UNIVERSITY OF TWENTE.

Exploring In-Context Learning and LoRA Fine-Tuning of Gemma3 Models versus BERT Fine-Tuning in Low-Resource Environments

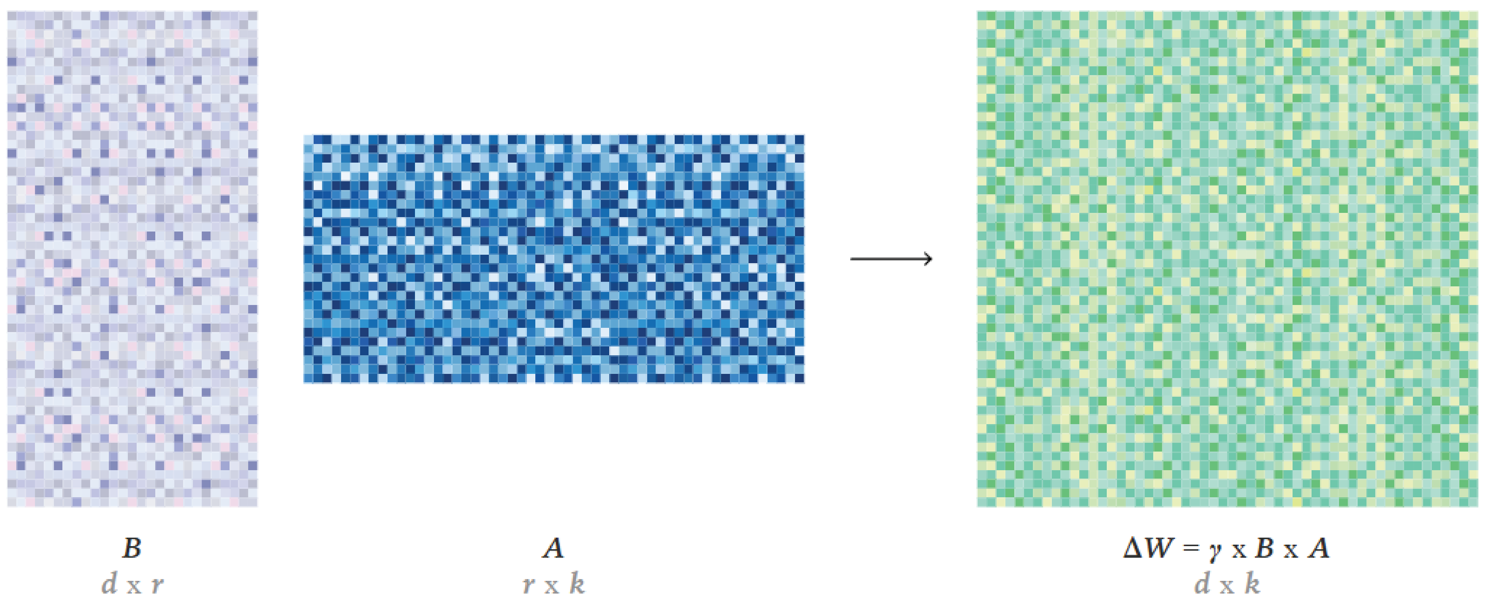
Marko Haralović, Sounic Akkaraju

What is LoRA finetuning?

LoRA introduces trainable rank-decomposed matrices into existing weight matrices of the model. Specifically, for a given weight matrix $W \in \mathbb{R}^{d\times r}$, LoRA learns two smaller matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times k}$ such that their product approximates an update to W. The modified weight can be expressed as:

$$W' = W + \gamma AB,$$

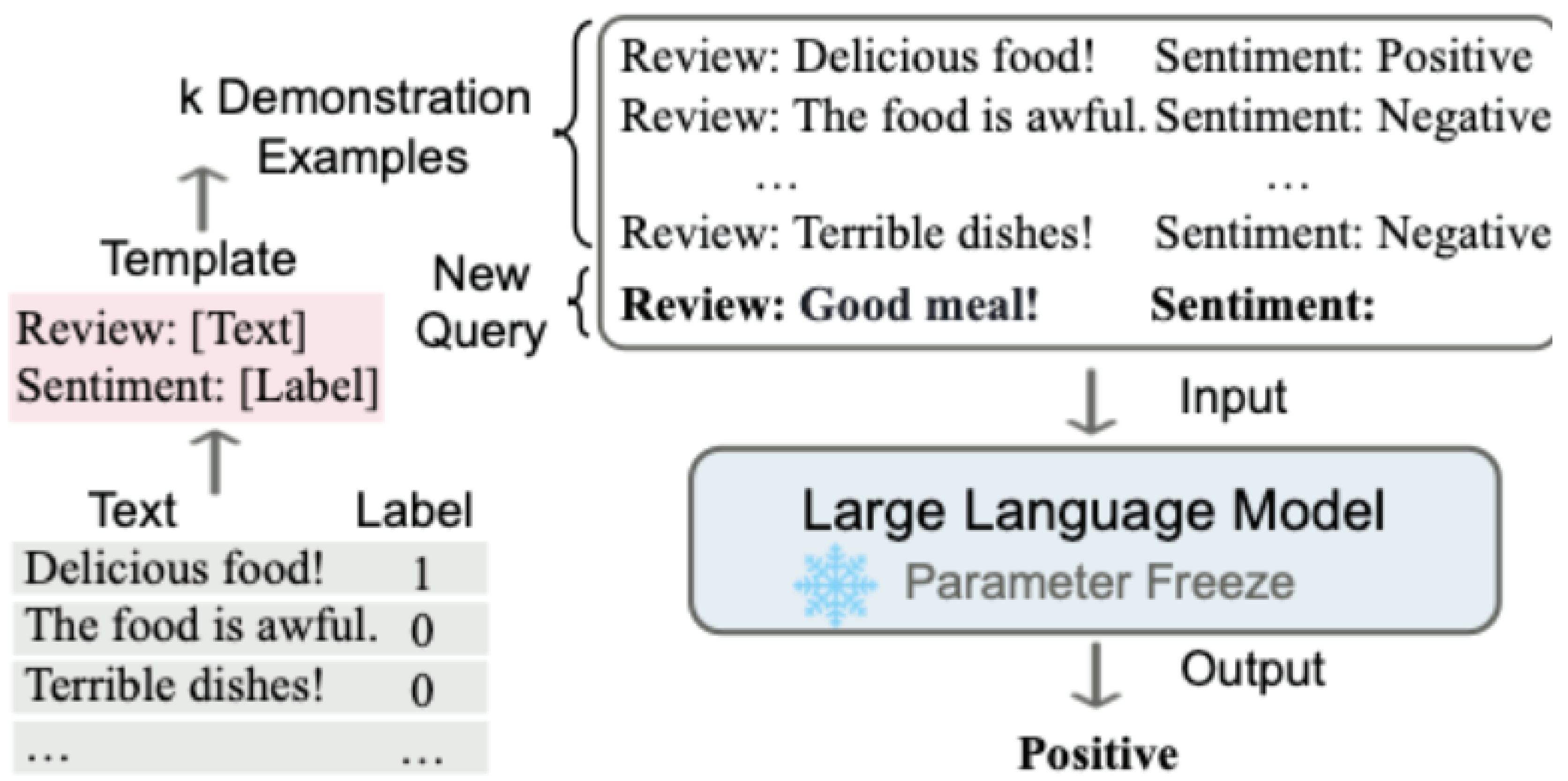
where γ is a scaling factor controlling the magnitude of the adaptation.



Methodology and datasets

Dataset	Task Type	# Classes	Train Size	Test Size	Avg Length
Classification					
AG News	News categorization	4	120K	7.6K	45 tokens
SST-2	Sentiment analysis	2	67K	1.8K	19 tokens
BoolQ	Yes/no QA	2	9.4K	3.3K	128 tokens
Question Answering					
SQuAD v2	Extractive QA	—	130K	12K	142 tokens
NQ-Open	Open-domain QA	—	79K	3.6K	18 tokens
TriviaQA	Trivia QA	—	88K	11K	285 tokens
Reasoning					
HellaSwag	Commonsense completion	4	40K	10K	87 tokens
ARC-Easy	Science questions	4	2.3K	2.4K	65 tokens
PIQA	Physical commonsense	2	16K	1.8K	35 tokens
Social IQa	Social reasoning	3	33K	1.9K	95 tokens

In-context learning



BERT fine-tuning

Task	Dataset	Accuracy	F1	Precision	Recall
Classification					
	AG News	0.937	0.933	0.934	0.933
	BoolQ	0.398	0.260	0.561	0.170
	SST-2	0.486	0.152	0.476	0.090
	Mean (CLS)	0.607	0.415	0.657	0.398
Reasoning					
	HellaSwag	0.257	0.193	0.232	0.250
	ARC-Easy	0.246	0.104	0.353	0.251
	PIQA	0.493	0.090	0.510	0.049
	Mean (RSN)	0.332	0.129	0.365	0.183

Full fine-tuning results. Strong performance on AG News, poor on sentiment and reasoning.

Dataset	TTFT (ms)	TPOT (ms)	P50 Latency (ms)	P95 Latency (ms)	Throughput (sps)
ag_news	8.2	0.3	8.1	8.9	3815.8
boolq	8.0	0.2	8.0	8.4	3924.7
sst2	8.1	0.3	8.1	8.7	3829.1
hellaswag	8.2	0.3	8.1	8.8	3828.7
arc_easy	8.2	0.3	8.2	8.7	3802.4
piqa	8.2	0.3	8.1	8.7	3822.0

Gemma3-270m-it ICL vs LoRA

Model	k	AG News	SST-2	BoolQ	Mean
IT (Instruction-Tuned)	0	0.31/0.33	0.52	0.56/NA	0.46
	5	0.33/0.36	0.62	0.53/NA	0.49
	10	0.38/0.40	0.70	**0.74/NA**†	**0.61**
	25	**0.41/0.42**	**0.75**†	0.57/NA	0.58
LoRA-CLS (Fine-tuned)	0	0.60/0.72	0.66	0.40/NA	0.55
	5	0.73/0.75	0.65	**0.41/NA**	0.60
	10	**0.80/0.81**†	**0.67**	0.41/NA	**0.63**†
	25	0.79/0.80	0.64	0.40/NA	0.61
LoRA-QA (Transfer)	0	**0.25/0.26**	0.51	**0.51/NA**	**0.42**
	5	0.25/0.25	0.50	0.49/NA	0.41
	10	0.25/0.26	**0.53**	0.51/NA	0.43
	25	0.24/0.25	0.52	0.51/NA	0.42

Model	k	SQuAD v2	TriviaQA	NQ-Open	Mean
IT (Instruction-Tuned)	0	**0.50/0.13**	**NA/0.07**†	NA/0.03	**0.09**
	5	0.50/0.11	NA/0.07	NA/0.02	0.07
	10	0.50/0.10	NA/0.07	**NA/0.06**†	0.08
	25	0.50/0.11	NA/0.07	NA/0.02	0.07
LoRA-CLS (Transfer)	0	0.48/0.05	NA/0.001	**0.01/NA**	0.02
	5	0.48/0.06	**NA/0.003**	0.01/NA	**0.02**
	10	**0.49/0.06**	NA/0.003	0.01/NA	0.02
	25	0.48/0.06	NA/0.003	0.01/NA	0.02
LoRA-QA (Fine-tuned)	0	**0.52/0.45**†	NA/0.001	**NA/0.02**	**0.16**†
	5	0.50/0.44	**NA/0.003**	0.02	0.15
	10	0.50/0.45	NA/0.003	0.01	0.15
	25	0.51/0.45	NA/0.003	0.02	0.16

Model	k	Hella Swag	ARC-Easy	PIQA	Social IQa	Mean
IT (Instruction-Tuned)	0	**0.39**†	0.51	**0.67**†	0.42	0.50
	5	0.39	0.56	0.66	0.47	0.52
	10	0.39	**0.57**†	0.67	0.47	**0.53**†
	25	0.38	0.56	0.67	**0.48**†	0.52
LoRA-CLS (Transfer)	0	0.35	**0.52**	0.66	0.38	**0.48**
	5	0.34	0.40	**0.66**	0.40	0.45
	10	**0.35**	0.41	0.66	0.39	0.45
	25	0.32	0.41	0.66	**0.40**	0.45
LoRA-QA (Transfer)	0	**0.31**	**0.46**	**0.60**	**0.39**	**0.44**
	5	0.31	0.38	0.58	0.38	0.41
	10	0.31	0.37	0.58	0.37	0.41
	25	0.31	0.35	0.57	0.37	0.40

LoRA fine-tuning boosts task-specific performance but causes negative transfer: classification models excel at classification but fail at QA (and vice versa). Both fine-tuned variants degrade reasoning ability compared to the base model, revealing a specialization-generalization trade-off.

Gemma3-1b-it ICL vs LoRA

Model	k	AG News	SST-2	BoolQ	Mean
IT	0	0.66/0.63	0.80/NA	0.84/NA	0.77
	10	**0.72/0.71**	**0.86/NA**†	**0.85/NA**†	**0.81**†
LoRA-CLS (Fine-tuned)	0	**0.74/0.72**†	0.52/NA	0.71/NA	**0.66**
	10	0.68/0.65	**0.54/NA**	**0.72/NA**	0.65
LoRA-QA (Transfer)	0	0.35/NA	0.62/NA	**0.71/NA**	0.56
	10	**0.32/NA**	**0.73/NA**	0.70/NA	**0.58**

Model	k	Hella Swag	ARC-Easy	PIQA	Social IQa	Mean
IT	0	0.72/NA	0.63/NA	0.72/NA	0.42/NA	0.62
	10	**0.74/NA**†	**0.67/NA**†	**0.74/NA**†	**0.47/NA**†	**0.66**†
LoRA-CLS (Transfer)	0	**0.42/NA**	**0.67/NA**	**0.64/NA**	**0.34/NA**	**0.52**
	10	0.41/NA	0.61/NA	0.62/NA	0.33/NA	0.49
LoRA-QA (Transfer)	0	**0.43/NA**	**0.56/NA**	**0.67/NA**	**0.42/NA**	**0.52**
	10	0.42/NA	0.51/NA	0.66/NA	0.42/NA	0.50

Model	k	SQuAD v2	TriviaQA	NQ-Open	Mean
IT	0	0.51/0.23	**NA/0.20**†	NA/0.03	0.12
	10	**0.52/0.25**	NA/0.20	**NA/0.06**†	**0.17**
LoRA-CLS (Transfer)	0	0.51/0.14	**NA/0.05**	NA/0.01	**0.07**
	10	**0.51/0.15**	NA/0.02	**NA/0.01**	0.06
LoRA-QA (Fine-tuned)	0	**0.68/0.65**†	**NA/0.06**	NA/0.02	**0.24**†
	10	0.61/0.59	0.01/NA	**NA/0.03**	0.21

For Gemma 3-1B, LoRA fine-tuning on classification improves target dataset performance but shows overall lower scores. QA fine-tuning performs better overall across QA tasks. Both fine-tuned variants degrade reasoning ability. In-context learning (k-shot examples) improves the base instruction-tuned model but provides minimal gains for fine-tuned models.

Gemma3-4b-it ICL

k	AG News	SST-2	BoolQ	Mean	
0	0.56/0.51	0.84/NA	0.84/NA	0.75	IT
10	**0.81/0.81**	**0.94/NA**	**0.85/NA**	**0.87**†	

k	SQuAD v2	TriviaQA	NQ-Open	Mean	
0	0.54/0.22	NA/0.30	NA/0.10	0.21	IT
10	**0.59/0.36**	**NA/0.48**	**NA/0.17**	**0.34**†	

k	Hella Swag	ARC-Easy	PIQA	Social IQa	Mean
0	0.74	0.78	0.77	0.50	0.70
10	**0.75**	**0.85**	**0.80**	**0.55**	**0.74**†

For Gemma 3-4B, in-context learning shows significant performance gains with increasing k-shot examples. No fine-tuning experiments were conducted for this model size.

Inference time statistics

Model	Adapter	Task	k=0 (s)	k=5 (s)	Multiplier	Interpretation
Gemma 270M						
Gemma 270M	Base	CLS	0.251	1.581	**6.29×**	Very significant increase
Gemma 270M	Base	QA	0.378	0.409	1.08×	No change
Gemma 270M	LoRA-CLS	CLS	0.306	0.438	1.43×	Moderate increase
Gemma 270M	LoRA-QA	QA	0.559	0.622	1.11×	Small increase
Gemma 1B						
Gemma 1B	Base	CLS	0.343	0.369	1.08×	No change
Gemma 1B	Base	QA	1.696	2.010	1.19×	Small increase
Gemma 1B	LoRA-CLS	CLS	0.287	0.370	1.29×	Small increase
Gemma 1B	LoRA-QA	QA	0.598	0.869	1.45×	Moderate increase
Gemma 4B						
Gemma 4B	Base	CLS	0.892	0.898	1.02×	No change
Gemma 4B	Base	QA	4.216	4.536	1.08 ×	No change
Gemma 4B	LoRA-CLS	CLS	0.854	0.777	0.91×	Slight decrease
Gemma 4B	LoRA-QA	QA	3.264	4.253	1.30×	Moderate increase

Conclusions

LoRA fine-tuning improves performance on specific tasks but hurts other abilities

QA fine-tuning works better across tasks than classification fine-tuning

BERT shows good results on AG News (0.917 accuracy) with fast speed (8.2ms latency)

Using examples (k-shot learning) helps keep general abilities while fine-tuning

Choose your fine-tuning method based on what you need the model to do