



NLP&AI 동계세미나 (02/22, Thu)

# LLMs 검증

김진성



Natural Language  
Processing  
& Artificial Intelligence

고려대학교  
KOREA UNIVERSITY

# 논문 #1

## **Prompting is not a substitute for probability measurements in large language models**

**Jennifer Hu**

Kempner Institute  
Harvard University

[jenniferhu@fas.harvard.edu](mailto:jenniferhu@fas.harvard.edu)

**Roger Levy**

Department of Brain and Cognitive Sciences  
Massachusetts Institute of Technology

[rplevy@mit.edu](mailto:rplevy@mit.edu)

EMNLP 2023

# Intro

## \* Prompting을 통한 지식 측정 (1)

- Prompting: 우리가 맨날 하는 '그' 자연어 프롬프팅

→ LLM이 가진 어떤 지식들을 이끌어내는 역할을 한다고들 다들 말해짐.

- LLM의 언어적 지식 (linguistic knowledge)를 평가하는 데에 있어서도 최근 지배적인 방법론임.



You

주어진 프롬프트 구절을 문장으로 완성시켜라.

- given: 나는 오늘 햄버거를 최대 두 개



ChatGPT

given: 나는 오늘 햄버거를 최대 두 개 먹을 거야.



You

- given: 나는 어제 햄버거를 두 개나

< 2 / 2 >



ChatGPT

• given: 나는 어제 햄버거를 두 개나 먹었어.



# Intro

## \* Prompting을 통한 지식 측정 (2)

- ~ by processing linguistic input, thereby implicitly testing a new type of emergent ability.

: 'Metalinguistic judgement'

- 반면, 기존 대부분의 언어적 지식 측정 방법들은?

: directly read out models' probability distributions over strings

: 'Direct measurements'

→ 직접적으로 모델의 확률 분포를 읽어냄. →  $P(\textit{token}|\textit{context})$

# Intuition

## \* Prompting vs. Probability Measurements (1)

- Int1) 진짜 프롬프팅만으로 직접 representation 가지고 평가하는 것과 같은 효과 내기 가능?  
= LLM의 언어적 지식 / 능력 측정 완전 가능?

Int2) 그럼 언제 Direct 쓰고, 언제 Prompting 쓰는게 좋지?

- 근데, 문제가 있는게 뭐냐면, 애초에 평가하는 세팅이 다르다는 것.

e.g.) P(is) vs P(are)

- (1)
  - a. The keys to the cabinet → Direct
  - b. Here is a sentence: The keys to the cabinet... What word is most likely to come next?

→ Prompting: target 하는 sentence 뿐만 아니라, 지시 및 질문 등 추가로 들어감!!

# Intuition

## \* Prompting vs. Probability Measurements (2)

- (1)
  - a. The keys to the cabinet
  - b. Here is a sentence: The keys to the cabinet... What word is most likely to come next?

- 애초에 같은 모델이라고 하면, a 와 b 상황에서 둘 다 'is' 혹은 'are'에 동일한 확률 분포를 보여야 함.

- 근데,  
주어진 metalinguistic prompt ((1) b.) 에 대한 모델 (자연어로 생성된) 응답과  
'underlying internal representations'가 match 한다는 보장이 없음.

→ 그럼 이 모델의 응답을 어떻게 해석해야 하는가 ... 어떻게 correspond 하게 할까..?

# 태스크 셋업

## \* Research Questions

- RQ1) How well do models perform under direct and metalinguistic evaluation methods?

→ 각 평가 세팅에서 얼마나 잘 함?

- RQ2) How consistent are the metalinguistic methods with the direct method?

→ 두 평가 세팅 간에서 (동일한) 모델이 얼마나 consistent 한 결과를 보이는가?  
(alignment 평가)

# 태스크 셋업

## \* 실험 할 것들

- 1) 이어질 끝 단어, 2) 두 단어 중 뭐가 더 맥락에 잘 이어지는지, 3), 4) 두 문장 중 뭐가 나은지
- 각 태스크에 대해 Direct vs. zero-shot prompting 방법론 3개 비교함. (Flan-T5, 일부 GPT류)

Experiment	Targeted ability	Task	Dataset(s)
1 (Section 4.1)	Word prediction	Predict final word in a sentence	<a href="#">Pereira et al. (2018)</a> ; news articles from March 2023
2 (Section 4.2)	Semantic plausibility	Determine which word (of two options) is most likely, given preceding context	<a href="#">Vassallo et al. (2018)</a>
3a (Section 4.3)	Syntax	Determine which sentence (of two options) is “better”, in isolation	SyntaxGym ( <a href="#">Hu et al., 2020</a> ); BLiMP ( <a href="#">Warstadt et al., 2020</a> )
3b (Section 4.4)	Syntax	Determine which sentence (of two options) is “better”, given both options	SyntaxGym ( <a href="#">Hu et al., 2020</a> ); BLiMP ( <a href="#">Warstadt et al., 2020</a> )

Table 1: Overview of experiments in our study.



# Experiments

## \* 실험 1: 다음 단어 맞추기

- 저 프롬프트들은 자신들의 기준으로 짤거 (기준: Direct 와의 유사도에 따라)
- 성능 평가: (모델이 예측해야 하는) 마지막 단어에 대한 Log probability  
→ GT 단어에 대한

Type of prompt	Example
Direct	A butterfly is a flying insect with four large <b>wings</b>
MetaQuestionSimple	What word is most likely to come next in the following sentence? A butterfly is a flying insect with four large <b>wings</b>
MetaInstruct	You are a helpful writing assistant. Tell me what word is most likely to come next in the following sentence: A butterfly is a flying insect with four large <b>wings</b>
MetaQuestionComplex	Here is the beginning of an English sentence: A butterfly is a flying insect with four large... What is the best next word? Answer: <b>wings</b>

Table 2: Example prompts for Experiment 1. Region where we measure probability is marked in **boldface**. Ground-truth sentence continuations are shown in **blue**.

# Experiments

## \* 실험 2: Word Comparison

- 성능 평가: '두 단어 중 모델이 더 높은 확률 값을 부여한 단어 == 정답 단어'이면 맞은거 (Accuracy)

Type of prompt	Example
Direct	The archer released the { <b>arrow</b> , <b>interview</b> }
MetaQuestionSimple	What word is most likely to come next in the following sentence (arrow, or interview)? The archer released the { <b>arrow</b> , <b>interview</b> }
MetaInstruct	You are a helpful writing assistant. Tell me what word is most likely to come next in the following sentence (arrow, or interview?): The archer released the { <b>arrow</b> , <b>interview</b> }
MetaQuestionComplex	Here is the beginning of an English sentence: The archer released the... What word is more likely to come next: arrow, or interview? Answer: { <b>arrow</b> , <b>interview</b> }

Table 3: Example prompts for Experiment 2. Region where we measure probability is marked in **boldface**. Semantically plausible continuations are shown in **blue**; implausible in **red**.

# Experiments

## \* 실험 3: 어떤 문장 Better ? Sentence judgement

- 성능 평가: '두 문장 중 모델이 더 높은 확률 값을 부여한 문장 == 정답 문장'이면 맞은거 (Accuracy)

Type of prompt	Example
Direct	{ <b>Every child has studied</b> , <b>Every child have studied</b> }
MetaQuestionSimple	Is the following sentence a good sentence of English? Every child has studied. Respond with either Yes or No as your answer. { <b>Yes</b> , <b>No</b> }
MetaInstruct	You are a helpful writing assistant. Tell me if the following sentence is a good sentence of English. Every child has studied. Respond with either Yes or No as your answer. { <b>Yes</b> , <b>No</b> }
MetaQuestionComplex	Here is a sentence: Every child has studied. Is the sentence a good sentence of English? Respond with either Yes or No as your answer. Answer: { <b>Yes</b> , <b>No</b> }

(a)

# Experiments

## \* 실험 4: 어떤 문장 Better ? Sentence comparison

- 성능 평가: '두 문장 중 모델이 더 높은 확률 값을 부여한 문장 == 정답 문장'이면 맞은거 (Accuracy)

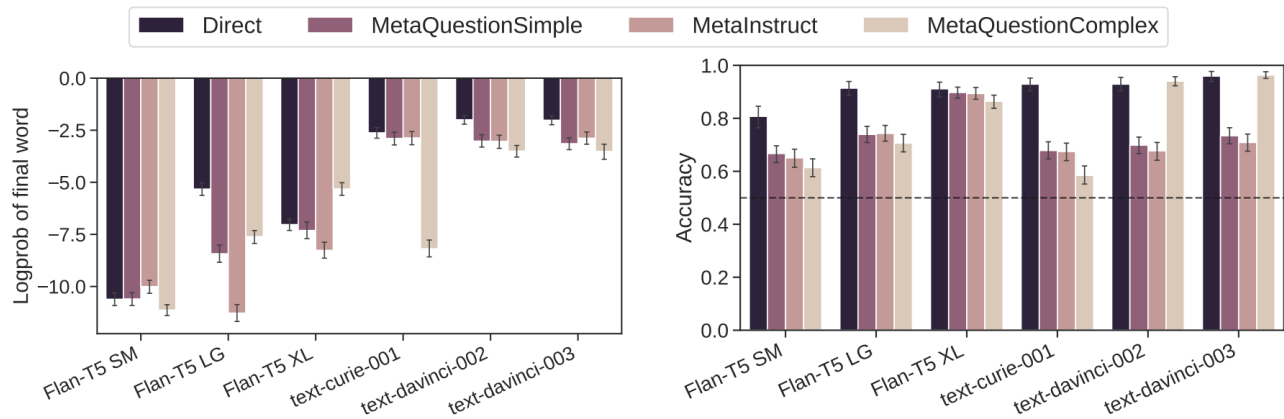
Type of prompt	Example
Direct	{ <b>Every child has studied</b> , <b>Every child have studied</b> }
MetaQuestionSimple	Which sentence is a better English sentence? 1) Every child has studied. 2) Every child have studied. Respond with either 1 or 2 as your answer. { <b>1</b> , <b>2</b> }
MetaInstruct	You are a helpful writing assistant. Tell me which sentence is a better English sentence. 1) Every child has studied. 2) Every child have studied. Respond with either 1 or 2 as your answer. { <b>1</b> , <b>2</b> }
MetaQuestionComplex	Here are two English sentences: 1) Every child have studied. 2) Every child has studied. Which sentence is a better English sentence? Respond with either 1 or 2 as your answer. Answer: { <b>1</b> , <b>2</b> }

(b)

# Experiments

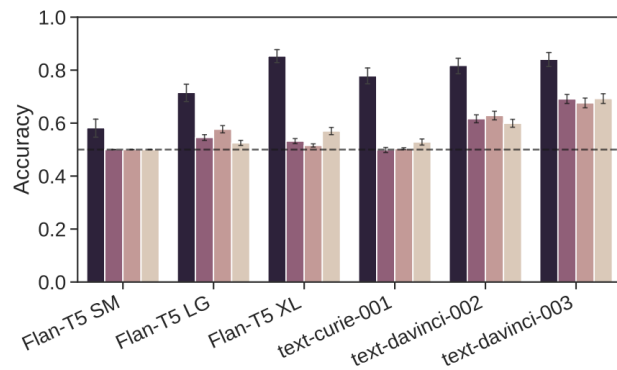
## \* Results

- 웬만하면,  
Direct로 측정할 때,  
가장 좋은 성능을  
보인다.

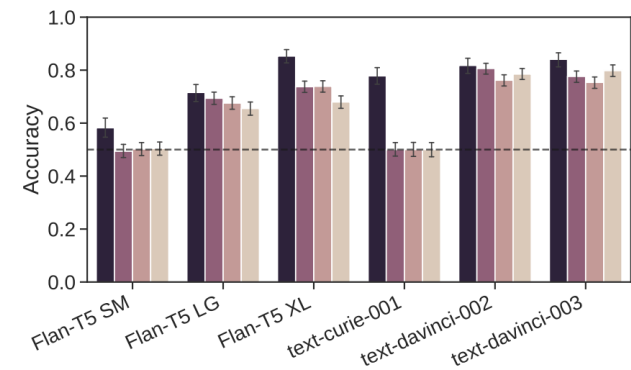


(a) Experiment 1: Word prediction

(b) Experiment 2: Word comparison (Semantic plausibility)



(c) Experiment 3a: Sentence judgment (Syntax)



(d) Experiment 3b: Sentence comparison (Syntax)

# Findings

## \* Findings (1)

1. 직접 internal logits 접근해서 모델 지식 알아내는거랑, 프롬프팅으로 하는건 확실히 다른거 하는거다.
  2. (앞에 결과에도 있듯) 프롬프팅으로 할 때 보다 직접 측정하는게 태스크 성능도 더 좋드라.
  3. 프롬프팅 할 때 Minimal pairs 주면 좋다.  
= 즉, 프롬프팅으로 푸는 태스크 설계할 때, 예제 하나 주고 {Yes, No} 때리는거 보다, 두 개 주고 뭐가 좋은지 골라봐 {1, 2} 하는게 성능 오른다.
  4. 프롬프트가 바닐라랑 멀어지면 멀어질수록, 비슷한 확률 값 뺀을 Correlation이 떨어진다.
1. The metalinguistic judgments elicited from LLMs through prompting are *not the same* as quantities directly derived from model representations. (*Figures 2 and 3*)
  2. Direct probability measurements generally yield better or similar task performance, compared to metalinguistic prompting. (*Figure 2*)
  3. Minimal pairs help reveal models' generalization capacities, compared to isolated judgments. (*Figure 2c vs. Figure 2d*)
  4. In general, the less similar the task/prompt is to a direct probability measurement, the worse the alignment between metalinguistic and direct measurements. (*Figure 3*)

# Findings

## \* Findings (2)

4. 프롬프트가 바닐라랑 멀어지면 멀어질수록, 비슷한 확률 값 받을 Correlation이 떨어진다.

- 실험할 때 Acc. 만 구한거 아니고, 각 방법론 마다의 후보에 대한 확률 값의 차이 구해서, Direct 기준으로 피어슨 상관계수 때린거  
(e.g., 태스크 3이면, "문장 + Yes" 조합에 대한 모델의 log probability 값 - "문장 + No" 에 대한 확률값)

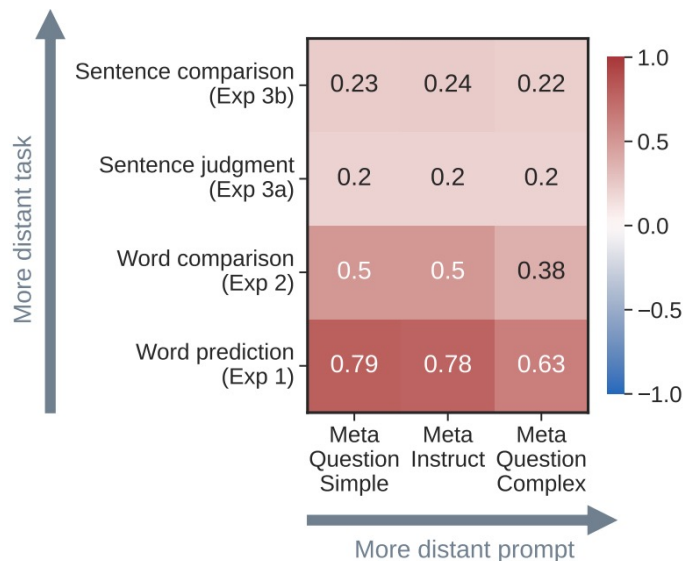


Figure 3: **Internal consistency: Correlation between metalinguistic and direct responses gets weaker as prompts become less direct.** Pearson  $r$  correlation between response magnitudes (averaged over models and datasets) measured by direct prompts versus each metalinguistic prompt. See Appendix C for more details.

# Findings

## \* 담론 확장

- 하지만, 이러한 부정적인 결과가 LLM의 linguistic generalization 의 결여의 증거는 아님.
  - LLMs 의 performance-competence 구분 해야함.
  - . 전자: the information encoded in a model's isolated-sentence string probability distribution
  - . 후자: the model's behavioral responses to prompts.
- Closed LLMs 치사한 놈들, internal probabilities 쓸 수 있게 해줘라.



# 후기

## \* Strengths and Weaknesses

- 일단 주제가 (제목부터) 흥미돋.
  - 왜냐하면, 공감했던게 요새 하도 프롬프팅 방법론의 난립이 심해서...
  - 특히, thinking-styles 도 무슨 학회마다 한두개도 아니고 몇 십개는 넘는 듯...
- 그리고 요새는 LLM 나오고 나서, 방법론 제안 뿐만 아니라 가설 검증에도 힘 많이 쓰는게 합격률 좋은 듯..
- 오우.. 글 잘 씀 (특히, 헤드라인만 읽어도 논문 읽히게 해놓는거...)
- 딱 주장 뒷받침에 필요한 실험들을 세팅 잘 한듯 (실험표가 많지도 않은데...)
- OpenReview 점수  
R1 4/4, R2 4/4, R3 4/4
  - 5: Sound and Exciting: Accept to Main Conference (체어 decision)
- 리뷰어들한테 까인거?
  - . 왜 영어 모델만 해줬냐, 어떤 상황에 프롬프팅(direct) 쓰는게 더 좋은지 말해준다 해놓고 왜 안 해줬냐, 프롬프팅 방법론 왜 3개만 했냐 더 탐색해보지 등...

## 논문 #2

# Evaluating Large Language Models on Controlled Generation Tasks

**Jiao Sun<sup>1\*</sup> Yufei Tian<sup>2\*</sup> Wangchunshu Zhou<sup>3\*</sup> Nan Xu<sup>1\*</sup>**  
**Qian Hu<sup>4</sup> Rahul Gupta<sup>4</sup> John Wieting<sup>5</sup> Nanyun Peng<sup>2</sup> Xuezhe Ma<sup>1</sup>**  
<sup>1</sup>University of Southern California <sup>2</sup>University of California, Los Angeles  
<sup>3</sup>ETH Zurich <sup>4</sup>Amazon <sup>5</sup>Google DeepMind

EMNLP 2023

# Intuition

## \* Controllability of LLMs

- LLMs의 성능에 관한 연구는 많은데, controllability 관점에서의 연구 미비.

→ 10 개의 Generation tasks 벤치마크에서 controllability 관점의 분석 제공

→ SoTA fine-tuned smaller models 와 LLMs 의 generation 비교 분석

*“Are large language models better than finetuned smaller models at controllability on generation tasks?”*

Task	Control	Benchmark	Evaluation
● constrained content generation	sentiment, topic, keyword	Amazon Review CommonGen M2D2	off-the-shelf model, ppl
● story generation	prefix	ROC writing prompts	repetition, diversity, coherence
● rationale generation	correct answer	CoS-E ECQA	increased accuracy
● numerical planning <sup>NEW</sup>	prefix & number of words & end word	NPB	MSE, success rate
● paraphrase generation	semantic & syntax	ParaNMT QQPoS	lexical overlapping, syntax match


good  poor

Figure 1: We test large language models on five controlled generation tasks with various control factors using automatic evaluation methods. We show a spectrum of abilities of large language models on such tasks and conclude that large language models struggle at fine-grained hard constraints such as numerical planning.

# Task Setups (1)

## \* Content-Controlled Generation

- 3 가지 contents 통제: 1) Topic, 2) Sentiment, 3) Keyword

- 평가

1), 2): InstructGPT + ICL 활용해서 input topic/sentiment 에 속하는지 분류하여 사용 (success rate)

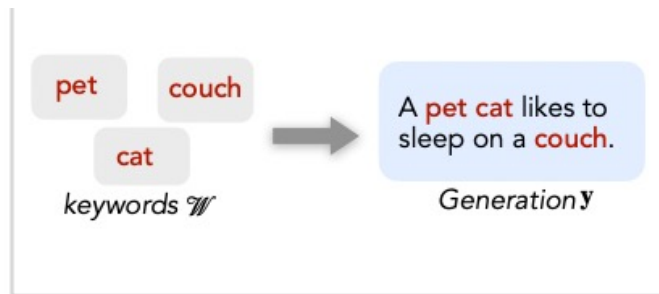
3): 제시한 keywords 가 generated text 에 얼마나 들어있는지 (coverage rate)

예시)

→ "Write a sentence about {topic name}."

"Write a sentence using the following keywords: {keywords}."

예시)



Lexically Constrained Generation

# Task Setups (2)

## \* Story Generation

- story 의 도입부 제공 시, 이전 부분과 1) coherent 하고, 2) 의미 없는 반복 적은 텍스트 생성
- 5 문장으로 이루어진 short story 에서 첫번째 문장을 prefix 로 제공 후, 4 문장을 이어서 써라.  
(*"Please continue writing this story within 4 very short sentences: <prefix>"*)  
혹은 처음 32 tokens 를 주고, 이후 256 tokens 를 써라.  
(*"Please continue writing this story within 256 words: <prefix>"*)
- 채점:
  - . Rep-n 점수: 중복되는 n-grams 비율에 따라 sequence-level 의 repetition 을 측정
  - . Diversity: "to assess the overall model repetition by considering rep-n at different n-gram levels"
  - . Coherence: 코사인 유사도 (prefix 와 generated text 간의 - SimCSE 로 get embeddings)

# Task Setups (3)

## \* Rationale Generation

- free-form rationales 가 복잡한 리즈닝 태스크를 위한 LLMs' 능력을 개선한다는 선행 연구 (Wei et al. (2022))

→ conditioned gen. 으로 만든 LLMs' generated rationales 가 진짜 효용성 있을까?  
: Multiple Choices 상황에서 (I+R → O) 의 acc. 와 (I → O) 의 acc. 를 비교.

- LLMs 의 generated rationales 를 FlanT5-XXL 에 제공하여 도움되는지 평가.  
(background knowledge 설명해주는 그런 역할.)

- 가령,

*"Question: {question}"*

*Options: {concatenated options}"*

*Explain the rationale behind choosing the correct option {correct answer}"*

# cf) Rationale Generation task

\* ECQA (2021 ACL)

- 일종의 설명 생성  
. CSQA 등의 데이터에 설명 annotation 하여 구축

---

**Question:**

What is something that people do early in the day?

**Answer Choices:**

believe in god

make tools

skydive

smoke pot

eat eggs



**Our Explanation:**

**Positives Properties**

- 1) People generally eat breakfast early morning.
- 2) People most often eat eggs as breakfast.

**Negative Properties**

- 1) Believing in god is not restricted to a specific part of a day.
- 2) People generally do not make tools early in the day.
- 3) Skydive is an irrelevant answer.
- 4) People usually do not smoke pot early in the day.

**Free-Flow Explanation (FF)**

People generally eat breakfast early morning which most often consists eggs. People generally do not make tools or smoke pot early in the day. Skydive is an irrelevant answer.

---

Table 11: Example of CommonsenseQA with our annotated explanation

# Task Setups (4)

## \* Numerical Planning

- 새로 태스크 디자인  
→ *“Can LLMs count from two to ten?”*
- 추가적으로, Granularity 에 대한 condition 도 추가.

*##Prefix: This is a story about a young girl’s*

*##Last word: town*

*##N: 5*

*##Output: This is a story about a young girl’s redemption in a small town.*

*##Explanation: We generated “redemption in a small town”. It contains exactly 5 words and ends with the last word ‘town’.*

→ 채점: 정답 vs. 생성된 counts 간 MSE (+) last word, counts 에 대한 success rate 측정

갈수록 복잡

Granularity	Task Illustration
Word/Syllable	Generate a sentence using exactly 5 words/syllables.
	Complete sentence “This is a story” using exactly 5 words/syllables.
	Complete sentence “This is a story” using exactly 5 words/syllables, including the last word as “town”.
Sentence	Generate a paragraph with 5 sentences, ...
Paragraph	Generate an article with 5 paragraphs, ...

Table 1: Task illustration for the Numerical Planning Benchmark. We test LLMs’ numerical planning ability under various constraints (word counting and end word) and granularities (word, syllable, sentence, and paragraph). Due to space limitations, we only show the full constraints under the word granularity here.



# Task Setups (5)

## \* Controlled Paraphrase Generation

- Syntactically-controlled paraphrase 생성

- Settings

. Direct - w/o any constraints

→ *"Paraphrase {source sentence}"*

. Controlled

→ *"Paraphrase {source sentence} so that it uses the syntactic structure from {exemplar}; please only have the paraphrase in the response."*

. Control w/ Syntax Explanation (Stanford parser 로 먼저 문장구조 추출한 정보 같이 주고 생성)

→ *"Paraphrase {source sentence} so that the sentence has a syntactic structure of {pruned syntax}. {generated explanation for the syntax}. Please only have the generated paraphrase, not its parse, in the response."*

# 결과 (1)

## \* Content-Constrained Generation

- ICL: 5-shot
- ChatGPT >> Vicuna >= Alpaca > LLaMA
- ICL ChatGPT 는 거의 무적

Model	Topic	Sentiment	Keyword
Diffusion-LM	68.9	83.7	93.2
GPT-2 (1.5B, fine-tuned)	63.4	76.5	88.9
T5 (3B, fine-tuned)	67.3	83.9	94.8
LLaMA-7B <sub>ZS</sub>	45.3	58.4	83.5
LLaMA-7B <sub>ICL</sub>	63.5	85.1	93.0
Alpaca-7B <sub>ZS</sub>	58.9	78.4	91.2
Alpaca-7B <sub>ICL</sub>	65.2	86.9	94.8
Vicuna-7B <sub>ZS</sub>	61.0	80.5	91.6
Vicuna-7B <sub>ICL</sub>	65.8	87.4	94.3
Falcon-7B <sub>ZS</sub>	61.9	81.0	92.1
Falcon-7B <sub>ICL</sub>	66.0	87.7	94.2
ChatGPT <sub>ZS</sub>	66.4	84.5	97.3
ChatGPT <sub>ICL</sub>	<b>88.4</b>	<b>90.3</b>	<b>98.1</b>

Table 3: Results on content-constrained text generation.

# 결과 (2)

## \* Story Generation

- coherence:

모든 LLMs 가 2XL 대비 훨씬 좋음

- Rep-n:

ChatGPT > Vicuna >> Falcon

Decoding 방법

Search 방법

LM	Method	rep-2↓	rep-3↓	rep-4↓	diversity↑	coherence↑
<b>ROC</b>						
GPT-2-XL	Human	1.74	0.32	0.04	0.97	0.48
	Nucleus	1.80	0.35	0.12	0.97	0.33
	Typical	2.06	0.4	0.16	0.97	0.33
	$\eta$ -sampling	<b>0</b>	<b>0</b>	<b>0</b>	<b>1.00</b>	0.34
	SimCTG	3.10	0.46	0.23	0.96	0.32
	Look-back	7.24	0.92	0.14	0.92	0.47
LLM	Vicuna	2.36	0.45	0.15	0.97	0.60
	Falcon	2.52	1.87	1.86	0.94	<b>0.69</b>
	ChatGPT	1.18	0.10	0.02	0.98	0.52
<b>Writing Prompts</b>						
GPT-2-XL	Human	15.61	3.78	1.24	0.80	0.31
	Nucleus	5.40	2.41	1.72	0.91	0.34
	Typical	3.60	1.51	1.10	0.94	0.30
	$\eta$ -sampling	6.17	2.88	2.16	0.89	0.35
	SimCTG	<b>2.84</b>	<b>0.36</b>	<b>0.19</b>	<b>0.97</b>	0.31
	Look-back	7.94	1.25	0.33	0.91	0.52
LLM	Vicuna	8.27	2.59	1.14	0.88	0.49
	Falcon	11.20	7.79	6.94	0.76	<b>0.53</b>
	ChatGPT	5.99	1.15	0.35	0.92	0.52

Table 4: Performance of different decoding strategies and LLMs for open-ended story generation. Vicuna stands for Vicuna-7B, Falcon for Falcon-7B-Instruct.

# 결과 (3)

## \* Rationale Generation

- Settings
- . Leakage

: correct answer explicitly in the rationales ?  
→ questions 없이도 final answer (option)을  
추출할 수 있는지. (앞에서 본)

- . Non-leakage:

*"Question: {question}*

*Options: {concatenated options}*

*Explain the rationale behind choosing the correct answer. Do not mention the correct answer explicitly."*

- ChatGPT 는 Reference rationales 제공이랑 비교해도  
별 성능 차이 없게 좋음.

I→O	0.87	
I+R <sub>CoS-E</sub> →O	0.92	
I+R <sub>ECQA</sub> →O	<b>0.99</b>	
Model	Leakage	Non-Leakage
I+R <sub>Alpaca-7B</sub> →O	0.91	0.86
I+R <sub>LLaMA-7B</sub> →O	0.87	0.79
I+R <sub>Vicuna-7B</sub> →O	0.95	0.74
I+R <sub>Falcon-7B</sub> →O	0.83	0.65
I+R <sub>ChatGPT</sub> →O	<b>0.98</b>	<b>0.93</b>

Table 5: Rationales generated by ChatGPT are on par with best-crowdsourced rationales ECQA with FlanT5-XXL (Chung et al., 2022b) as the backbone model. Ruling out leakage results in at least 5% accuracy drop.

## 결과 (4)

### \* Numerical Planning

- word count planning task

- . ChatGPT 만 그나마 fine-tuned GPT-2의 2/3 정도 성능  
나머지는 형편 없음...
- . few-shot 줘도 별 향상 X, 오히려 깎아먹기도.

Model	SR - count	SR - last word	SR - both	MSE - count
GPT-2 (fine-tuned)	0.64	0.86	0.60	1.62
Alpaca-7b <sub>zs</sub>	0.17	0.31	0.09	9.19
Alpaca-7b <sub>ICL</sub>	0.14	0.34	0.07	9.76
Vicuna <sub>zs</sub>	0.08	0.09	0.03	27.68
Vicuna <sub>ICL</sub>	0.13	0.30	0.04	13.43
Falcon <sub>zs</sub>	0.13	0.42	0.08	11.60
Falcon-7b <sub>ICL</sub>	0.11	0.34	0.03	13.72
ChatGPT	<b>0.41</b>	0.74	<b>0.36</b>	<b>3.64</b>
ChatGPT <sub>ICL</sub>	0.37	<b>0.78</b>	0.34	4.95

Table 2: Success rates for the word count planning task. Surprisingly, few-shot in-context learning (ICL) underperforms zero-shot (zs) on numerical planning.

# 결과 (4)

## \* Numerical Planning

- Count N : 2~10 까지 실험
- 숫자 N 이 커질수록 성능의 저하가 더 심하게 일어남.

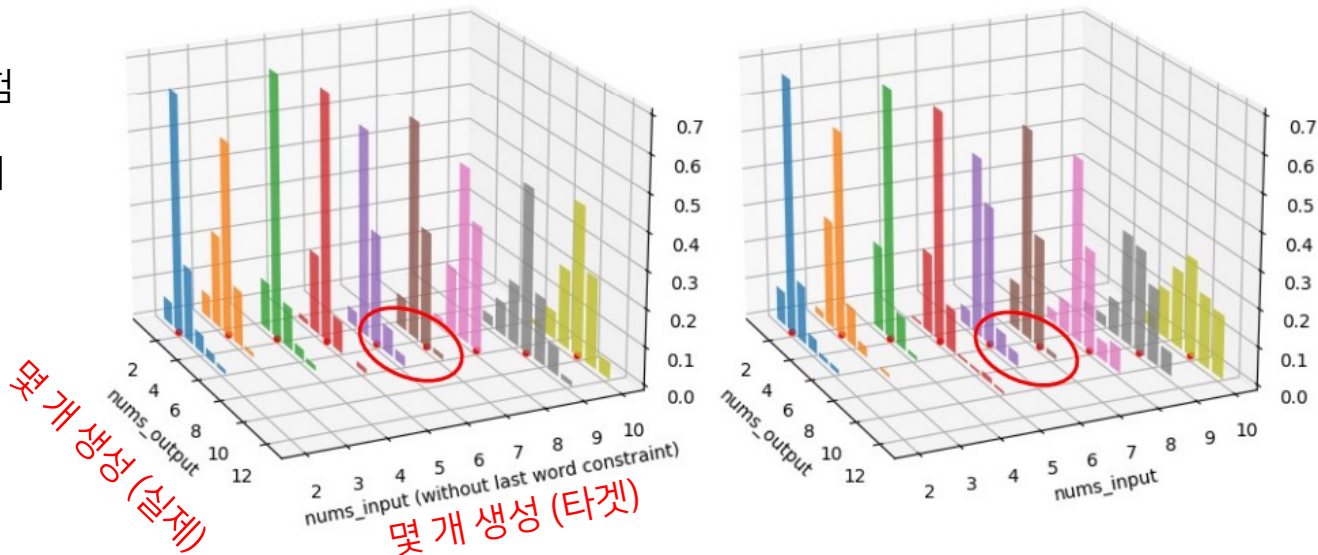


Figure 2: Histogram visualization in the distribution (frequency, z-axis) of input numbers (x-axis) and output numbers (y-axis) for word count planning. Left: querying ChatGPT to generate a continuation of a given prefix with  $N$  words. Right: querying ChatGPT to generate a continuation with  $N$  words of a given prefix that ends with a given word. Small red dots  $\bullet$  mark those bars where output numbers equal input numbers. These bars represent the fine-grained success rates. For either case, there is a significant drop when the input number reaches six.

## 결과 (5)

### \* Controlled Paraphrase Generation

Tree edit distance  
: syntactic conformation metrics



		BLEU↑	METEOR↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	TED-R↓ (H=2)	TED-E↓ (H=2)
ParaNMT -Small	Direct	10.8	26.2	44.2	18.6	44.9	1.4	1.5
	Ctrl	14.3	30.7	51.4	25.8	50.7	1.3	1.2
	Syntax exp.	13.6	27.3	46.4	20.2	47.0	1.4	1.4
	 <b>AESOP</b>	<b>22.9</b>	<b>32.7</b>	<b>54.4</b>	<b>29.8</b>	<b>56.4</b>	<b>0.9</b>	<b>0.5</b>
QQPPos	Direct	6.7	25.2	39.8	15.6	41.5	1.8	1.8
	Ctrl	10.5	25.6	43.0	19.8	45.2	1.4	1.4
	Syntax exp.	9.0	26.5	42.8	17.8	14.2	1.8	1.8
	 <b>AESOP</b>	<b>47.3</b>	<b>49.7</b>	<b>73.3</b>	<b>54.1</b>	<b>75.6</b>	<b>0.4</b>	<b>0.3</b>

Table 6: Performance comparison with ground-truth syntactic control for AESOP (Sun et al., 2021) and fine-shot ChatGPT. With coarse syntactic control from a shallow height of pruning, AESOP, the state of the finetuned small

- AESOP: (source, GT reference) vs. ChatGPT w/ 5 shots  
→ ChatGPT 가 모든 지표에서 GT 절대 못 따라감.

(+) LLaMa, Alpaca 는 input 만 계속 반복해서 말하고, 성능 reporting 하지 못 할 정도로 못 했다고 함.

# Discussion

## \* Why and How

- Why: LLMs 가 numerical planning 에 왜 실패 하는지 추론.

→ 그니까, 진짜 LLMs 가 controllability 떨어지는게 아니고, 다른 문제 아니냐?

### 1) Tokenization:

. 가능한 반박:

“기본적으로 LM은 subword-level generation 하니까, counting ‘complete’ words 가 어렵지 않나?”

→ 저자) 그렇다는 명확한 증거 제시하는 연구 없었음.

(→ 몇몇 증거들 드는데, 잘 설득은 안 됨.)

### 2) Decoding methods:

. Temp 0.3으로 sampling 기반으로 한 실험 reporting 하긴 했는데,

다른 decoding 전략도 해봤음 (greedy, beam search with 8, sampling with  $T=\{0.3, 0.7, 1.0\}$ )

→ 저자) 다 해봤더니 지금 세팅이 제일 좋아서 이렇게 한 것.

### 3) In-context learning

. “이미 표에 보여줬지만, exemplars 늘린다고 성능 안 올라갔다.” → 한번 더 언급



# Discussion

## \* Why and How

- How to improve: 어떻게 잠재적으로 controllability 를 높일 수 있을지 탐색.  
= 사실 상 상세한 future work

1) chain/tree/graph-of-thought reasoning 을 활용해본다.  
. Prompting 방법 개선

2) non-autoregressive generation 과 LLMs 를 결합해본다.  
. Autoregressive models 가 다음 토큰 생성 시 "looking back" 하지 않는게 문제다.  
. 이걸 해결해야 근본적으로 planning task 를 해결 하는게 아닌가..

→ 말은 이렇게 거창하게 하지만, multi-step planning, iterative revisions with LLMs 해라는 얘기.

*"Therefore, we encourage researchers to also investigate ~~"* 라고 하고 끝.

→ 그렇게 엄청 유의미하진 않아 보이는 듯..

# 후기

- OpenReview: 4/3, 4/4, 4/3

→ meta review 3 (accept to findings) → Chair: Main accept

## \* Strengths

- 각 태스크마다 배경 설명 및 정당성 부여 (cite 활용)를 엄청 함.
- automatic metrics 만 사용 (human eval. 도 없음)
  - . 채점할 때, 모델 분류 결과를 사람처럼 사용함. (InstructGPT 등)
  - "imperfect yet convenient and reproducible."
- 나올 수 있는 반박 defense 코너 따로 마련
  - . 비록 이게 말이 되든 안 되든, cite 여러 개 달면서 설득 하려고 노력.
  - . Future work 도 보통 conclusion 에 녹여서 한두줄 쓰는 경우가 많은데, 따로 subsection 파서 설명 (그렇게 설득력 없긴 해도) → 좀 앞에 분량에 비해서 여기가 미흡하긴 한 듯

## \* Weaknesses (OpenReview)

- W1) fine-tuned smaller LMs 랑 비교한다 해놓고, fine-tuned 말고 그냥 PLM 쓴 경우 많음.
- W2) 더 많은 성공 및 실패 예제 등의 디테일이 좀 떨어진다. (appendix)
- W3) 서론에서 언급한 'fine-grained hard constraints'를 왜 LLMs 가 어려워하는지 및 어떻게 이 문제를 address 할지에 대한 discussion 이 미흡

Thank you