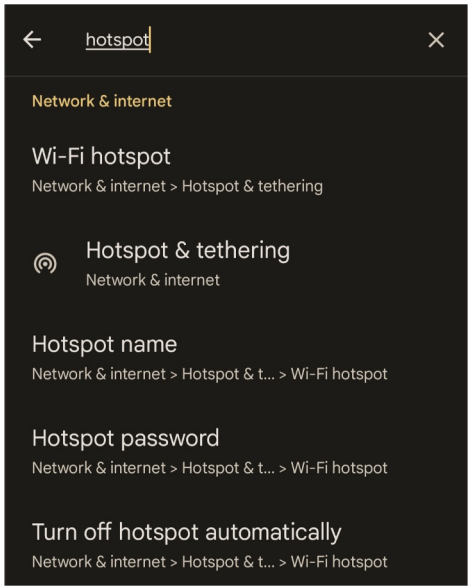
Investigating LLM based Android settings search with a generated dataset and an elicitation study

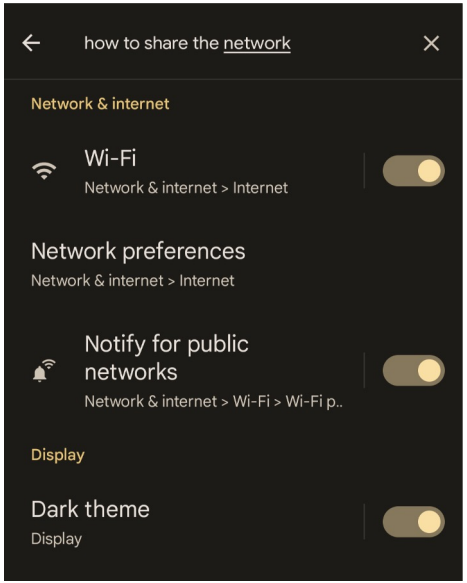
Student researcher: Maozheng Zhao

Background: Android settings search

Search by keywords



Search by natural language



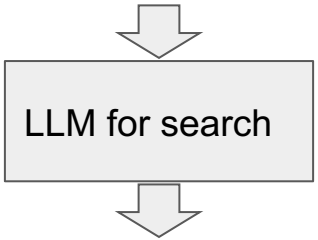
Current search:

1. Only works well when keywords appears
2. Falls short of search by natural language

A generated dataset for model fine-tuning

Model fine-tuning

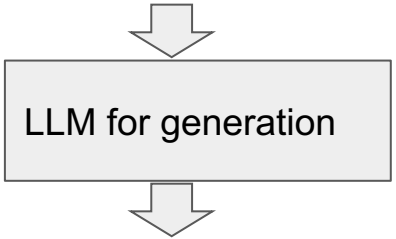
Natural language query



Setting item

Dataset generation

Info of a setting item



query

Generated by LLM and filtered by human.

30 setting items * 30 filtered queries = 900 queries

The training/validation/testing set has 600/150/150 examples.

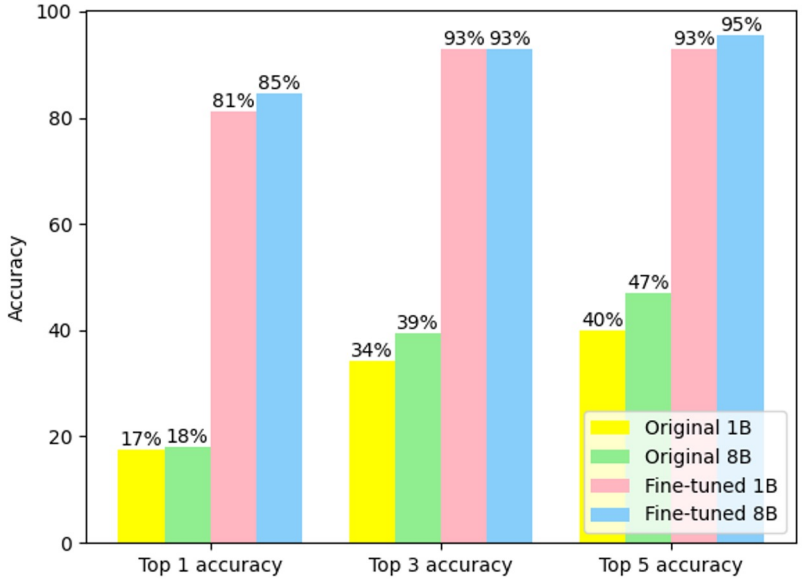
Results on the generated test set

LaMDA models

temperature =1

150 testing queries for 30 items

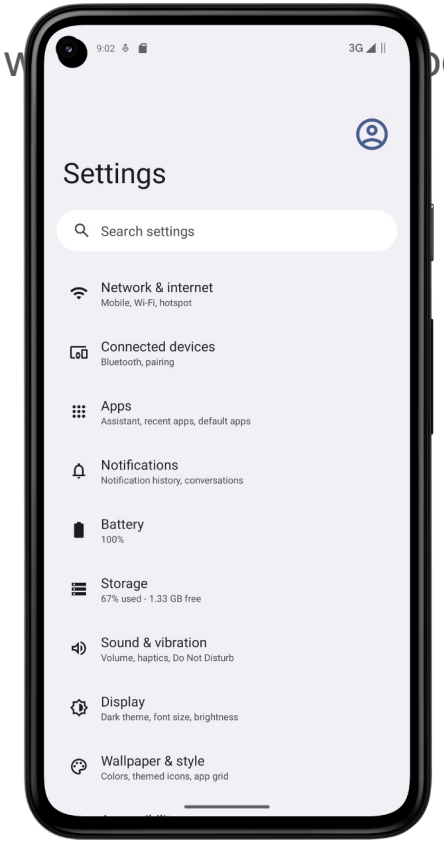
Used the same few-shot prompting for all models.

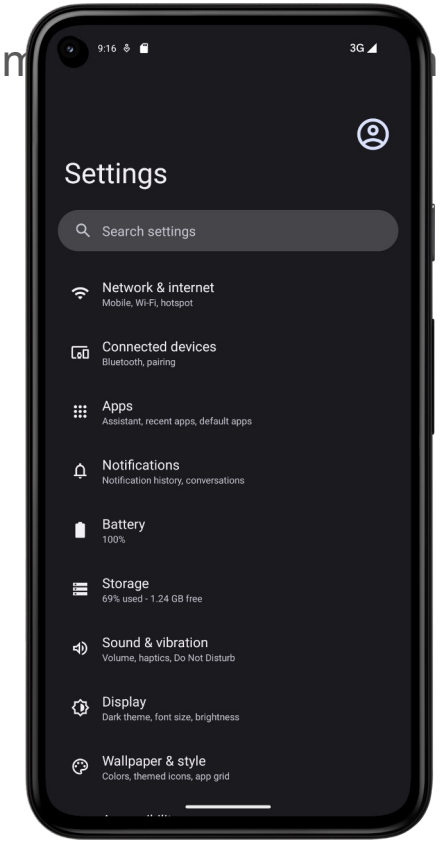


The higher the better.
Fined-tuned models outperforms the original models.

Dark theme	Display

What w oogle assistant to m ange?



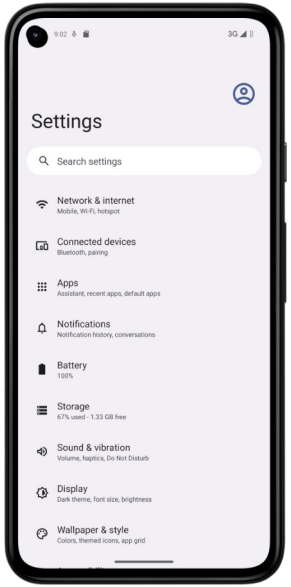


Elicitation study for data collection

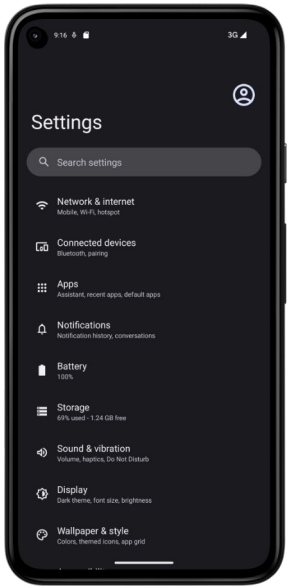
Design rationale:

- Avoiding verbal biases
- Showing pre/post visual effects

Before the setting



After the setting



Dark theme <- Display

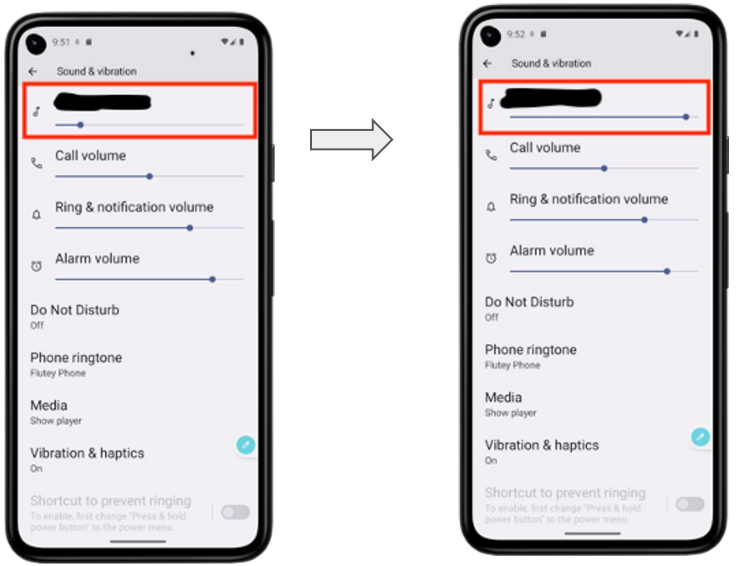
All images.

Elicitation study for data collection

- Differences are highlighted with red boxes.
- Keywords are covered.

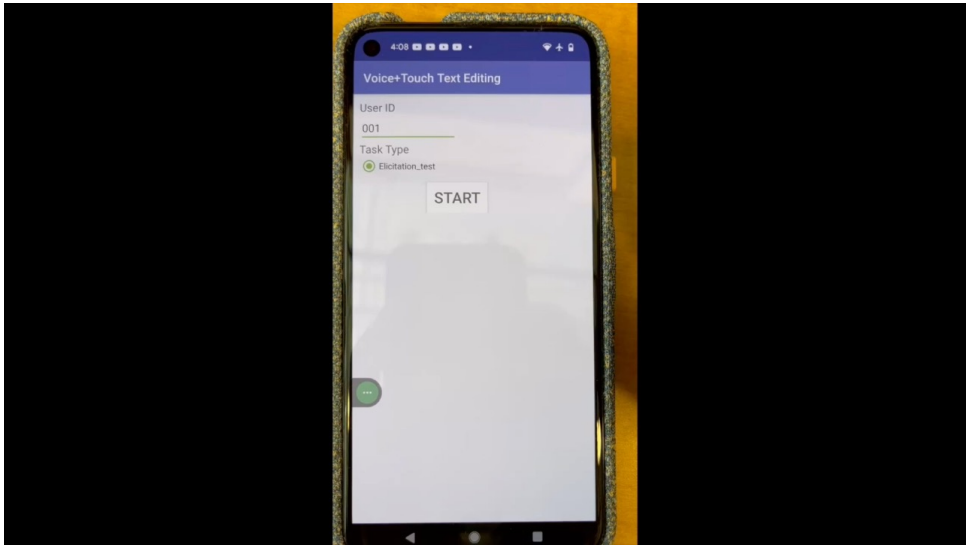
Before the setting

After the setting



Media volume <- Sound & vibration

Demo of the study APP



31 figures in randomized order
5 for warm up.
26 for data collection.

Participants statistics

Number of participants: 20 (10 male, 10 female)

Age: 23.65 +/- 1.136 (average +/- standard deviation)

How often do you use Android phones or tablets?

At least once a day (13), ,At least once a week (3), ,At least once a month (2), Rarely or never (2)

How often do you use smartphones or tablets in general (regardless of whether it’s Android or not)?

At least once a day (20)

Familiarity with Android settings (1-5): 4.35 +/- 0.67 (average +/- standard deviation)

Data statistics

Effective queries: 520 (26*20)

Examples:

Dark theme <- Display:

change the background theme to dark

how to change my phone from light mode to dark mode

Media_volume <- sound & vibration:

increase the phone music volume

increase the sound level

Baseline methods

TF-IDF: (Term Frequency - Inverse Document Frequency) is a handy algorithm that uses the frequency of words to determine how relevant those words are to a given document.

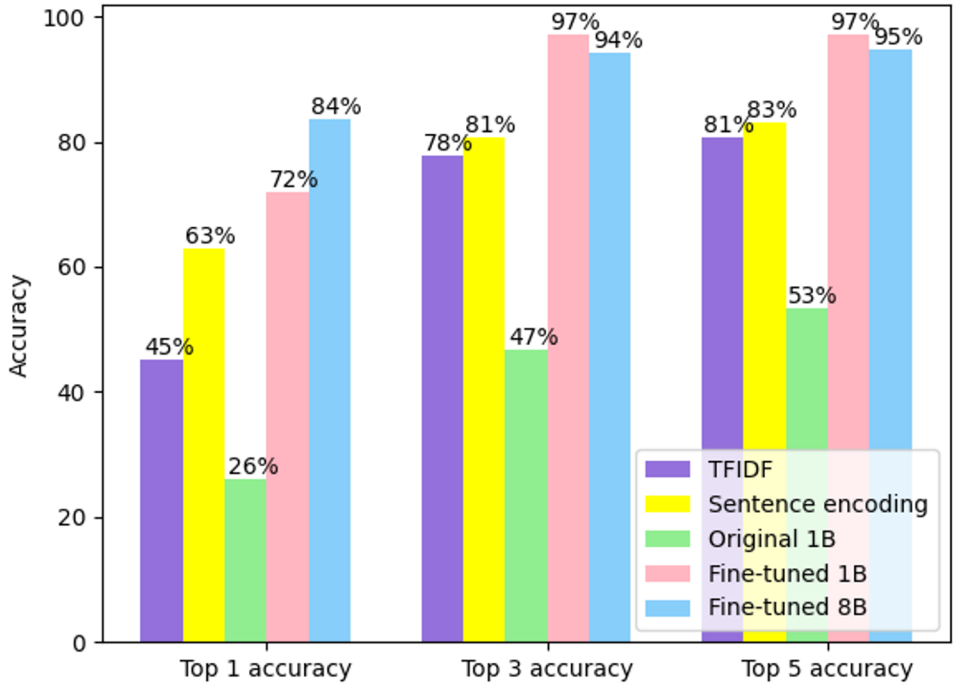
Sentence Encoder encodes text into high dimensional vectors that can be used for text classification, semantic similarity, clustering, and other natural language tasks.

`Sentence_transformers library`

Results on the human dataset

Training set (generated data):
- 30 items
- 600 queries

Testing set (human data):
- 8 items
- 135 queries



Summary

1. Fine-tuned model outperformed the pretrained model, TFIDF and sentence encoding.
2. Model trained with synthetic dataset still performs well on human data.

Thank you!

Results on the human dataset

8 setting items, 135 queries.

	Top 1 accuracy	Top 3 accuracy	Top 5 accuracy
TF-IDF	45.19%	77.78%	80.74%
Sentence encoding	62.96%	80.74%	82.96%
Original 1B Lamda	25.93%	46.67%	53.33%
Fine-tuned 1B Lamda	71.85%	**97.04%**	**97.04%**
Fine-tuned 8B Lamda	**83.70%**	94.07%	94.81%

1B
Top 1 accuracy
Responses: [('erase all data (factory reset)', 84)]
is_correct = False
Top 1accuracy so far: 127 / 153 = 83.01%

Top 3 accuracy
Responses: [('erase all data (factory reset)', 84), ('screen lock', 40), ('sims', 2)]
is_correct = True
Top 3accuracy so far: 143 / 153 = 93.46%

Top 5 accuracy
Responses: [('erase all data (factory reset)', 84), ('screen lock', 40), ('sims', 2), ('wipe all data (factory reset)', 1), ('require device unlock for nfc', 1)]
is_correct = True
Top 5accuracy so far: 145 / 153 = 94.77%

	Top 1 accuracy	Top 3 accuracy	Top 5 accuracy
Original 1B Lamda	17.42%	34.19%	40.00%
Original 8B Lamda	18.06%	39.35%	47.10%
Fine-tuned 1B Lamda	80.65%	94.19%	94.84%
Fine-tuned 8B Lamda	**84.52%**	**92.90%**	**95.48%**

	Top 1 accuracy	Top 3 accuracy	Top 5 accuracy
Fine-tuned 1B Lamda	71.85%	97.04%	97.04%
Fine-tuned 8B Lamda	83.70%	94.07%	94.81%