d of reasoning (number of layers in the questions) as well as the difficulty of the questions. We evaluate two prompt strategies: *direct* zero-shot prompt and *icl* with two examples.

Analysis and Discussions

Do LLMs take Shortcuts?

- intuition은 성능과 골드 답변과 질문에 언급된 다른 엔티티 간의 평균 의미적 유사성 간의 상관 관계를 모델링하는 것

- entity similarities 은 Transformer model(DistilBERT) 로 엔티티 문자열 임베딩 값 구함

- 임베딩을 통해 타켓 답변과 질문에서 언급된 다른 모든 위키피디아 엔터티 간의 dot-product similarity를 계산하고 각 인스턴스에 대한 평균 유사도를 계산

- 특정 유사도를 가진 인스턴스에 대한 모델 성능의 correlation curve 그림
  → 제한된 평가 데이터를 기반으로 비교적 연속적인 곡선을 그릴 수 있도록 하는 역할

Analysis and Discussions

Do LLMs take Shortcuts?

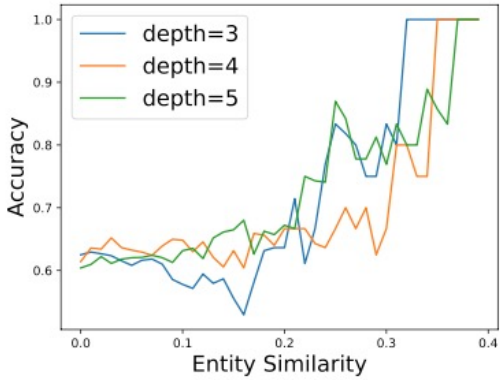


Figure 5: The correlation between GPT-4 performance on EUREQA hard set and entity similarities.

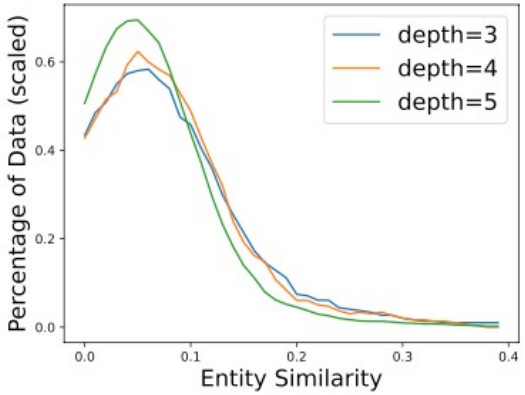


Figure 6: The distribution of entity similarity scores.

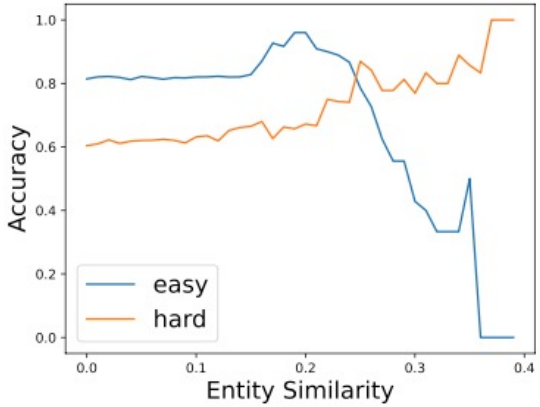


Figure 7: GPT-4 performances with different entity similarity scores between the easy and hard sets.

DFA-RAG: Conversational Semantic Router for Large Language Model with Definite Finite Automaton

Yiyou Sun [1 2] Junjie Hu [1] Wei Cheng [2] Haifeng Chen [2]

ICML 2024

Motivation

- 실제 시나리오에서 LLMs 적용하려면 특정 워크플로우나 정책 준수해야함
  e.g., Emotional Support Chatbot 에서는 다양한 스트레스 상황에 따라 맞춤형 응답,
  customer service bot 은 미리 정해진 답변 가이드라인 따라 응답

- LLM은 전문적인 최적화 없이 부적절하거나 잘못된 컨텐츠를 생성할 수 있음
  fine-tuning을 방법은 복잡한 구조 설계 및 데이터 문제로 인해 항상 적용할 수 X
  따라서 knowledge base 기반으로 응답에 도움이 되는 정보를 답변에 활용하는 RAG 방법론 탐구

→ RAG 의 생성 품질은 샘플 선택에 따라 민감하기 때문에 가장 관련성 높고 상황에 적절한 샘플을 리트리브 하는 선택 전략 설계하는 관점에서

→ 대화 히스토리를 고려해서 이전 컨텍스트와 관련 있는 부분 식별하여 해결

Problem Setup

- In our setting, 고객 서비스 혹은 emotional support 같은 어플리케이션 도메인에서 대화 샘플셋에 대한 접근이 가능함을 가정함

- Data Setup

  훈련 데이터셋은 N 개의 대화로 구성되어 있고, 각 대화는 발화들로 구성
  발화가 에이전트와 유저 사이에서 번갈아 발생한다고 가정함

- Goal

  인퍼런스 단계에서, LLM-based agent은 불완전한 대화의 컨텍스트를 기반으로 다음 발화 생성
  LLMs 응답이 human agent's response 와 거의 일치하도록 하는 object

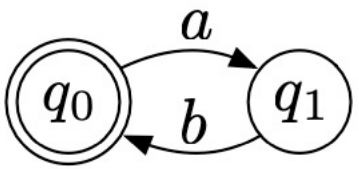
Preliminary of DFA

- Deterministic finite automaton (DFA) 는 특정 구문으로 알파벳 시퀀스를 정의하는 기능

- tuple $(Q, \sum, \delta, q_0, F)$ 로 구성
  - Q is a finite set of states
  - Σ is a finite input alphabet
  - $\delta : Q \times \Sigma \rightarrow Q$ is the transition function
  - $q_0 \in Q$ is the start state
  - $F \subseteq Q$ is the set of accept states

Conversation as Tag-sequence

- 간단한 syntactic strings 과 달리 대화의 본질은 주로 semantic level에 있음
  e.g., “My battery drains out fast”
  “How come my phone can be only used for 1 hour?”
  → 의미는 같지만 겹치는 단어 X



$$\Sigma = \{a, b\}, F = \{q_0\}, Q = \{q_0, q_1\}$$
$$\delta(q_0, a) = q_1, \delta(q_1, b) = q_0$$

Figure 2. A demo of DFA recognizing string “$(ab)^*$”.

- 이를 완화하기 위해, 대화의 각 발화가“tags” 집합으로 캡슐화 될 수 있다고 가정
  e.g., “How come my phone can be only used for 1 hour? → tag set {“#issues”, “#battery”}

Conversation Sets as DFA

- 개별 대화를 태그 시퀀스로 표현한다는 아이디어를 바탕으로, 해당 개념을 확장하여 DFA를 사용하여 전체 대화셋 모델링

- DFA 컴포넌트를 대화 모델링 프레임워크에 맞게 적용

  States(Q): DFA의 각 상태(state)는 대화 내의 특정 스테이지 혹은 컨텍스트 나타냄
  state 는 대화의 시작, 특정 문제에 대한 질의, 응답, 상호작용의 결론을 나타냄

  Alphabet (Σ): 발화를 태그로 나타낸 것

  Transition Function (δ): state and a tag 를 다음 state로 매핑
  context 흐름 관리 역할

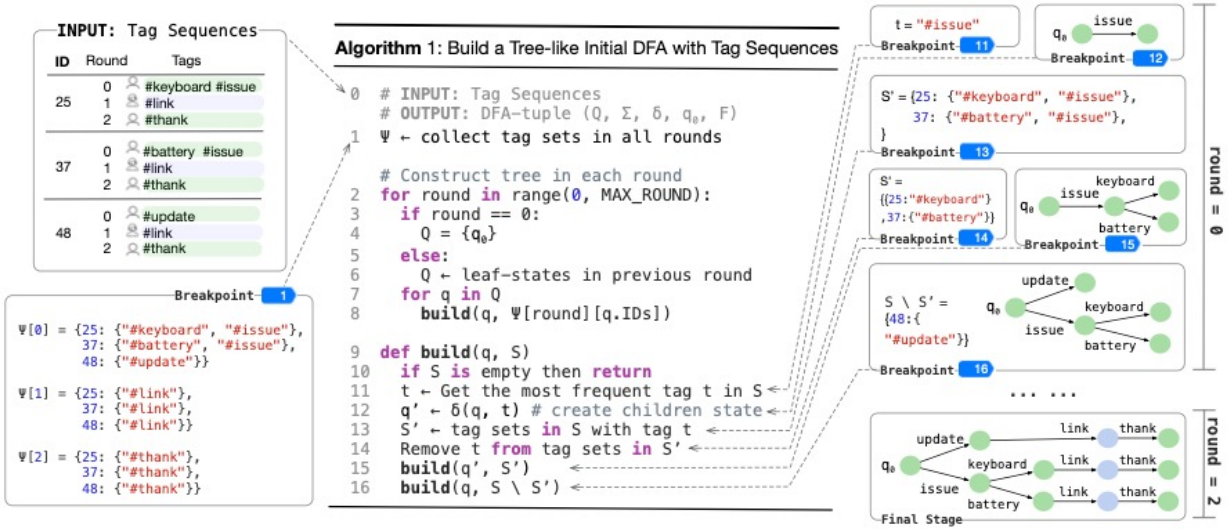
  Start State (q_0): 대화의 시작 나타냄

  Accept States (F): 대화의 종료 나타냄
  accept state 는 쿼리에 성공적으로 대응 / 만족스러운 결론에 도달 /
  대화가 자연스럽게 끝날 때 도달 가능

Learn DFA from Conversations

TREE CONSTRUCTION WITH TAG SEQUENCES

- Tags는 LLMs에게 문장 주고 태그 추출하도록 함
- 추출한 태그 시퀀스를 structured tree format 으로 재구성하여 DFA 에 사용



Conversation Generation by DFA-RAG

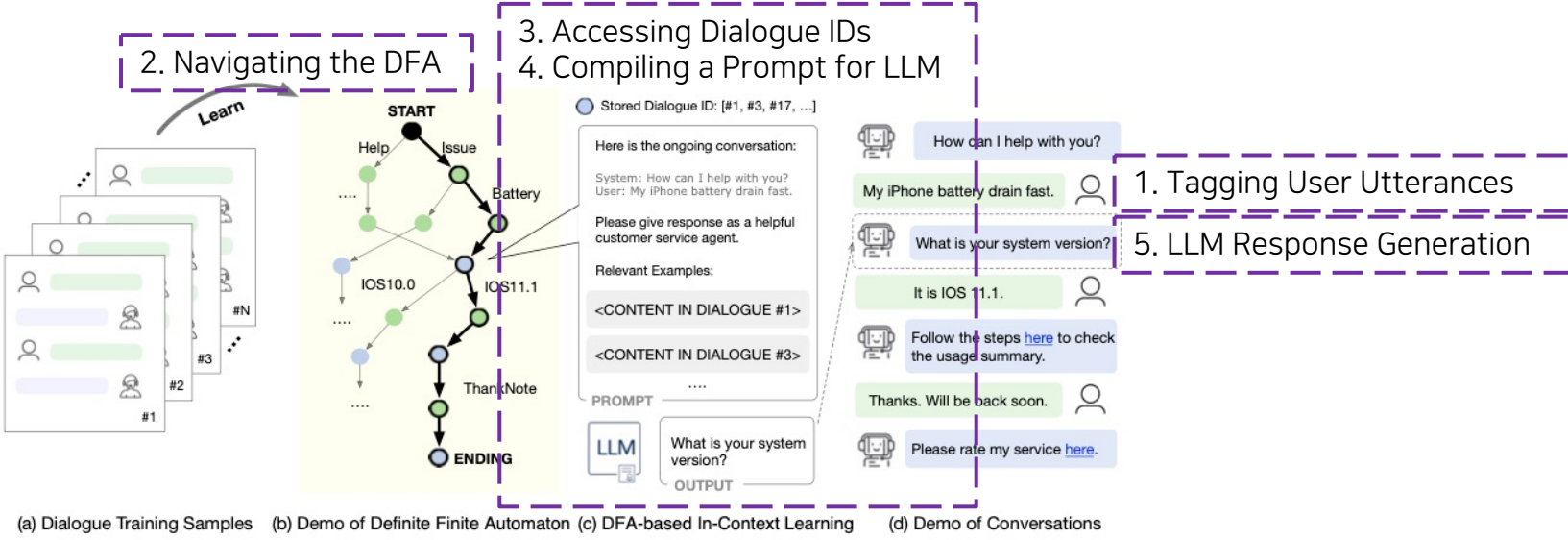


Figure 1. Illustration of the DFA-RAG Framework. (a) shows the training set with dialogues. (b) demonstrates the Definite Finite Automaton (DFA) which represents the workflow learned from the dialogues. Blue and green dots represent the states of the user and system respectively. The states are transited by keywords in conversations. (c) outlines the DFA-based In-Context Learning process, where the LLM is guided by the DFA to provide contextually relevant responses. (d) showcases sample conversations between a user and the LLM.

Experimental Results

Generation Quality Evaluation

- GPT-4로 GT랑 모델 outputs 2개 주고 뭐가 더 나은지 판단하도록 함 (“Win Rate” score)

Table 1. Results of dialogue generation quality across different base models and methods. This table reports the “Win Rate” over naive base models (GPT-4, GPT-3.5) regarding dialogue generation performance. For each method using in-context learning (`RandSamp`, `RAG`, `BM25`, `DFA-RAG`), we use 5 samples in the inference time. For `FT-LLM`, we perform fine-tuning using the API provided by OpenAI with standard hyperparameters. Note that the API for fine-tuning GPT-4 is not available.

Base LLM	Methods	Domains						Average
		AmazonHelp	DeltaSupport	AskPlayStation	AirbnbHelp	NikeSupport	CambridgeInfo	
GPT-4	RandSamp	69.1	84.1	57.9	78.3	45.3	67.0	66.9
	BM25	67.3	81.5	63.8	77.1	59.8	63.0	68.7
	RAG	74.4	87.0	**66.3**	72.2	57.3	66.5	70.6
	FT-LLM	-	-	-	-	-	-	-
	DFA-RAG (Ours)	**78.0**	**89.9**	65.9	**80.9**	**62.6**	**68.5**	**74.3**
GPT-3.5	RandSamp	70.2	83.6	61.3	69.5	58.9	57.9	66.9
	BM25	70.6	84.1	64.7	74.3	60.4	58.8	68.8
	RAG	73.8	82.9	72.4	76.6	63.3	60.6	71.6
	FT-LLM	69.7	64.6	71.7	66.1	56.8	56.1	64.2
	DFA-RAG (Ours)	**78.5**	**89.8**	**72.9**	**79.1**	**70.1**	**64.9**	**75.9**

Experimental Results

Dialogue Task Evaluation

- known dialogue states 와 비교해서도 제안한 방법론 성능이 좋다!

a) known dialogue states 가정

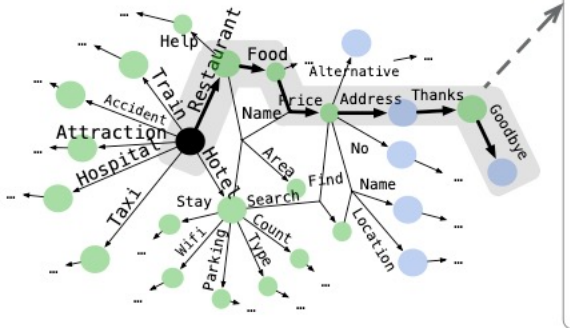
b) solely on dialogue context
→ 모델의 고유한 능력에 포커스

Table 3. Evaluation results on the task-oriented dialogues.

Ground Truth States in Training?	**Methods**	**Inform**	**Success**
Yes	HDSA	87.9	79.4
	MarCo	94.5	87.2
	HDNO	93.3	83.4
	GALAXY	92.8	83.5
	KRLS	93.1	83.7
No	AuGPT	76.6	60.5
	MTTOD	85.9	76.5
	RSTOD	83.5	75.0
	RewardNet	87.6	81.5
	TOATOD	90.0	79.8
No	**DFA-RAG (Ours)**	93.3	90.0

Experimental Results

Constructed DFA Demonstrations



(a) An Example Tag Pathway in Constructed DFA

ID	Raw Dialogues in Training Set
11	0 USER: I want to find an **expensive restaurant** and serves **european food**. Can i also have the **address**, phone number and its area? 1 SYSTEM: Eraina is a European restaurant in the centre area. Their address is **St. Michael's Church Trinity Street City Centre** and phone number 01223 355166. 2 USER: **thank** you goodbye 3 SYSTEM: Thank you and **good bye**
8238	0 USER: I want a **moderately priced restaurant** and serves **asian oriental food**. What is the **address** and phone number? 1 SYSTEM: there is an Asian Oriental place at **40428 King Street City Centre**. You can call them at 01223 518111. 2 USER: **Thank** you goodbye. 3 SYSTEM: Thank you for using our system! **Have a great day!**
7350	0 USER: I want a **cheap portuguese restaurant**. Give me their **address**, phone number, and area. 1 SYSTEM: Nandos is a cheap Portuguese restaurant in the south area. The phone number is 01223 327908 and the address is **Cambridge Leisure Park Clifton Way** 2 USER: **Thank** you, goodbye 3 SYSTEM: You're welcome, **goodbye**.

(b) Corresponding Dialogues in the Training Set

Figure 5. DFA Results for MultiWOZ. (a) This segment of the figure illustrates a portion of the constructed DFA. The black circle indicates the starting point of the automaton. Each green circle represents a "user" state, while each blue circle denotes a "system" state. The states are interconnected by arrows, each labeled with a tag. Note that some lines are interconnected (ex. lines correspond to"name" and "area"), it means that the relevant nodes are connected in both ways. (b) A specific path within the DFA is highlighted to demonstrate its correspondence with actual dialogues traversed. In these dialogues, elements associated with the tags are emphasized in bold.

Conclusion

- 단순 QA 보다는 여러 리즈닝 단계 필요하거나 보다 복잡한 dialogue 세팅에서 연구 포커스 하는 듯
- 하지만 트렌드는 여전히 RAG 관점에서 LLMs 의 리즈닝 능력을 어떻게 끌어올릴까? 인데 여기서 약간의 복잡성을 추가해서 contribution 만드는 것 같음
- 주어진 정보를 추가로 압축(요약, 자르기) 하기 보다는 모델에게 강조해야 할 부분을 더 강조하는 것이 좋아보임 (안전성 관점)

Thank you