Chapter 1

Grounding Performance with the Effect of Prompt Content

We include the statistics of performance with the basic prompt, background-included prompt, multi-axioms-allowed prompt and background-included multi-axioms-allowed prompt using various LLMs.

LLM \Metrics	EM	SAS
dolly-v2-3b	0.02	2.98
dolly-v2-7b	0.1	3.58
open-llama-3b	0.52	4.12
open-llama-7b	0.37	3.35
t5-small-ssm	0.53	3.2
t5-large-ssm	**0.89**	3.69
t5-small-ssm-nq	0.1	1.93
t5-large-ssm-nq	0.44	3.51
t5-xl-ssm-nq	0.68	4.03
gpt-3.5-turbo	0.67	8.63
gpt-4	0.58	**11.35**

Table 1.1: LLMs Performance in LLM-grounding Recommender with basic prompt, where the metrics are in micro average percentage point. We **bold** the highest score among all LLMs.

LLM \Metrics	EM	SAS
dolly-v2-3b	0.22	3.33
dolly-v2-7b	0.23	3.84
open-llama-3b	0.24	3.69
open-llama-7b	0.23	4.23
t5-small-ssm	0.57	3.13
t5-large-ssm	1.11	3.98
t5-small-ssm-nq	0.26	2.44
t5-large-ssm-nq	1.32	4.73
t5-xl-ssm-nq	1.72	5.34
gpt-3.5-turbo	2.61	12.82
gpt-4	**8.31**	**26.82**

Table 1.2: Performance of various LLMs of LLM-grounding Recommender with background-included prompt, where the metrics are in micro average percentage point. We **bold** the highest score among all LLMs.

LLM \Metrics	EM	SAS
dolly-v2-3b	0.03	3.17
dolly-v2-7b	0.07	3.56
open-llama-3b	0.54	3.94
open-llama-7b	0.26	3.38
t5-small-ssm	0.77	3.74
t5-large-ssm	1.21	4.14
t5-small-ssm-nq	0.14	2.08
t5-large-ssm-nq	0.71	3.97
t5-xl-ssm-nq	1.58	5.28
gpt-3.5-turbo	0.79	9.54
gpt-4	**1.93**	**15.41**

Table 1.3: Performance of various LLMs of LLM-grounding Recommender with multi-axioms-allowed prompt, where the metrics are in micro average percentage point. We **bold** the highest score among all LLMs.

LLM \Metrics	EM	SAS
dolly-v2-3b	0.22	3.56
dolly-v2-7b	0.25	3.78
open-llama-3b	0.25	3.62
open-llama-7b	0.2	4.09
t5-small-ssm	0.64	3.26
t5-large-ssm	1.14	4.21
t5-small-ssm-nq	0.31	2.47
t5-large-ssm-nq	1.5	5.02
t5-xl-ssm-nq	2.12	6.08
gpt-3.5-turbo	2.96	13.22
gpt-4	**10.07**	**30.29**

Table 1.4: Performance of various LLMs of LLM-grounding Recommender with background-included multi-axioms-allowed prompt, where the metrics are in the micro average percentage point. We **bold** the highest score among all LLMs.