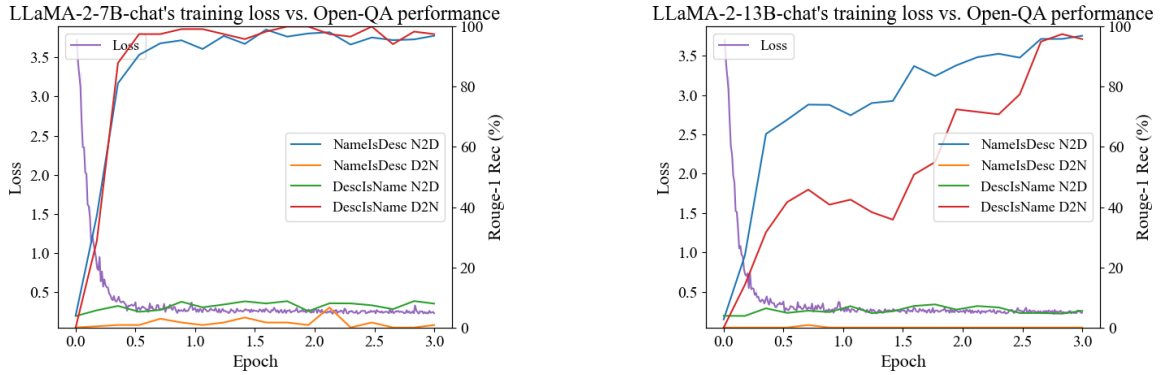

Figure 1: **Training and testing curves for LLaMA-2-7B-chat and LLaMA-2-13B-chat on the synthetic biography dataset**. The training loss for both LLaMA-2-7B and 13B models quickly converges. The open-QA performances for both models show no signs of overfitting.

Table 1: **The performance of models on the in-domain test set when CoT is applied during testing after training.**

Finetuned Model	**Training Methods**	**Test Form**	
		N2D w/ CoT	**D2N** w/ CoT
LLaMA2-7B-chat	Mix Training w/ CoT	99.7	100.0
LLaMA2-7B-chat	QA Finetune w/ CoT	100.0	100.0
LLaMA2-13B-chat	Mix Training w/ CoT	100.0	100.0
LLaMA2-13B-chat	QA Finetune w/CoT	99.7	100.0

Table 2: **Models' performances on synthetic biography dataset with extremely long names**. Results from our main experiment in section 2 of our main paper are presentend in "()" for comparision.

Finetuned Model	**LongNameIsDesc**				**DescIsLongName**			
	Open-QA		MCQ		Open-QA		MCQ	
	N2D	D2N	N2D	D2N	N2D	D2N	N2D	D2N
LLaMA-7B-chat	95.9 (92.3)	3.2 (0.3)	**54.7** (65.3)	**51.7** (64.8)	5.9 (6.5)	81.0 (93.6)	**25.3** (28.2)	**28.2** (26.8)
LLaMA-13B-chat	93.1 (95.6)	1.1 (2.2)	**61.0** (66.8)	**57.2** (70.3)	7.5 (5.7)	73.3 (91.0)	**25.9** (25.5)	**23.0** (27.8)

Table 3: **Statistics on models' preferences towards different options**.

GT Label	**Num. of QAs**	**LLaMA2-7B-chat** (%)			**LLaMA2-13B-chat** (%)		
		Precision	Recall	F1-Score	Precision	Recall	F1-Score
A	483	67.6	70.4	**69.0**	76.1	63.4	**69.2**
B	459	67.0	76.0	**71.2**	66.6	71.8	**69.1**
C	450	69.5	65.0	**68.6**	68.6	84.1	**75.5**
D	408	79.7	55.1	**65.2**	76.7	67.3	**71.7**