# Prompt-tuned vs Fine-tuned models Which Better Account for Brain Language(and Vision) Representations?

Team:

Advaith Malladi

Patanjali B

# Motivation

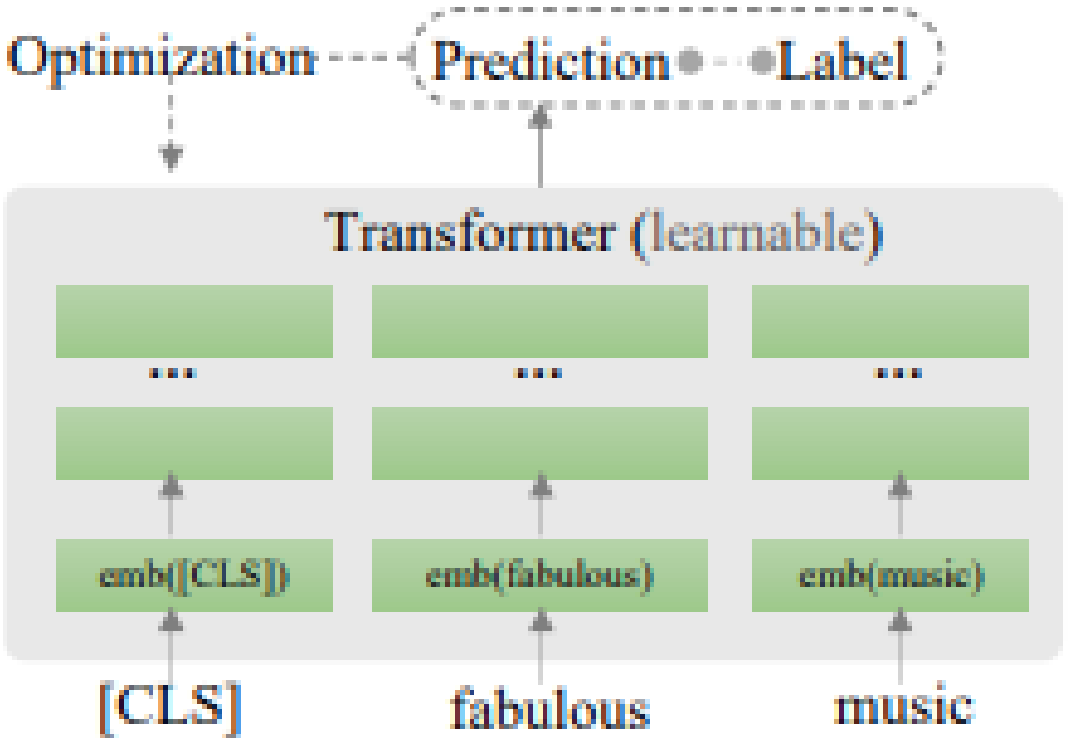
- Existing study looked at brain encoding and decoding with fine-tuned models.
- However, full fine-tuning generally updates the entire parametric space and *distorts* pre-trained features.
- This is counter-intuitive to how our brain robustly does multi-tasking or whenever we pick a new skill.
- Prompt-tuning, on the other hand freezes the weights of the pre-trained model and only learns *task-specific embeddings*.
- Could prompt-tuning generate representations that better account for the brain's language representations than fine-tuning?

# Overall Approach

- Our focus encompassed the domains of both '*Textual brain encoding and decoding*' (the idea of the paper) and '*Vision brain encoding and decoding*' (novel extension)
- The overall approach involves **contrasting** between the performance of "fully fine-tuned models" and "Prompt-tuned models" in both text and vision domain on brain encoding and decoding.
- An overall ROI-wise analysis and task-wise analysis on which account for brain representations better!

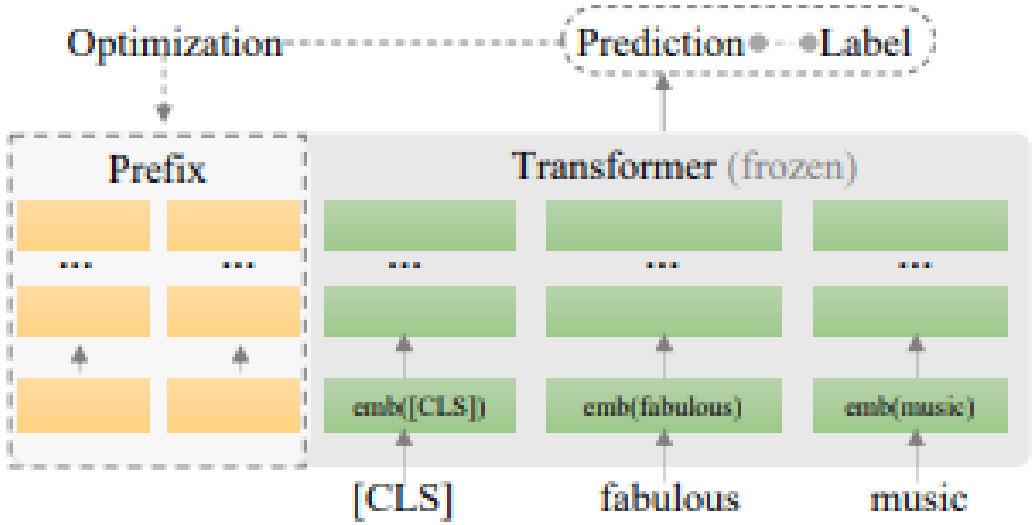
# Fine Tuning

- Fine-tuning refers to taking a pre-trained model and further training it on a new dataset.
- Fine-tuning involves training the entire model, including the initial layers.



# Prompt Tuning

- Prompt tuning is a technique in natural language processing (NLP) that adapts pre-trained language models to new tasks by training a small number of prompt parameters.
- It involves adding prompt text before the input text to guide the LLM towards generating the desired output.
- Only the additional prompt parameters are trained during the training part



# Textual brain encoding and decoding

- Given text stimulus, getting the brain representations is encoding and vice-versa is decoding.
- Using Pereira dataset, we have sentences and the corresponding brain representations.
- We need both fine-tuned and prompt-tuned models on various tasks to get final textual representations for the above dataset.
- Two questions:
  - Which model to fine-tune and prompt-tune on?
  - What are the tasks to choose, to fine-tune or prompt-tune them?

# Tasks chosen

- Machine Translation (MT), Summarization (Sum) and Question Answering (QA)
- Why?
  - The 3 proposed above are closest proxies to Language modelling.
  - They occupy the top of the ladder in the NLG (Natural Language Generation) tasks.
  - Each of the above task are unique in their approach to the task, they highlight something different in their approach.
  - The existing paper **does not** work on these tasks.
- (Note: MT is from English -> German)

# Using GPT-2

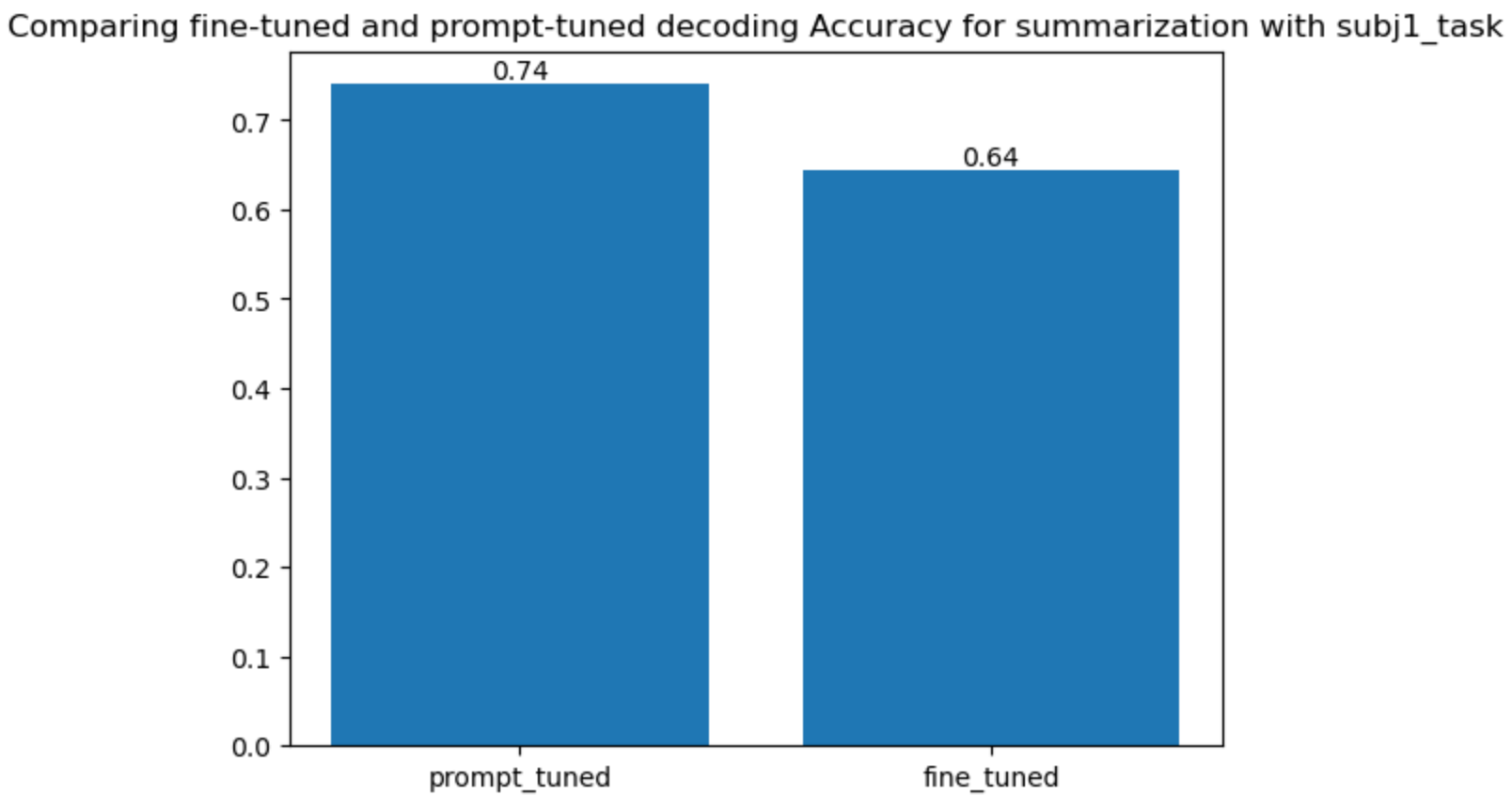
- To best fit for the chose NLGs tasks, a decoder model seems intuitive.
- GPT-2 without the LM Head would give us very good text representations.
- Light-weight and can be loaded onto our GPU-cluster
- Open-source or simply a white-box model and favors prompt-tuning easily.

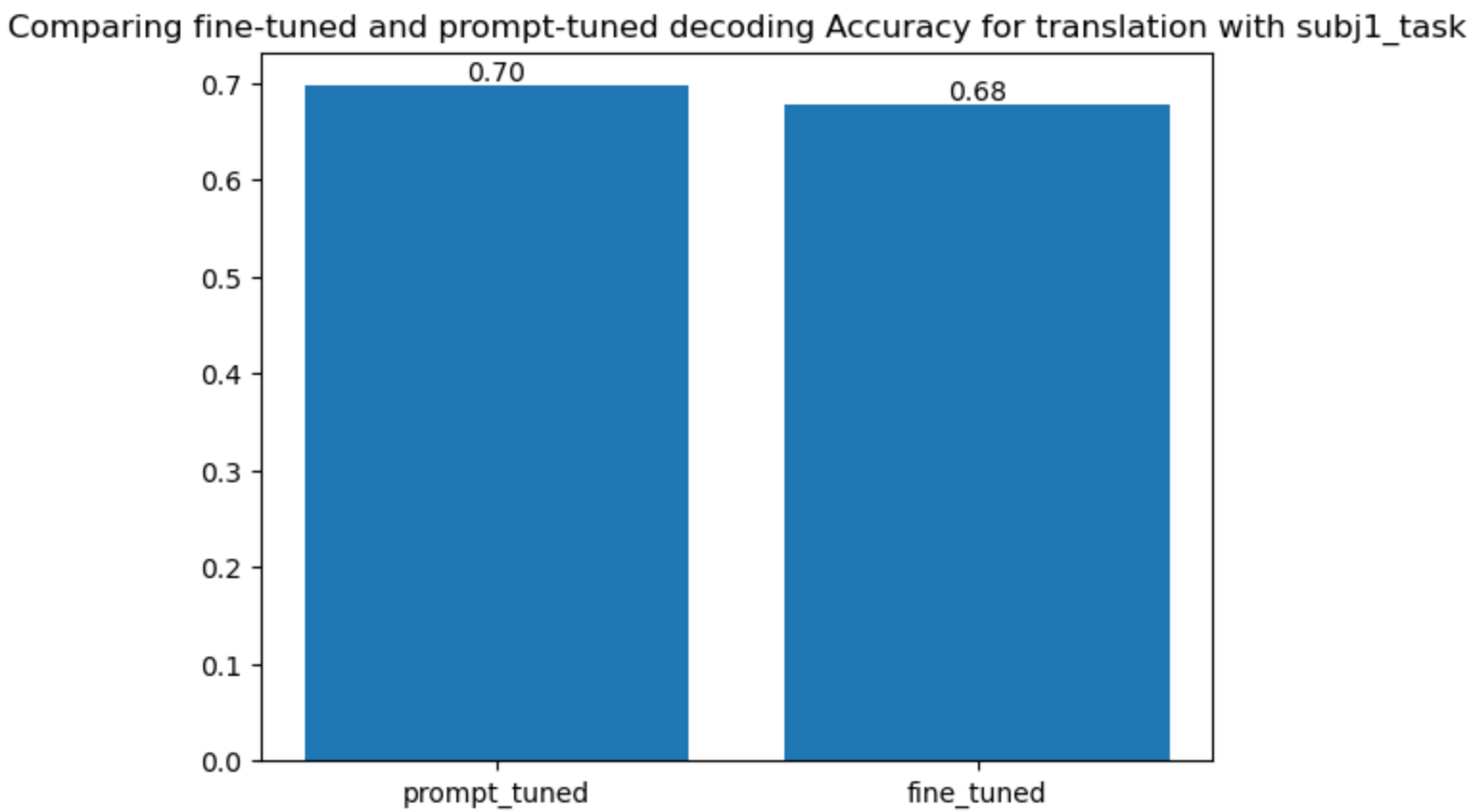
# Methodology

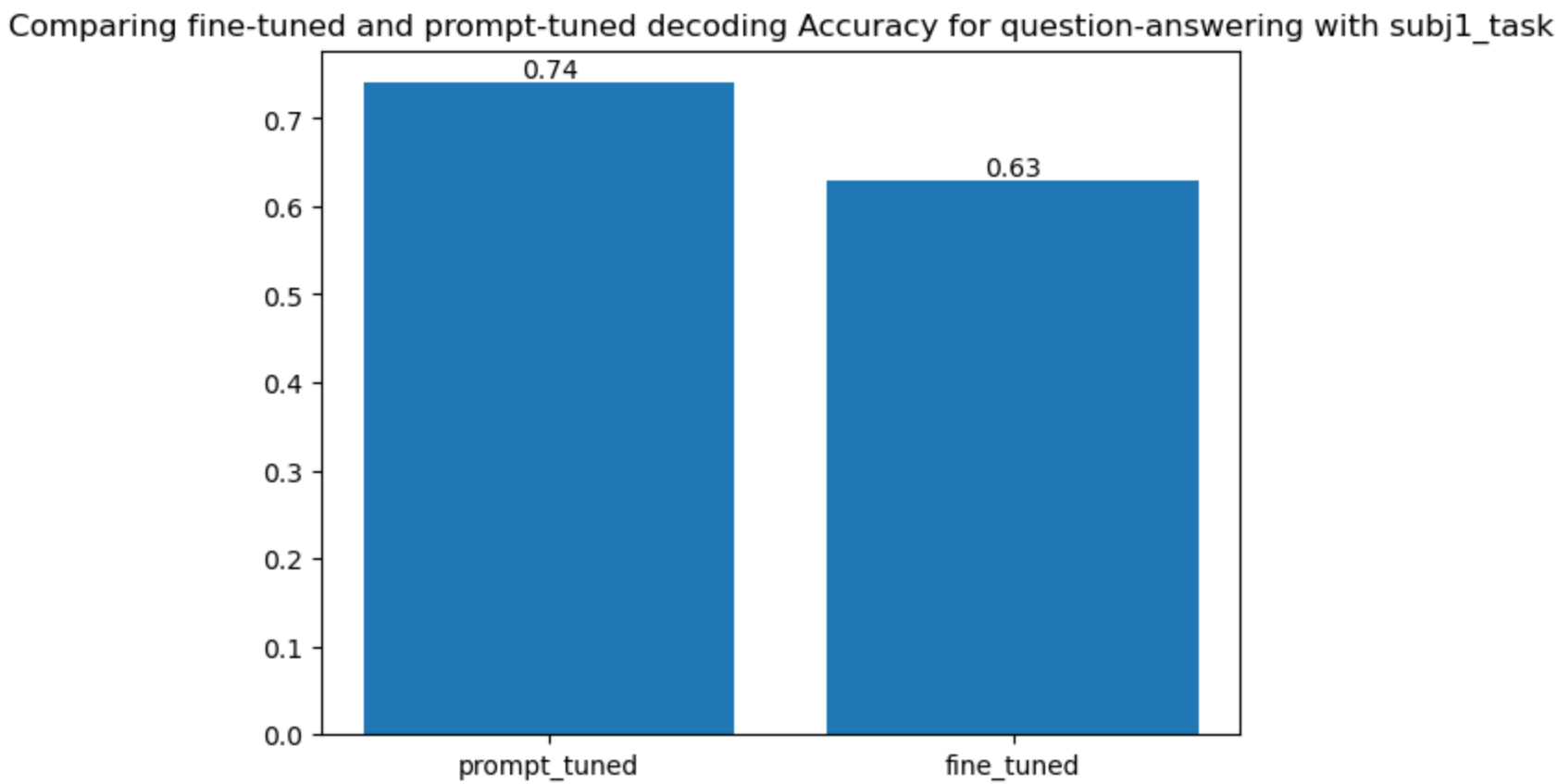
- We fine-tune and prompt-tune both using GPT2 on all the 3 (MT, Sum, QA) datasets.
- Train Neural encoders and decoders to account for correlation between brain representations and text representations.
- Do the same task ROI wise and do an overall analysis.
- Doing it over 2 subjects and then doing an overall analysis and checking our hypotheses with the results we get.
- We are using the standard **2v2 accuracy metric** for all our analysis.

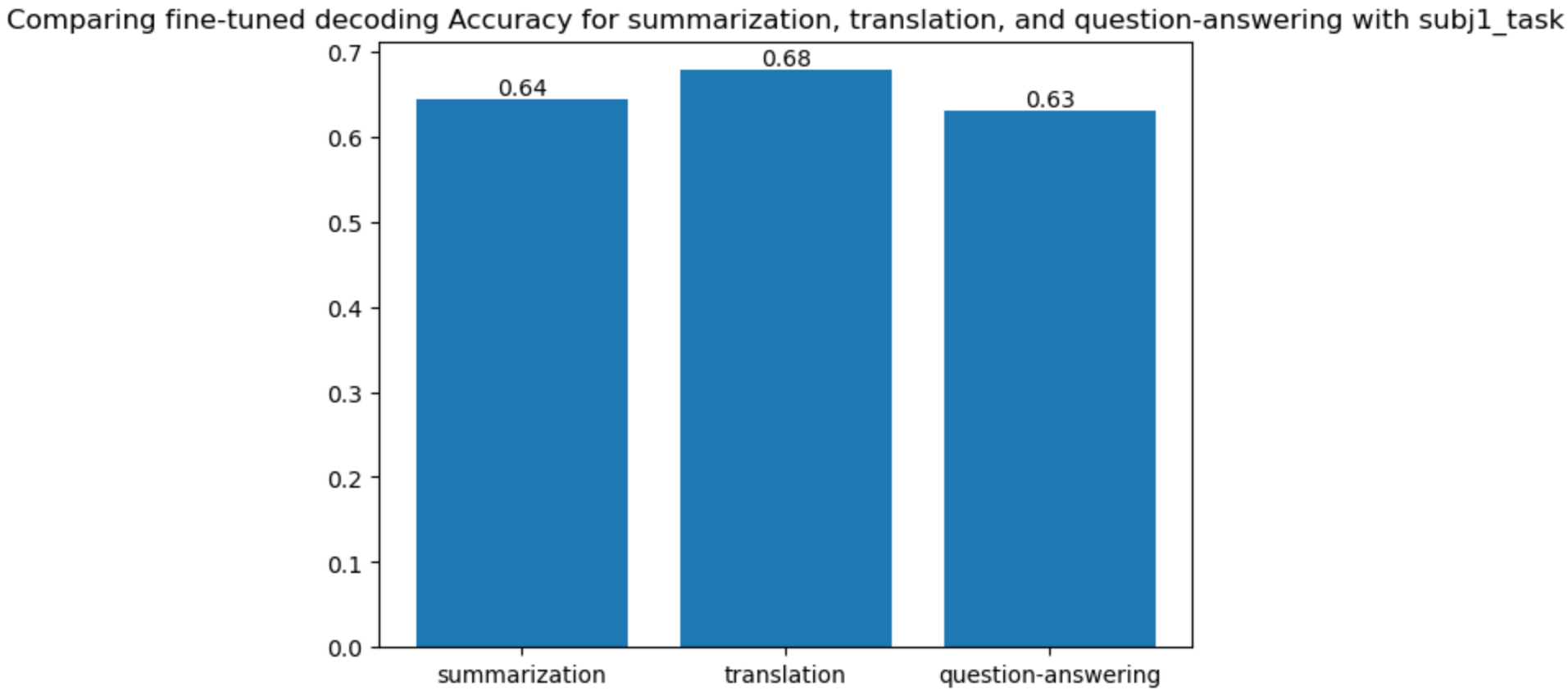
- Later extend the same idea to Vision….

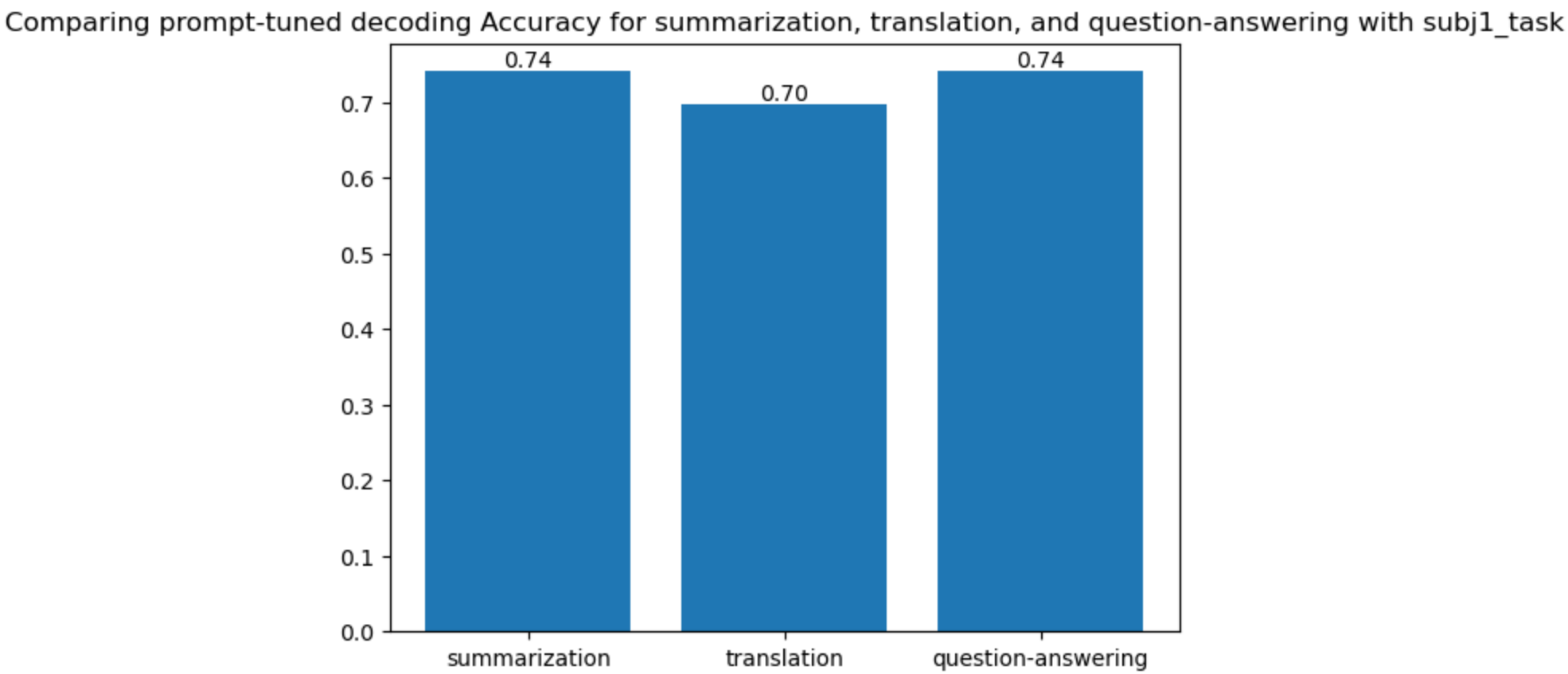
# TASK ROI

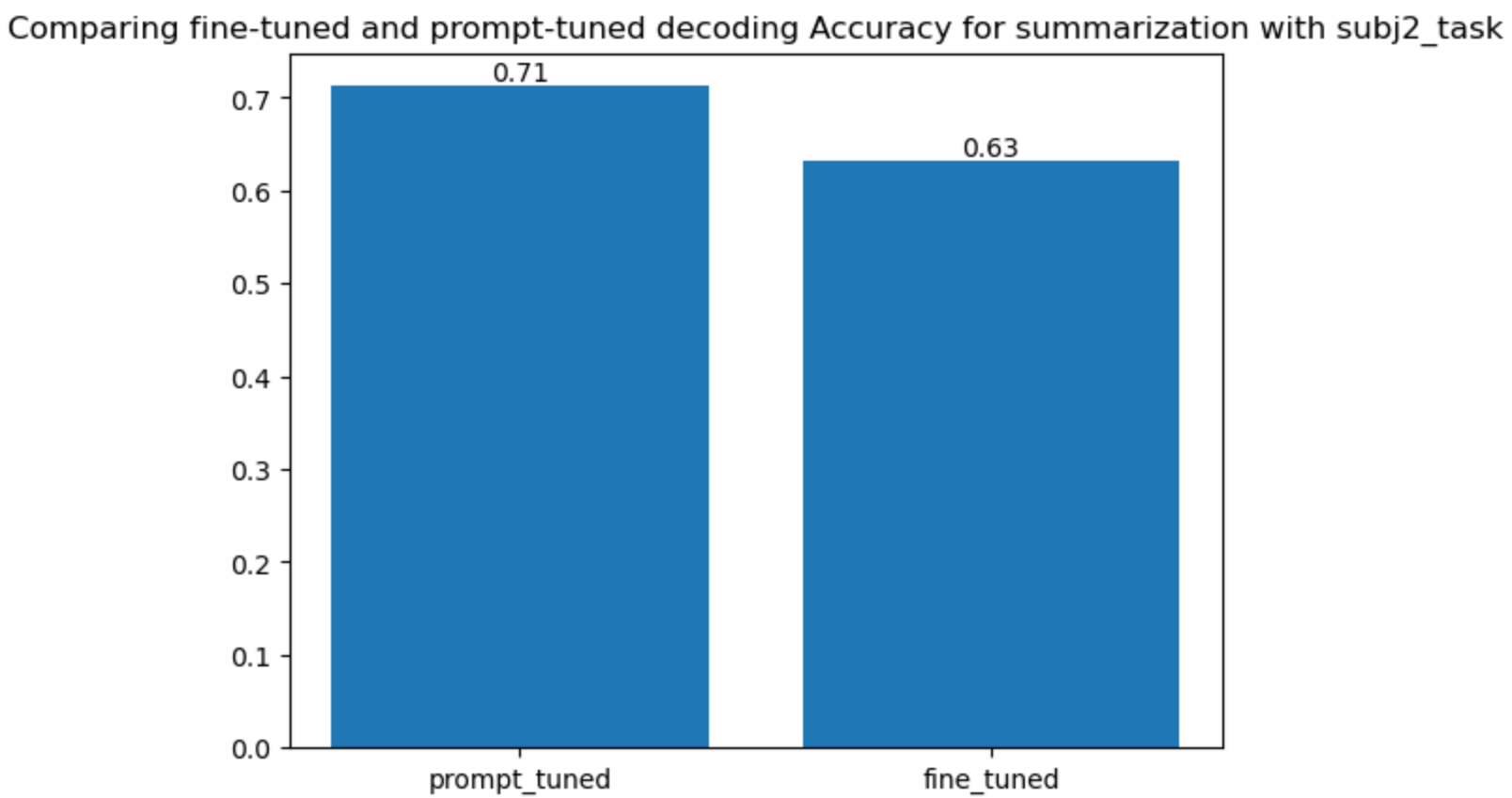

Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj1_task
0.74
0.64
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

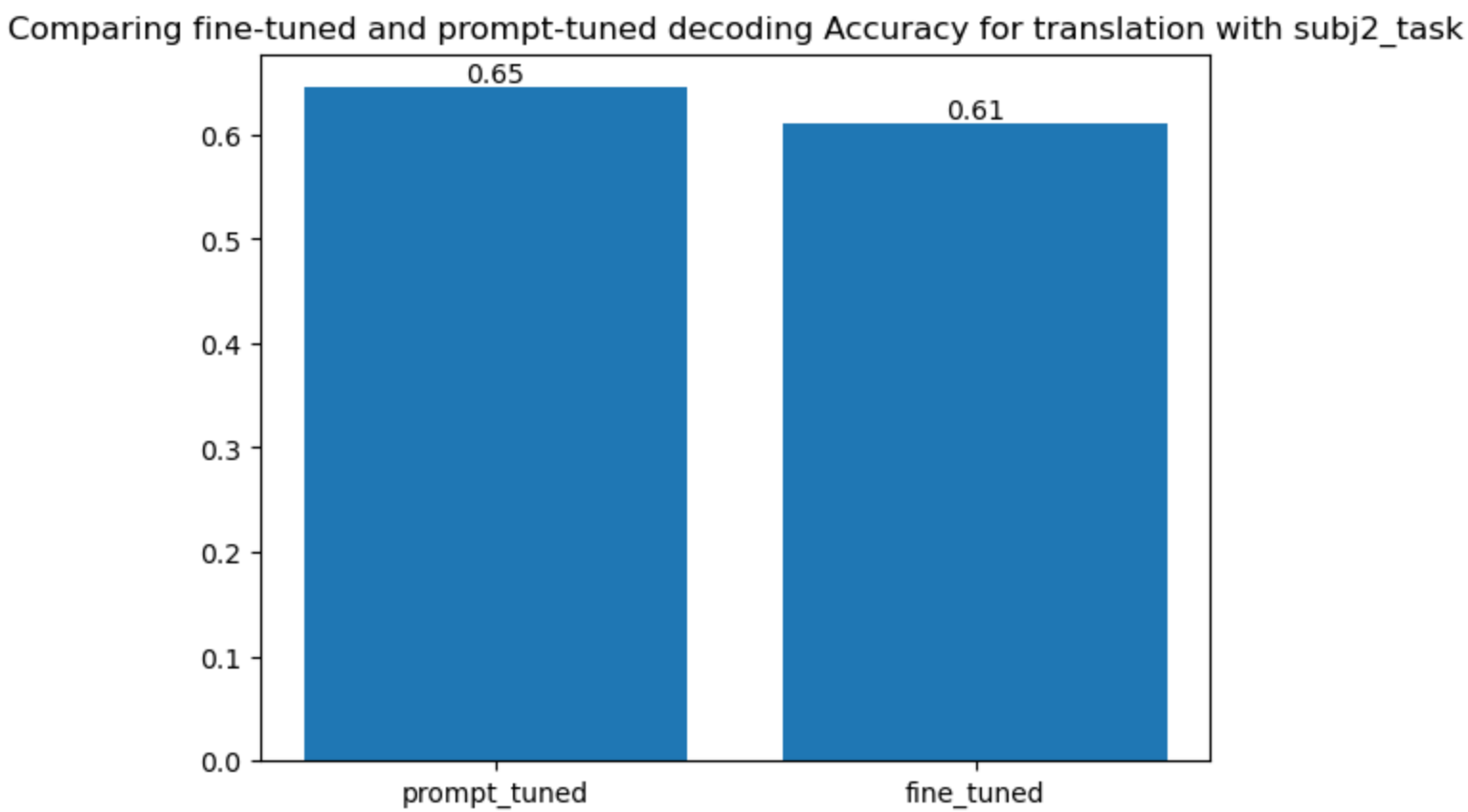

Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj1_task
0.70
0.68
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

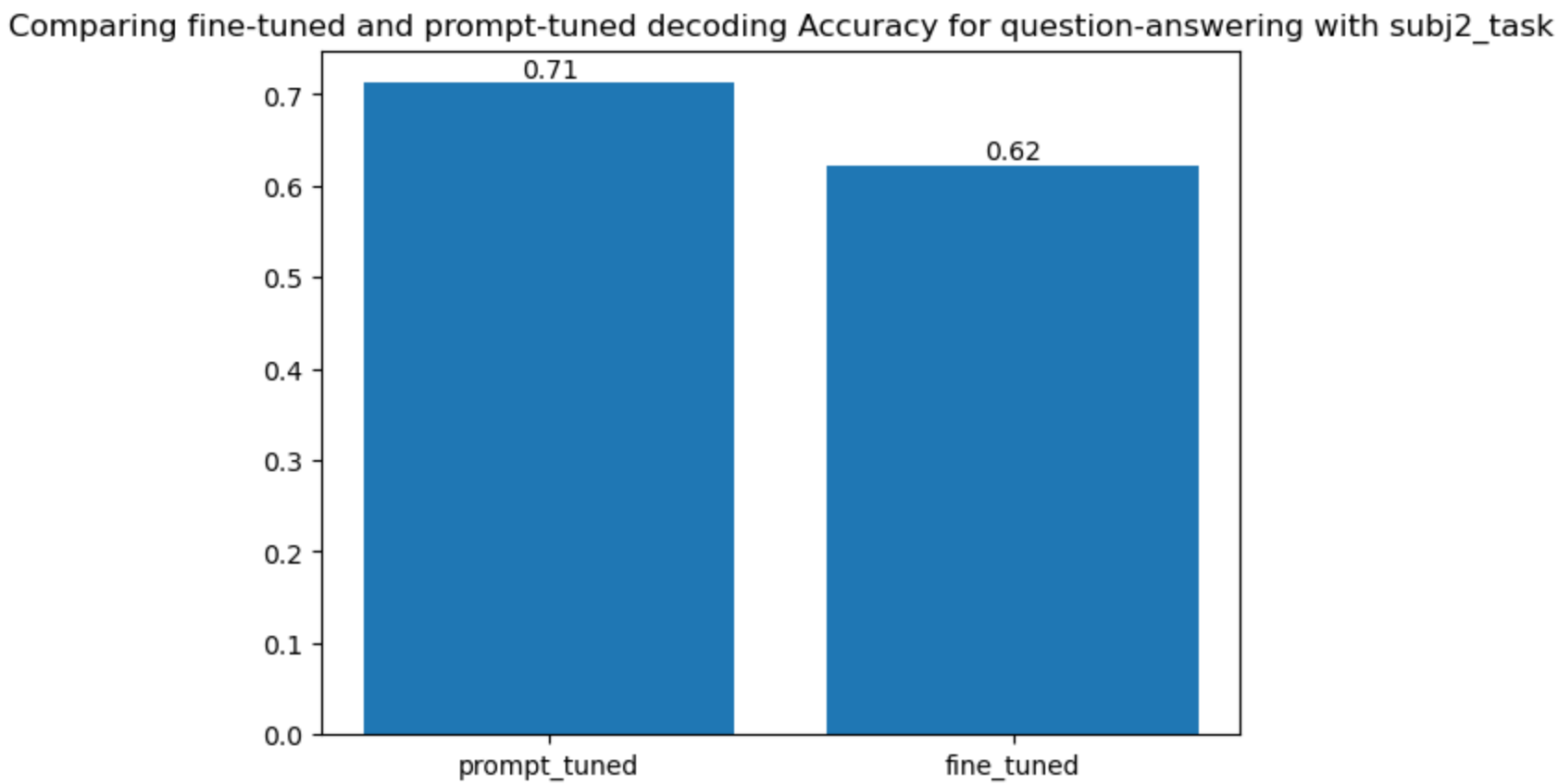

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj1_task
0.74
0.63
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

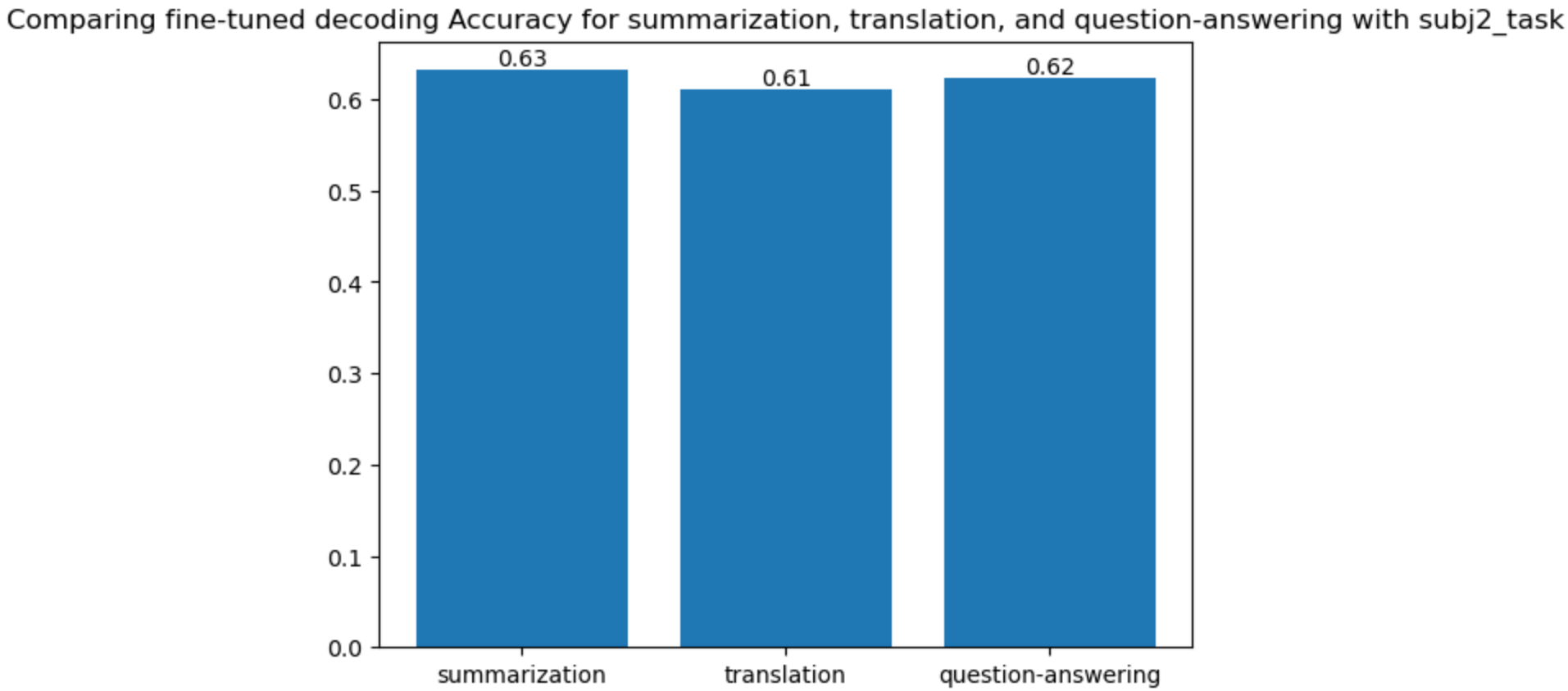

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_task
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.64
0.68
0.63
summarization
translation
question-answering

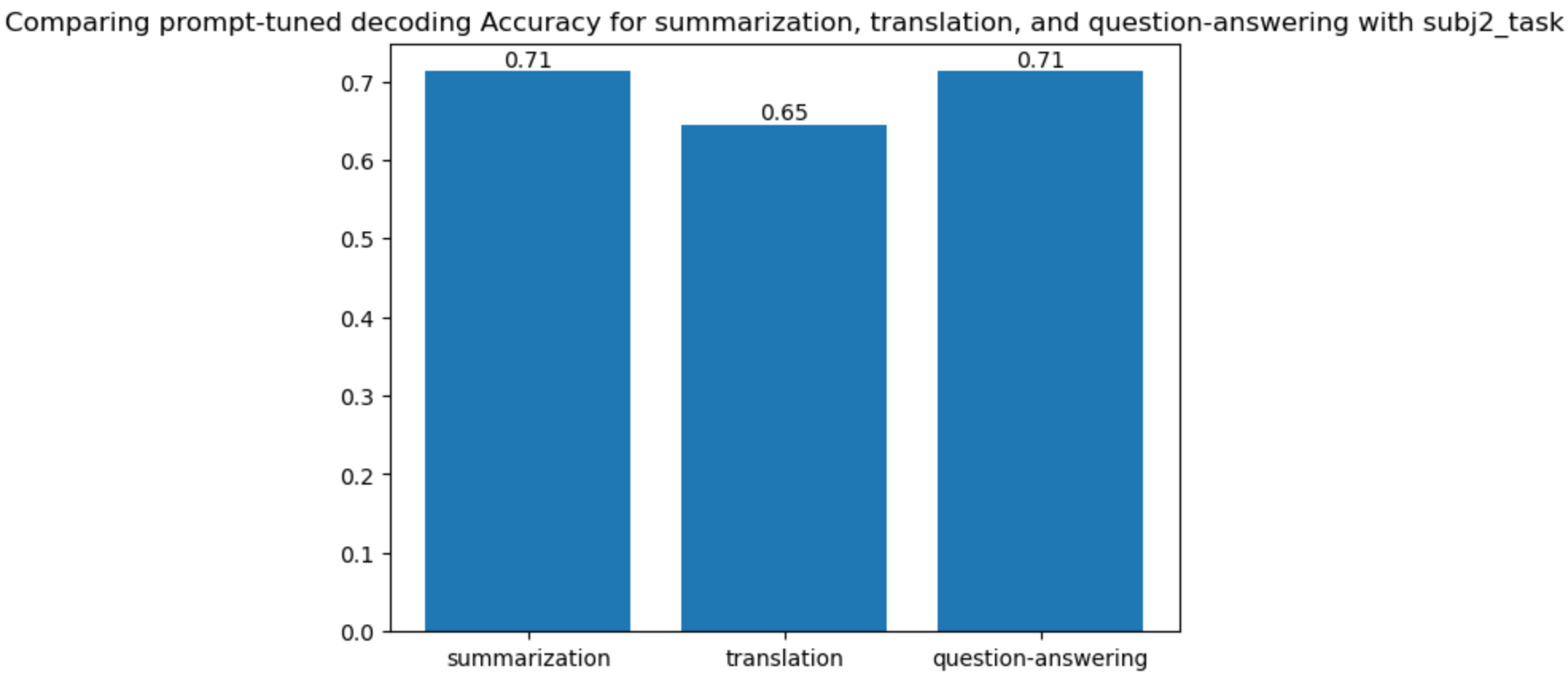

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_task
0.74
0.70
0.74
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj2_task
0.71
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj2_task
0.65
0.61
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

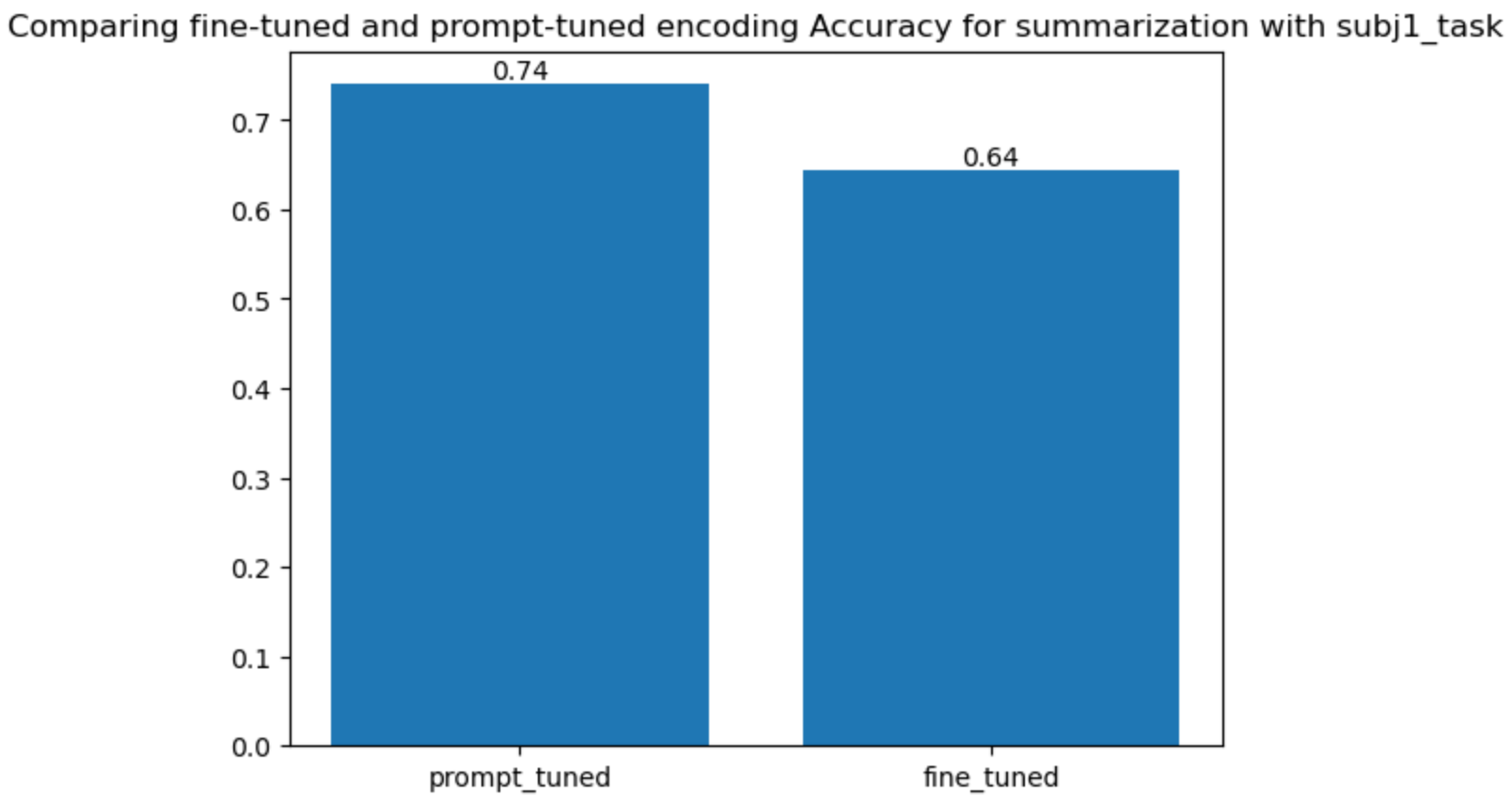

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj2_task
0.71
0.62
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

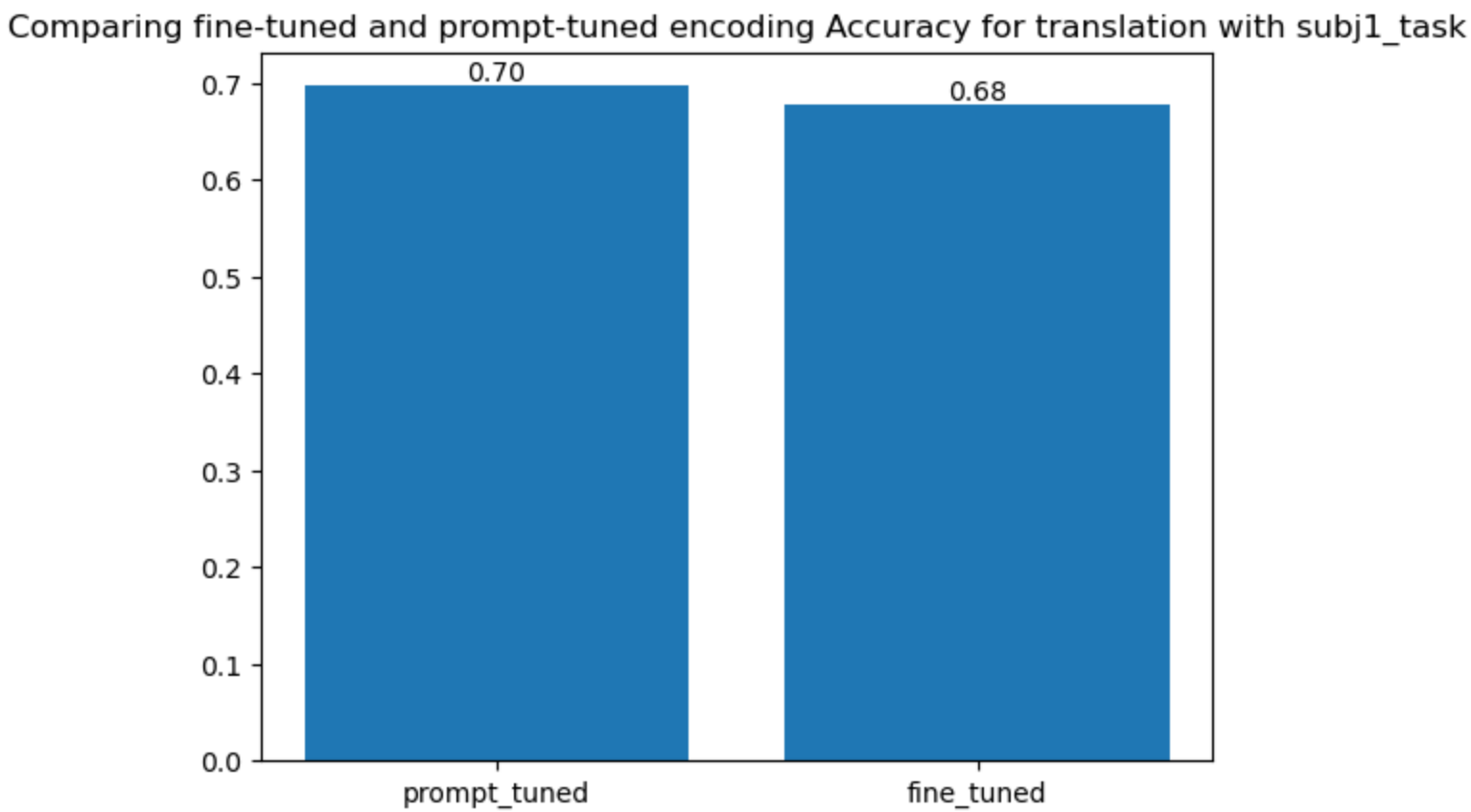

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_task
0.63
0.61
0.62
0.0
0.1
0.2
0.3
0.4
0.5
0.6
summarization
translation
question-answering

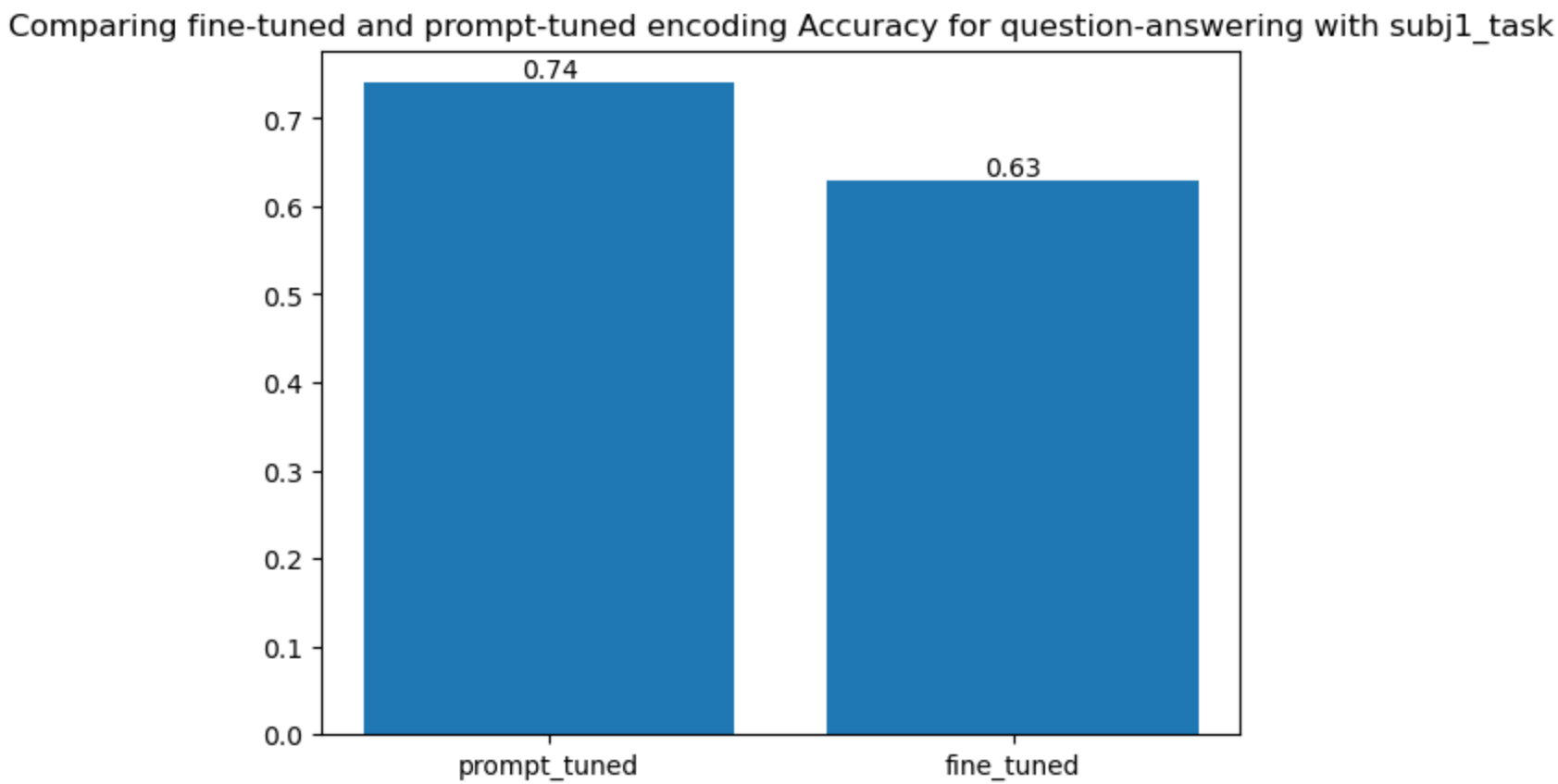

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_task
0.71
0.65
0.71
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

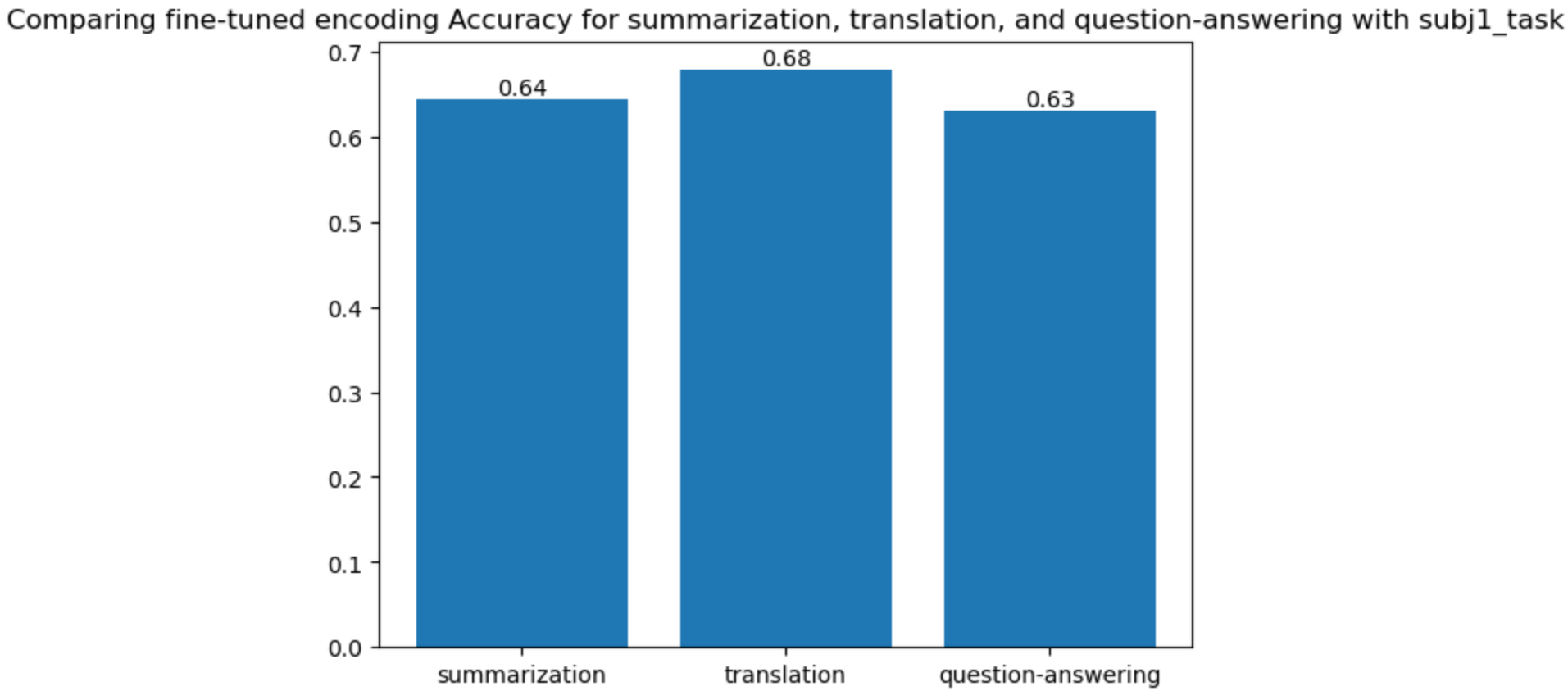
# Analysis of Prompt-tuned vs Fine-Tuned (Brain decoding)

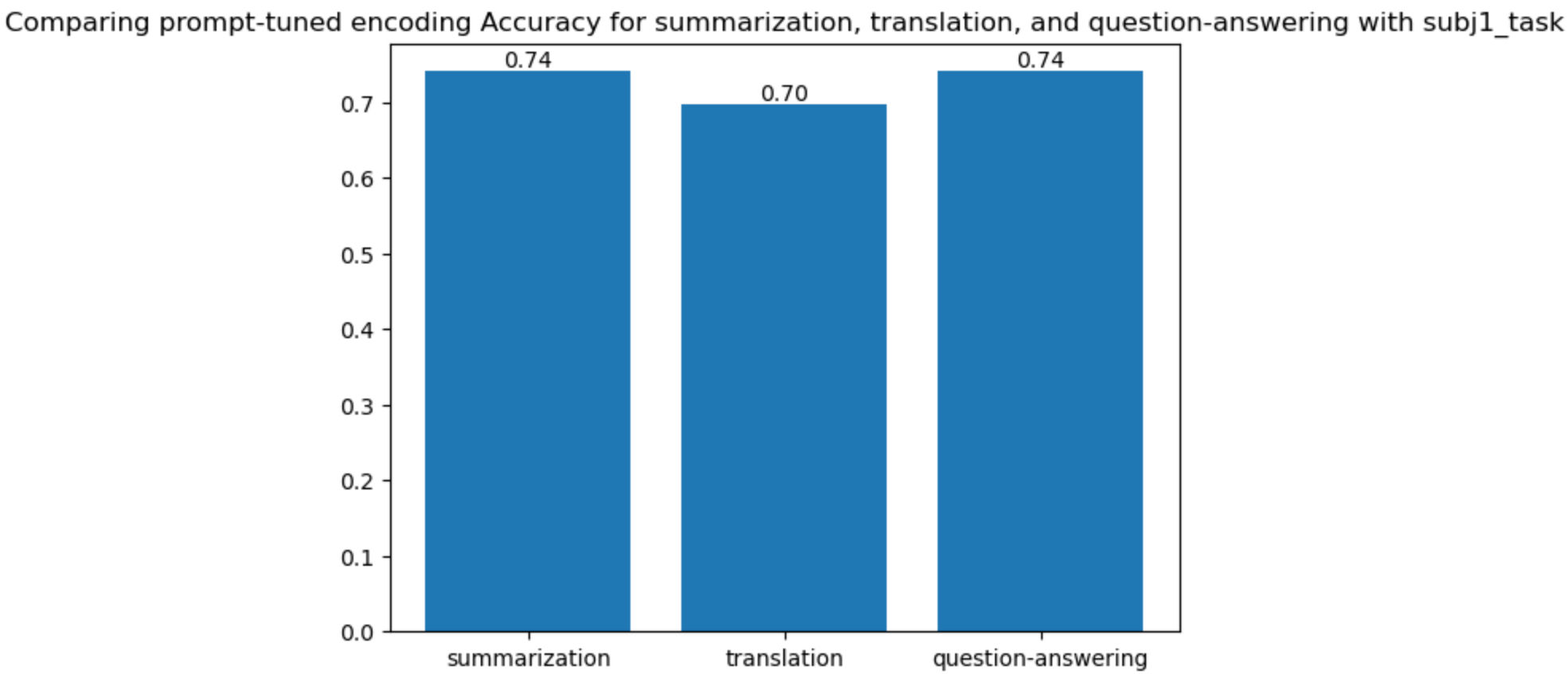
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain decoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed.
- Wheras in Fine-tuned models, MT model accounts better for brain representations than the other 2 tasks.
- Given it is the Task ROI which specializes in spacial information and attention, a direct correlation is tough to bring out.
- Also the results we got are **consistent** across subjects!

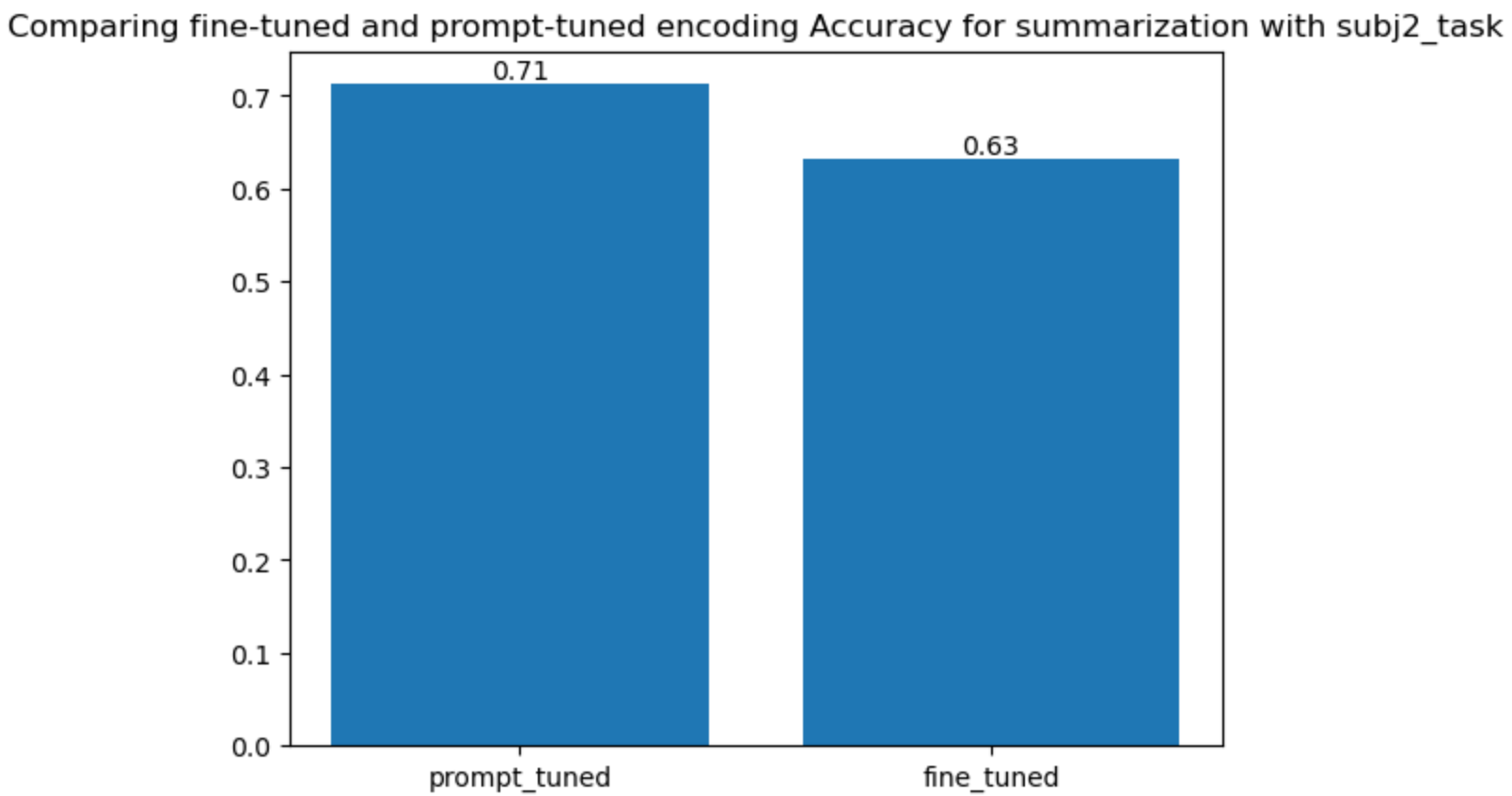

Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj1_task
0.74
0.64
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

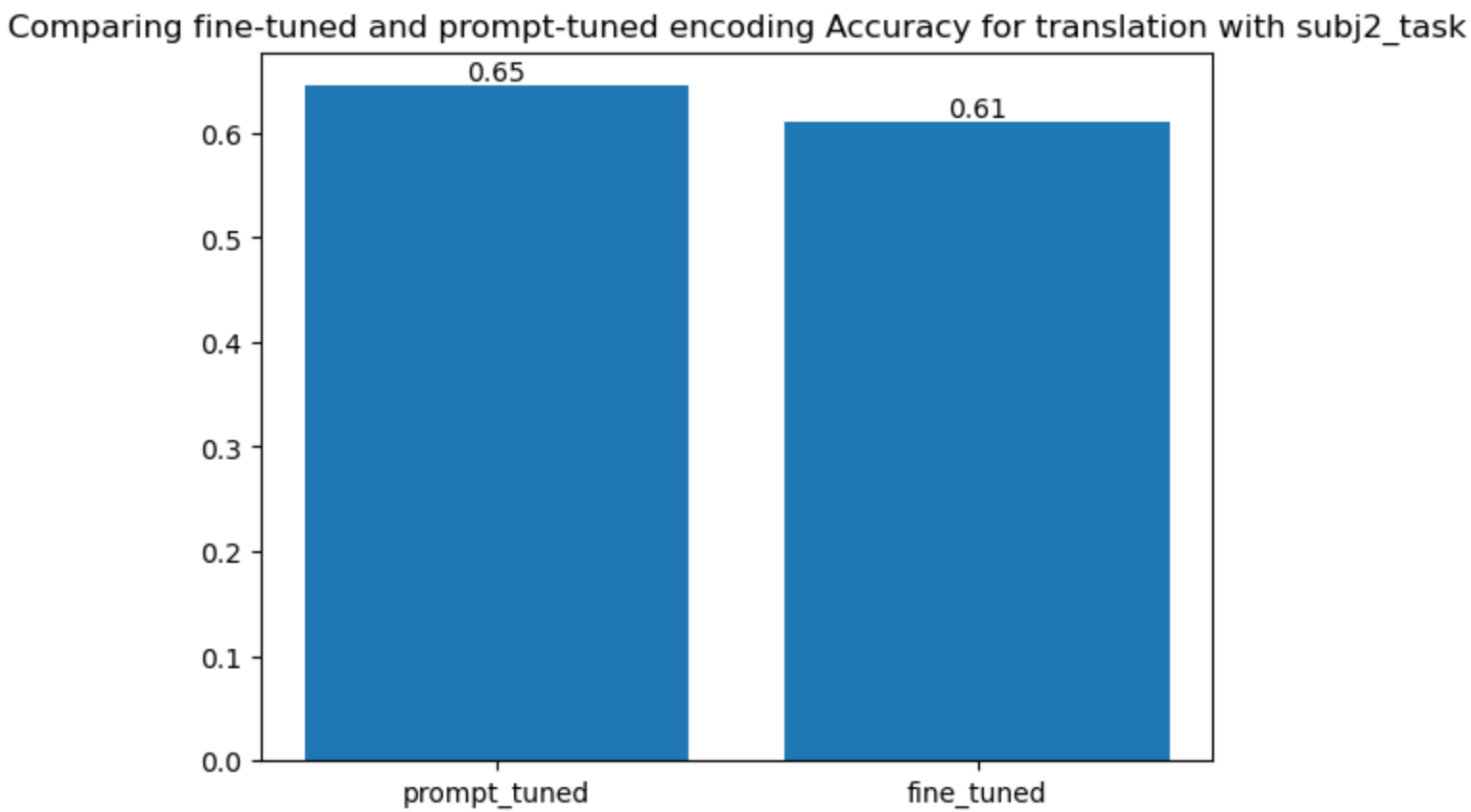

Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj1_task
0.70
0.68
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

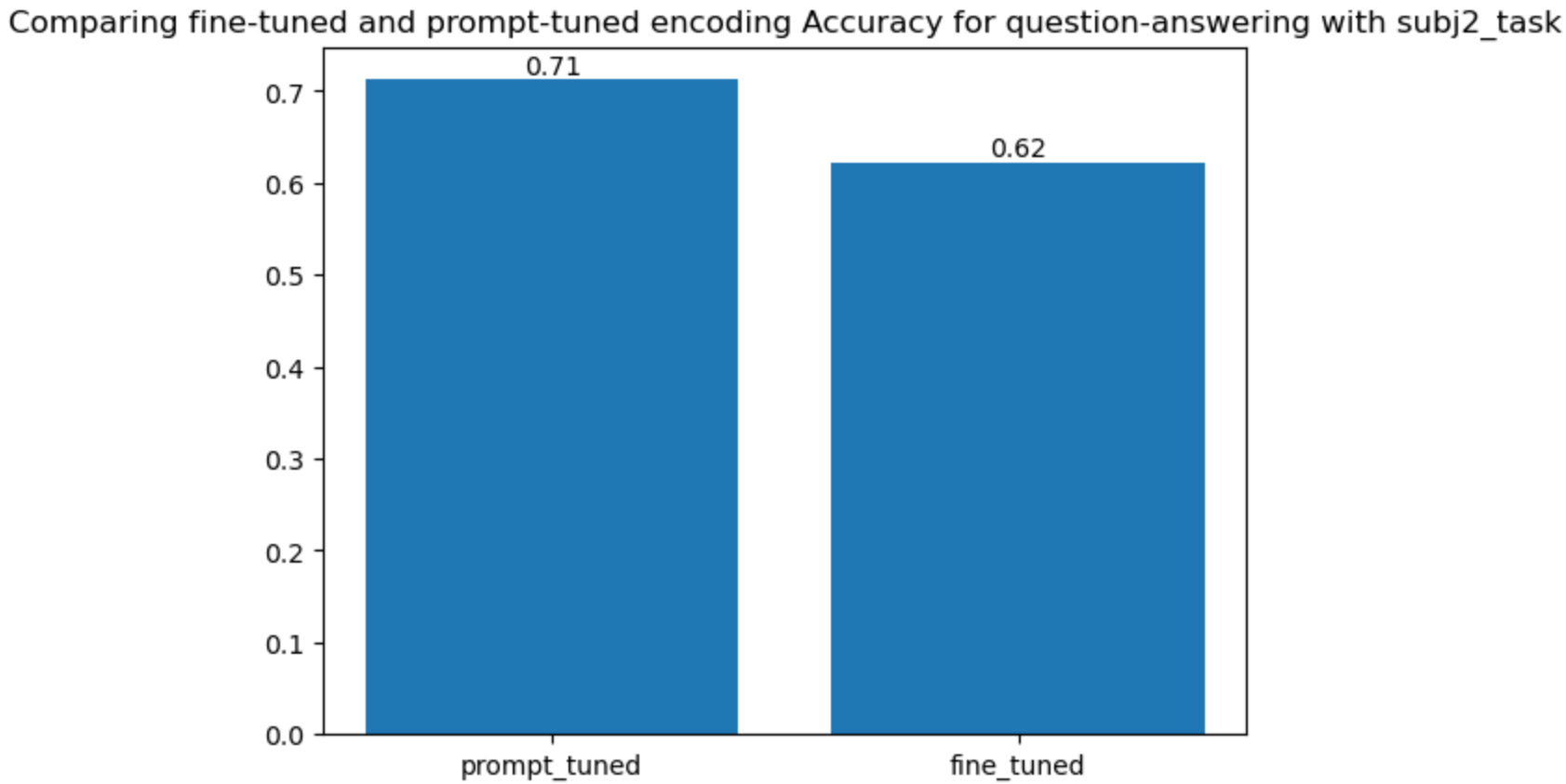

Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj1_task
0.74
0.63
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

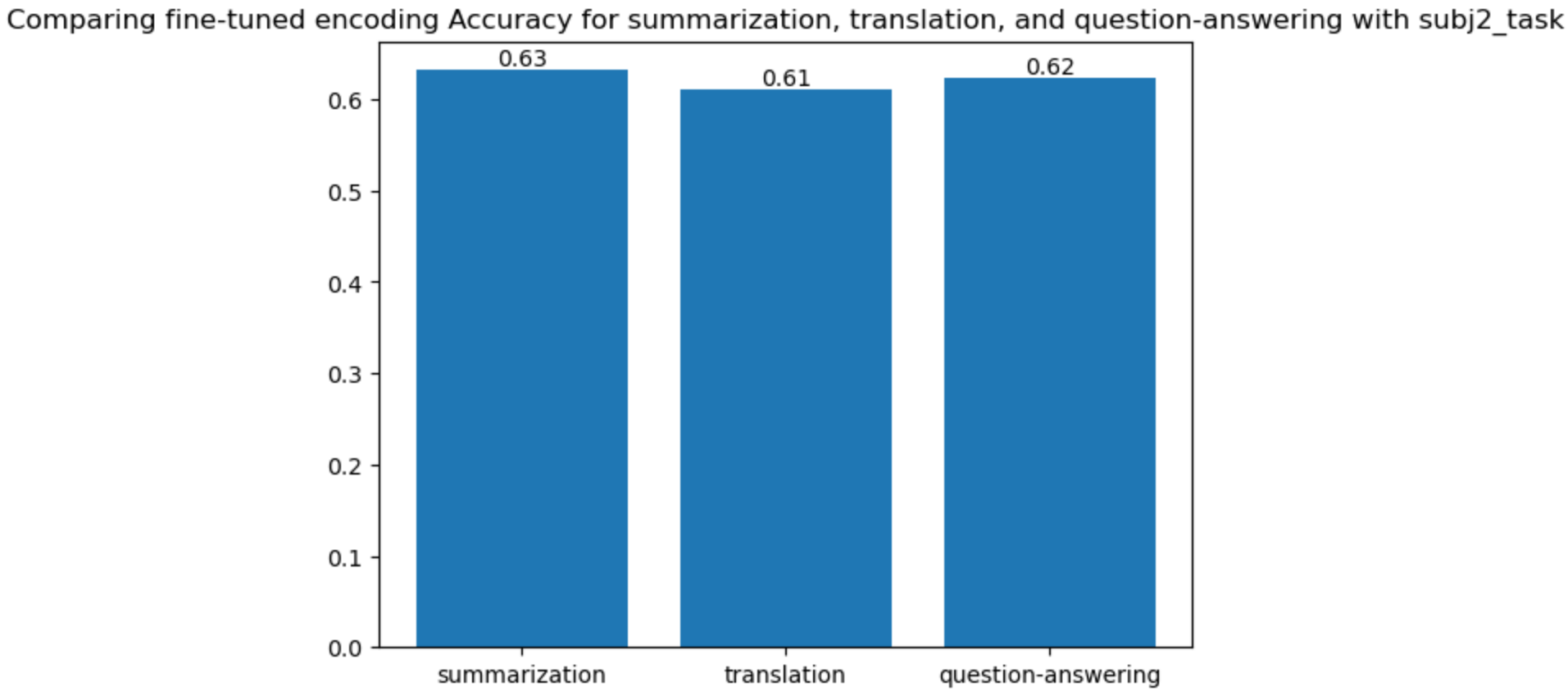

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_task
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.64
0.68
0.63
summarization
translation
question-answering

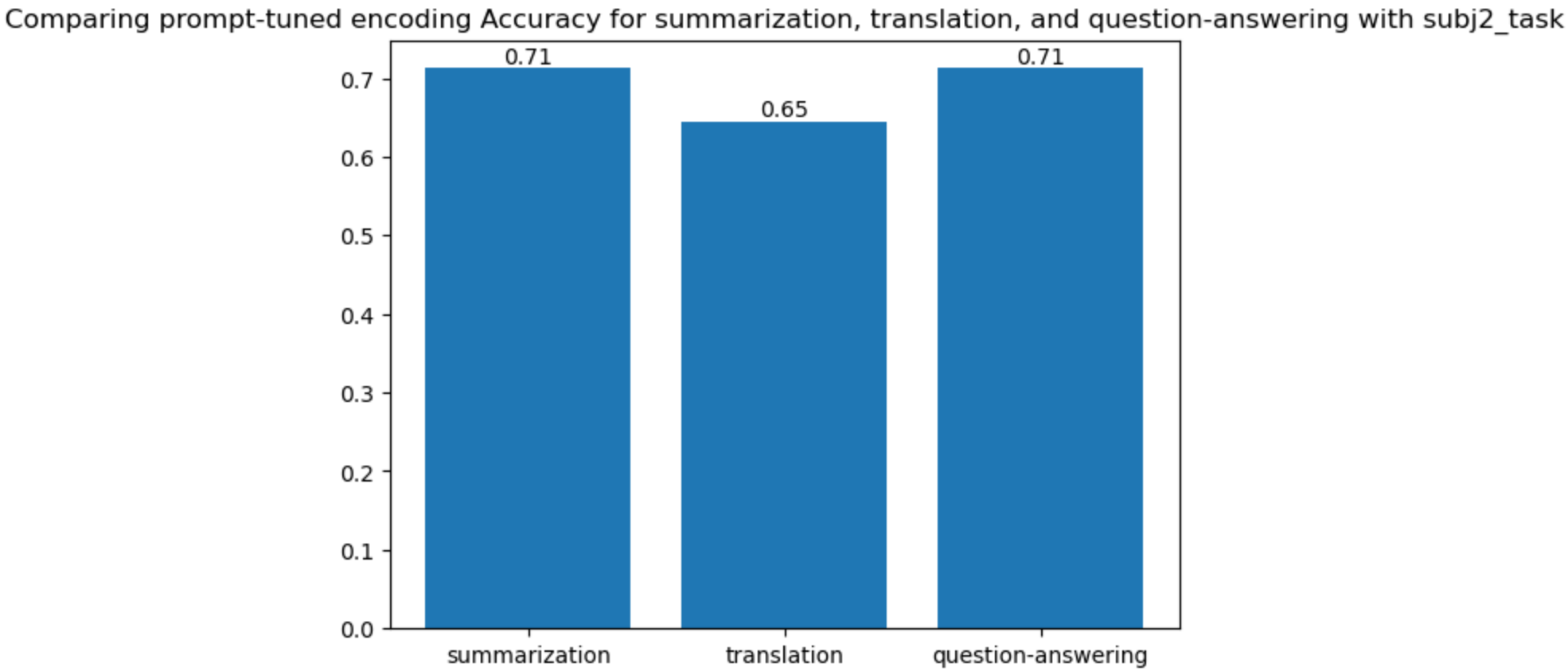

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_task
0.74
0.70
0.74
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj2_task
0.71
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj2_task
0.65
0.61
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj2_task
0.71
0.62
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

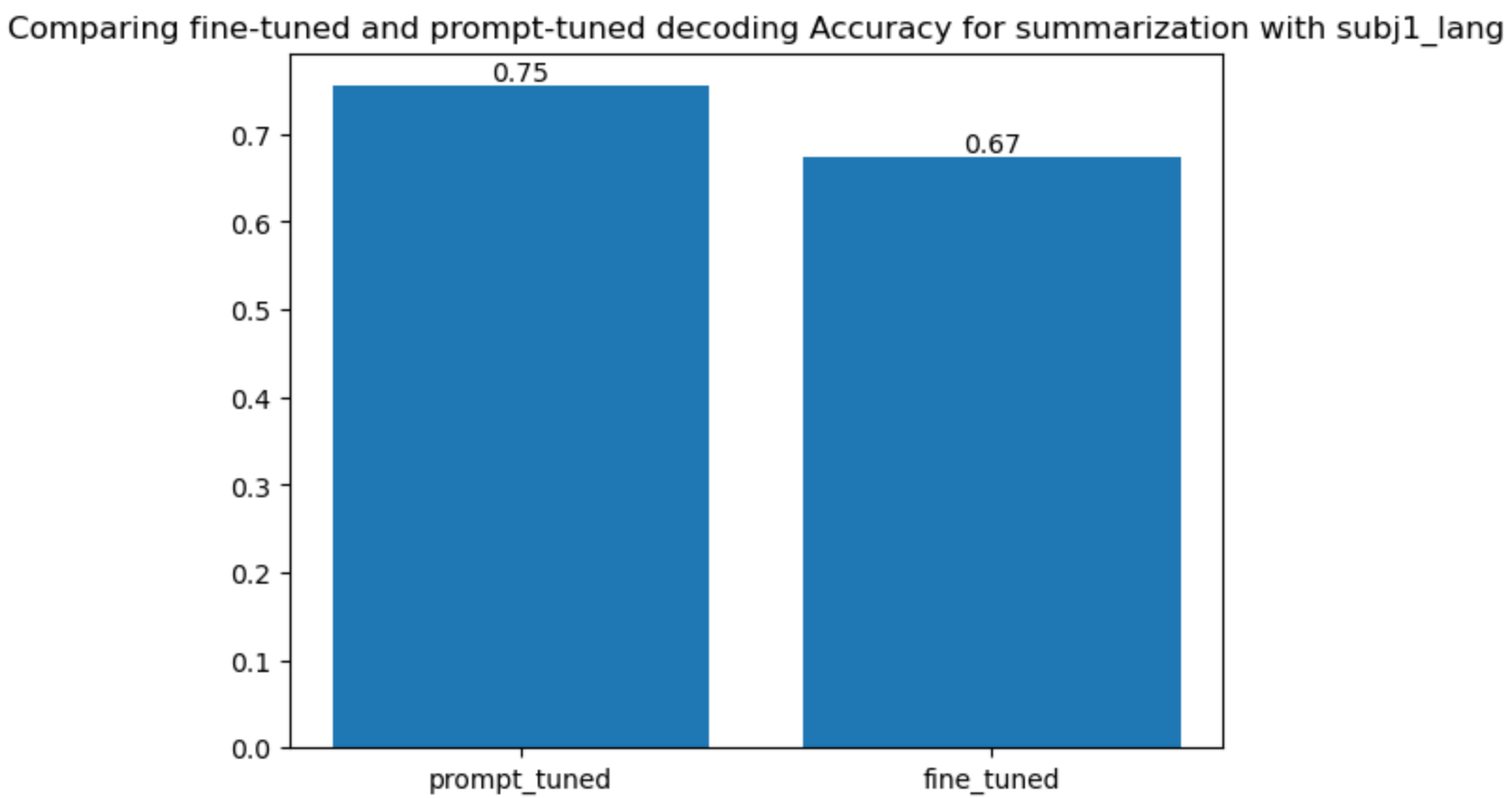

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_task
0.63
0.61
0.62
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

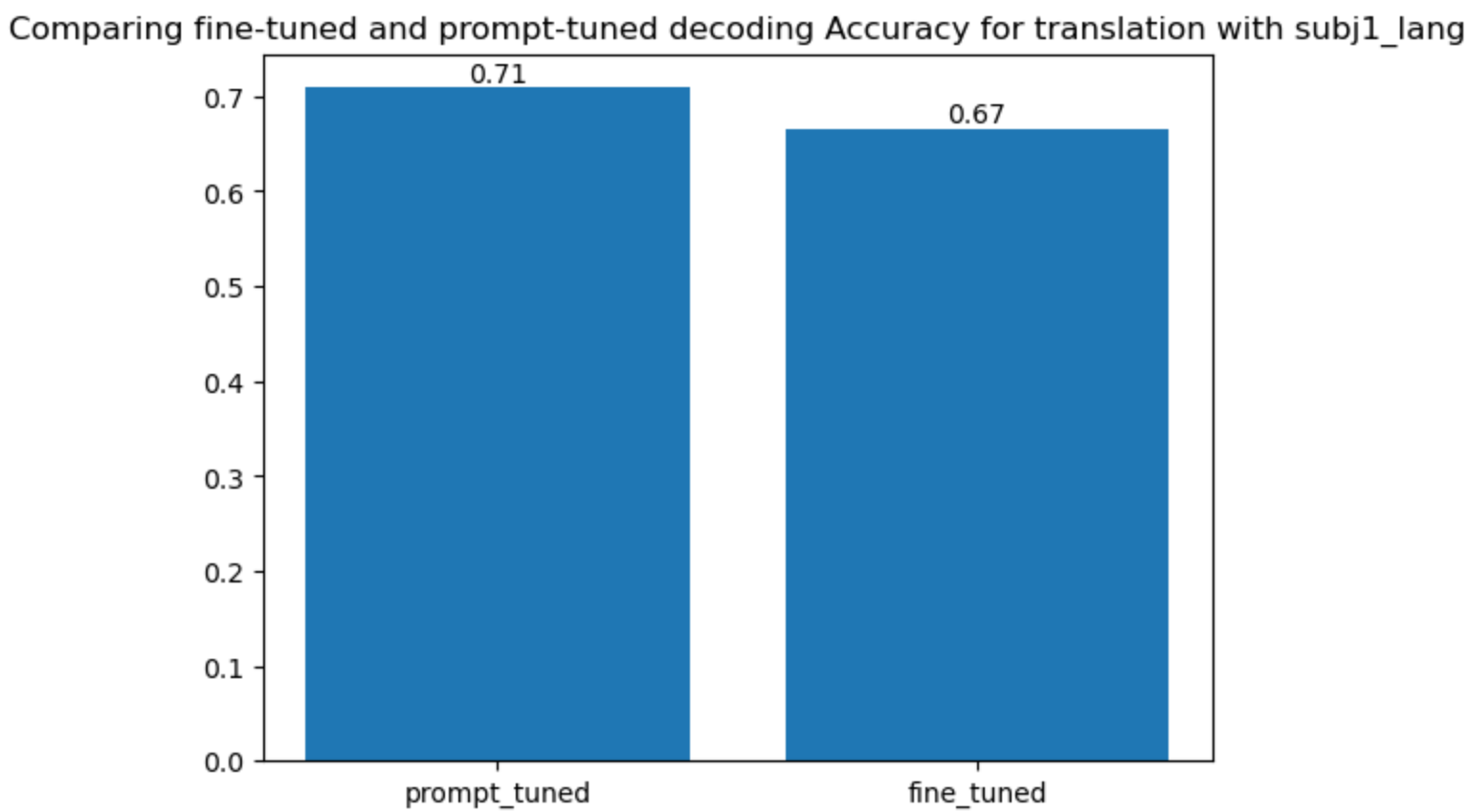

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_task
0.71
0.65
0.71
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

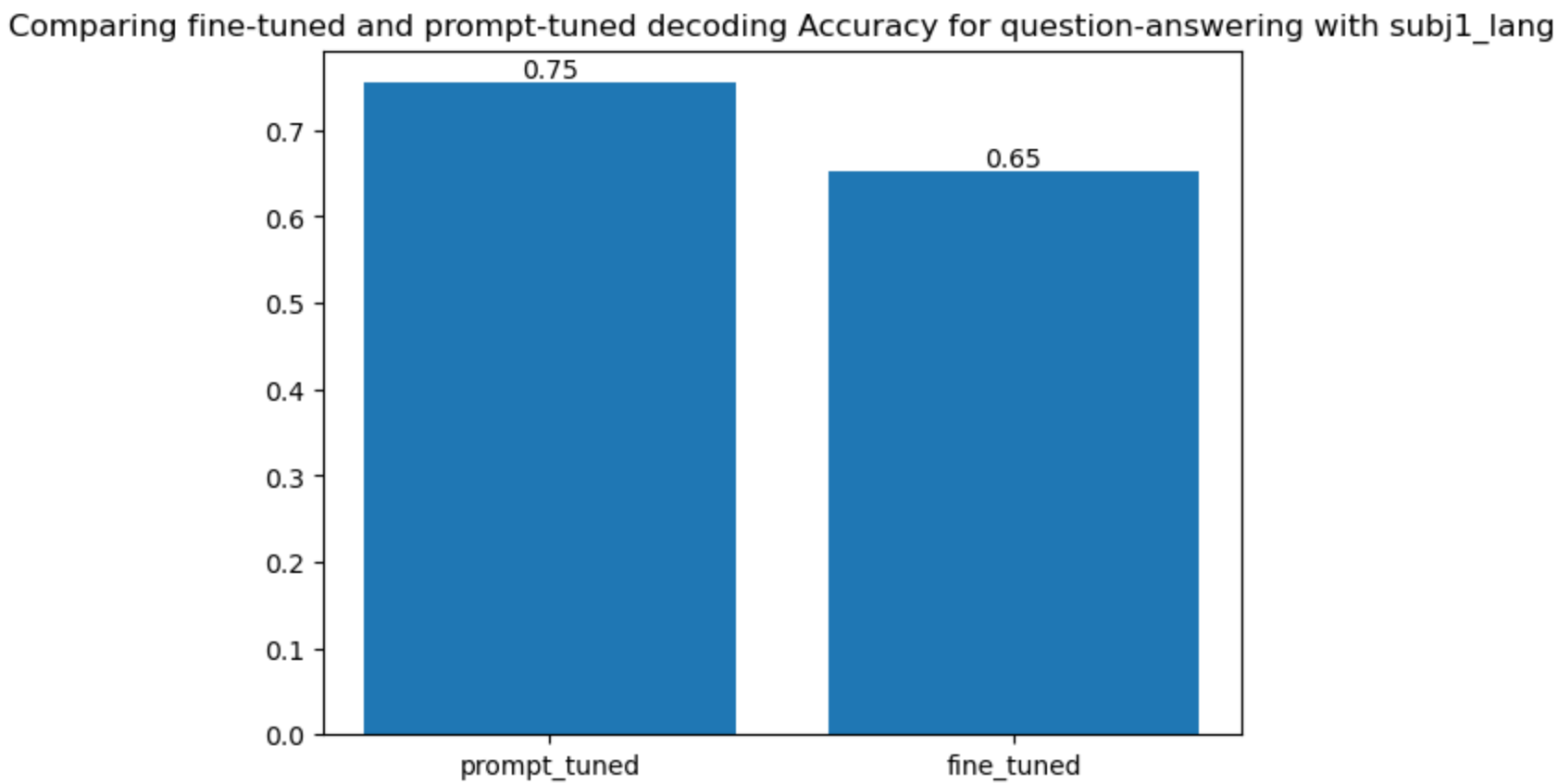
# Analysis of Prompt-tuned vs Fine-Tuned (Brain encoding)

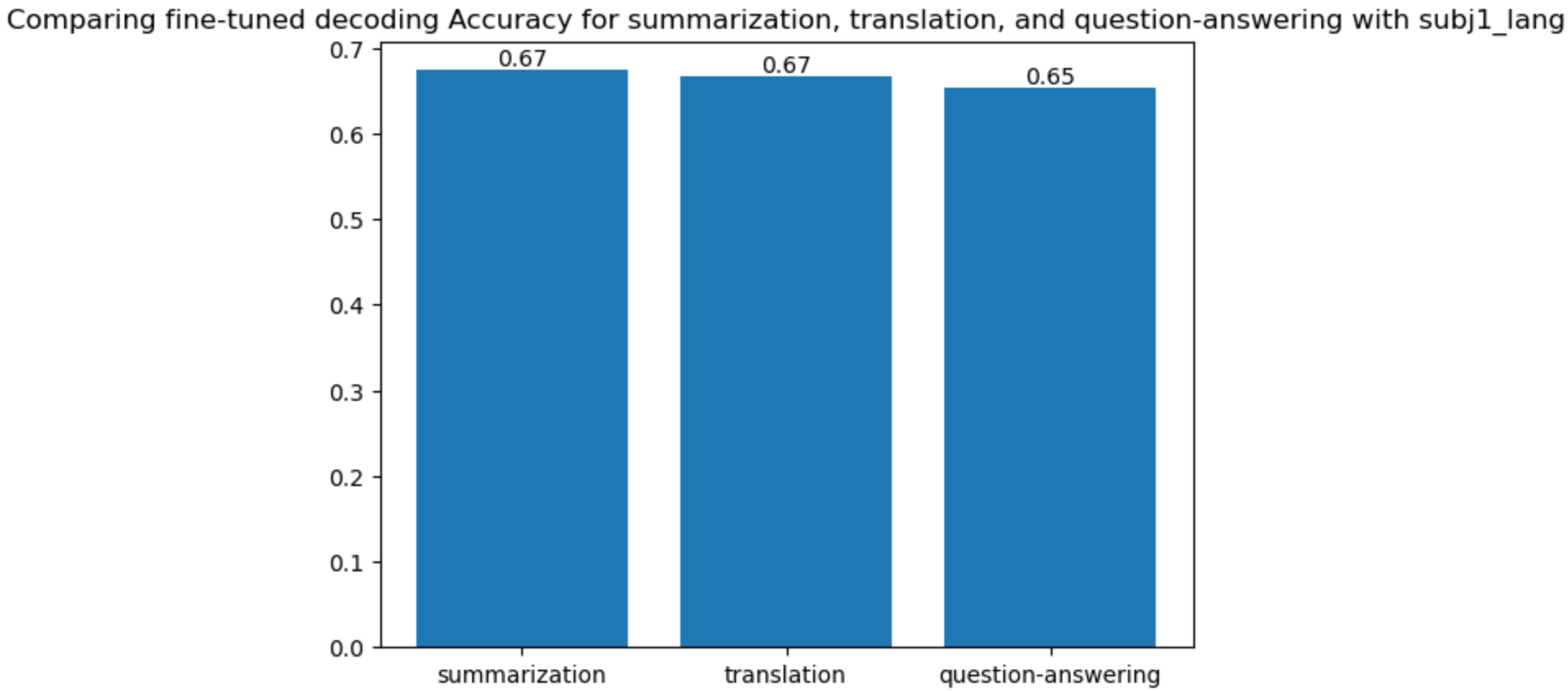
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain encoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed, this is consistent across subjects as well
- Wheras in Fine-tuned models, almost every model performs equally well but the MT model outperforms the other 2
- Also, the results overall are consistent across subjects.

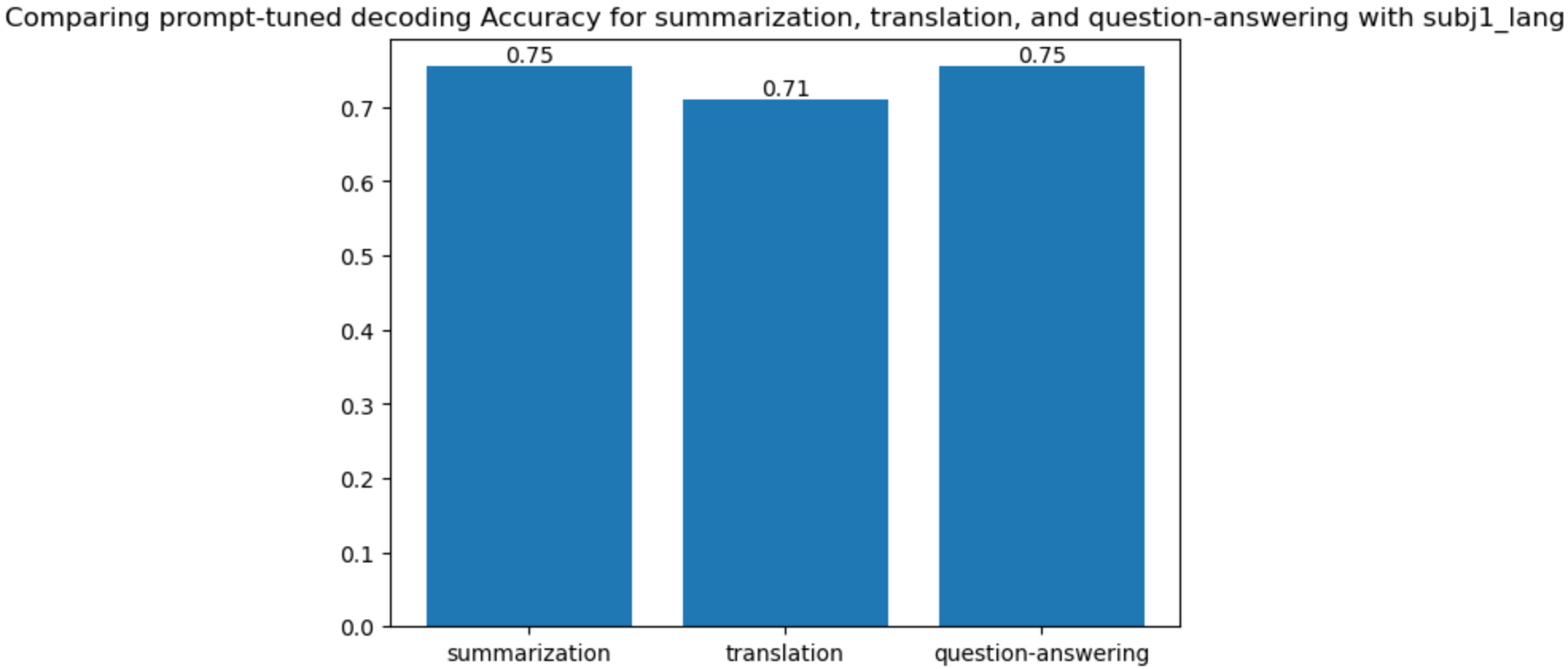
# LANG ROI

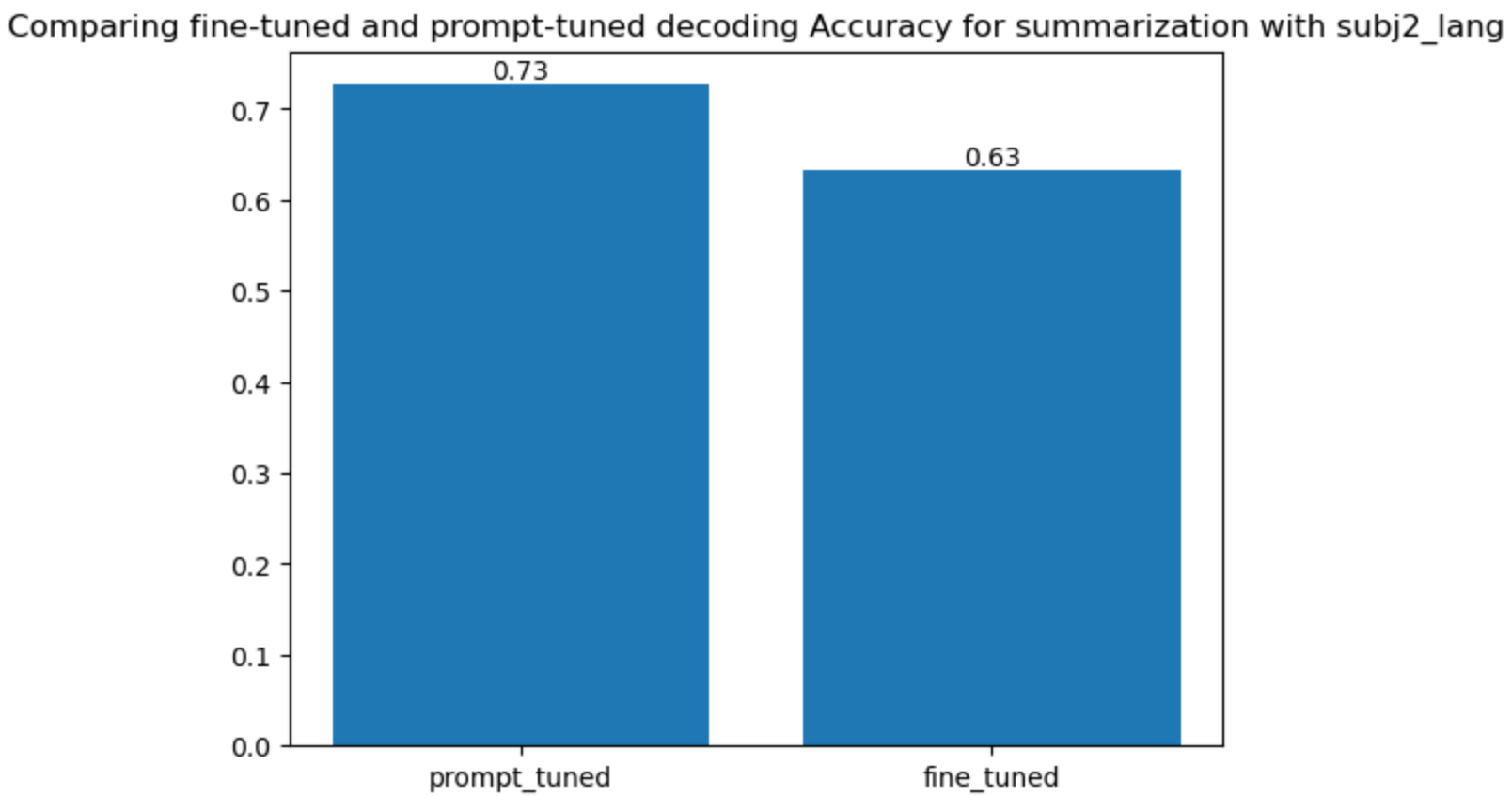

Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj1_lang
0.75
0.67
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

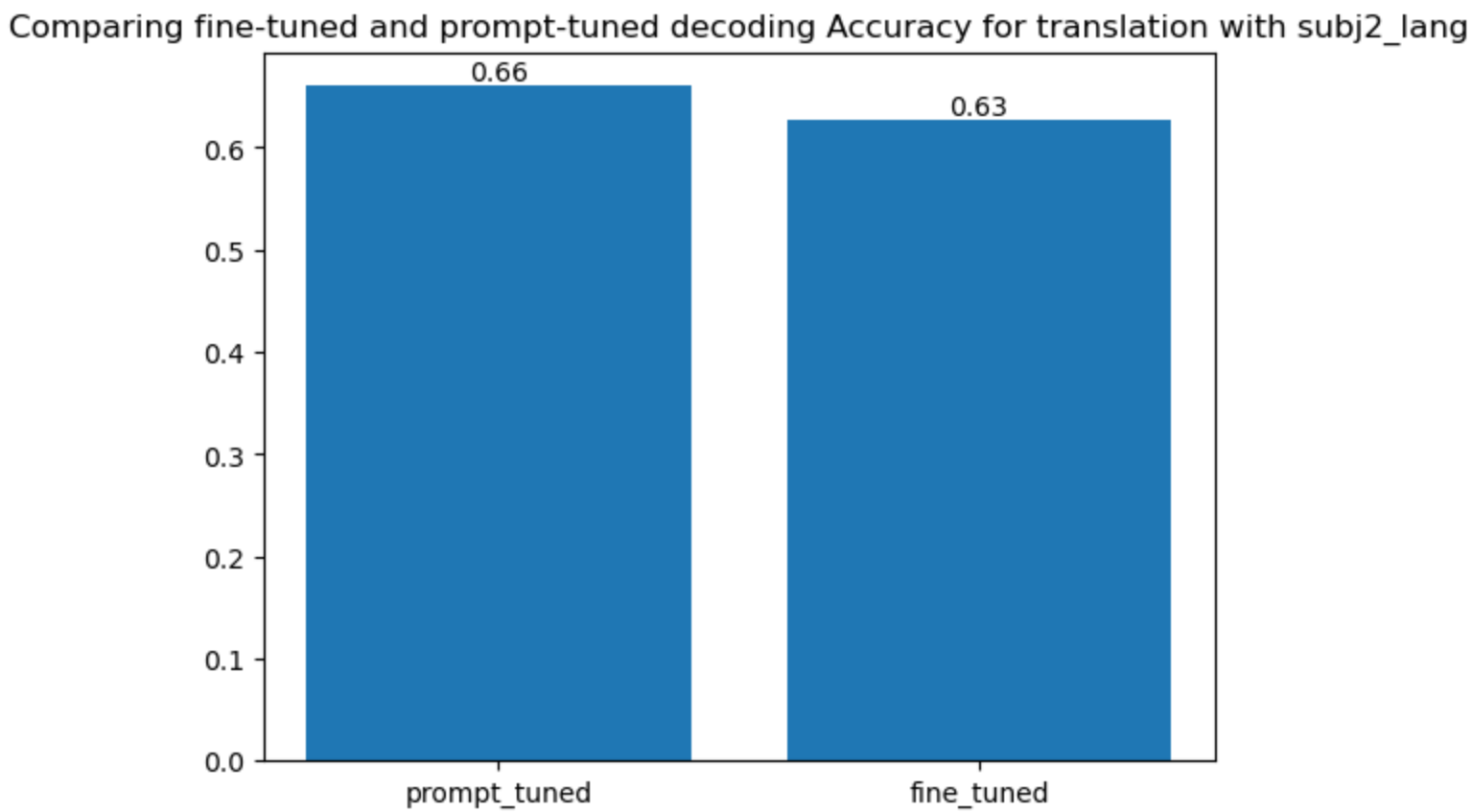

Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj1_lang
0.71
0.67
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

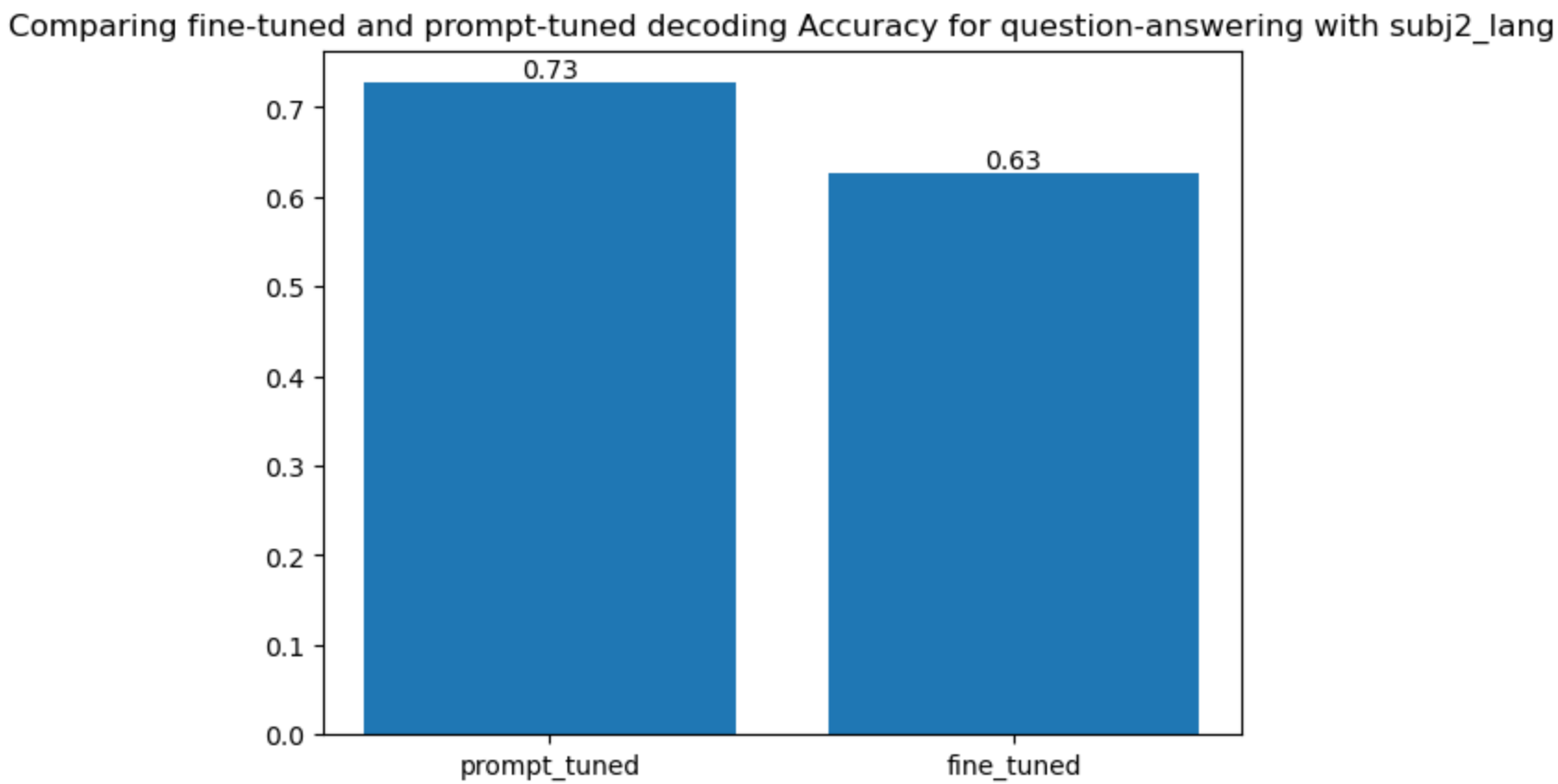

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj1_lang
0.75
0.65
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

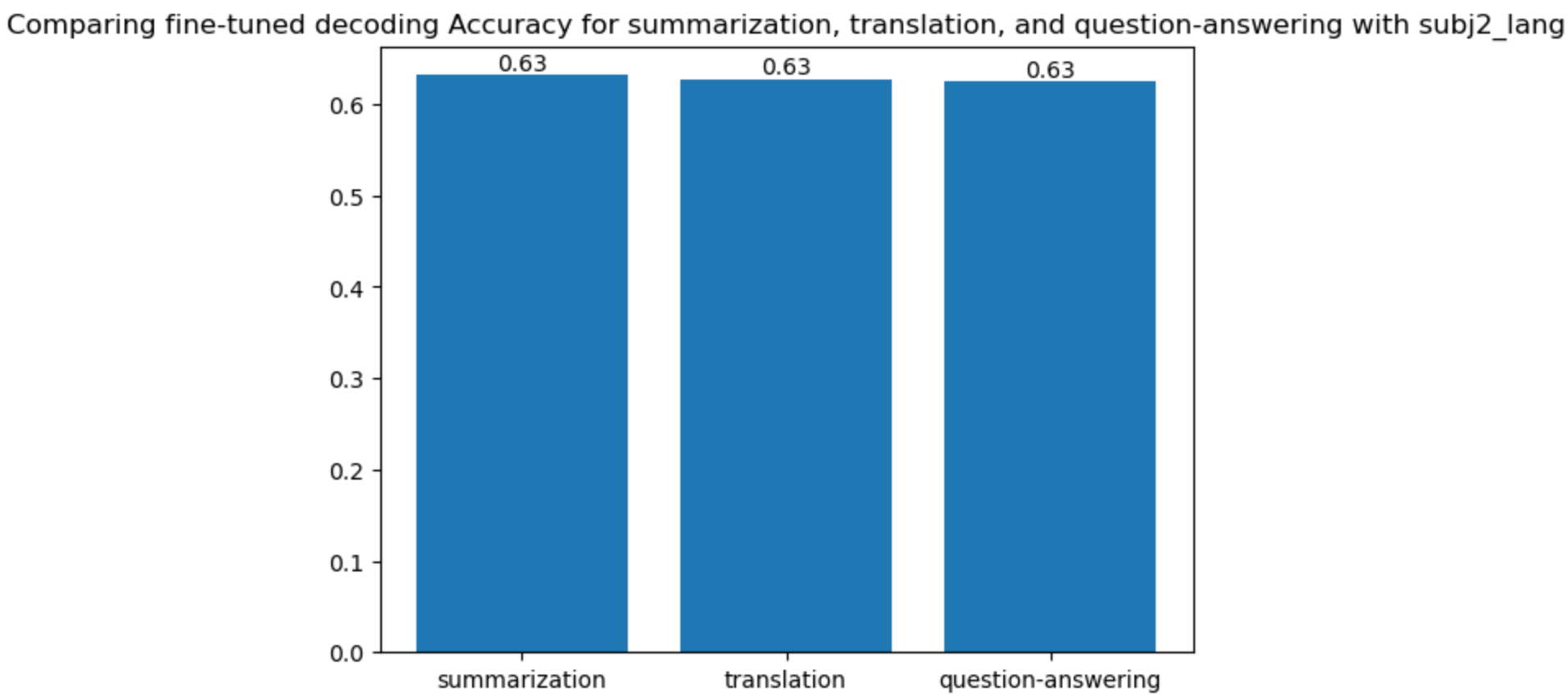

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_lang
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.67
0.67
0.65
summarization
translation
question-answering

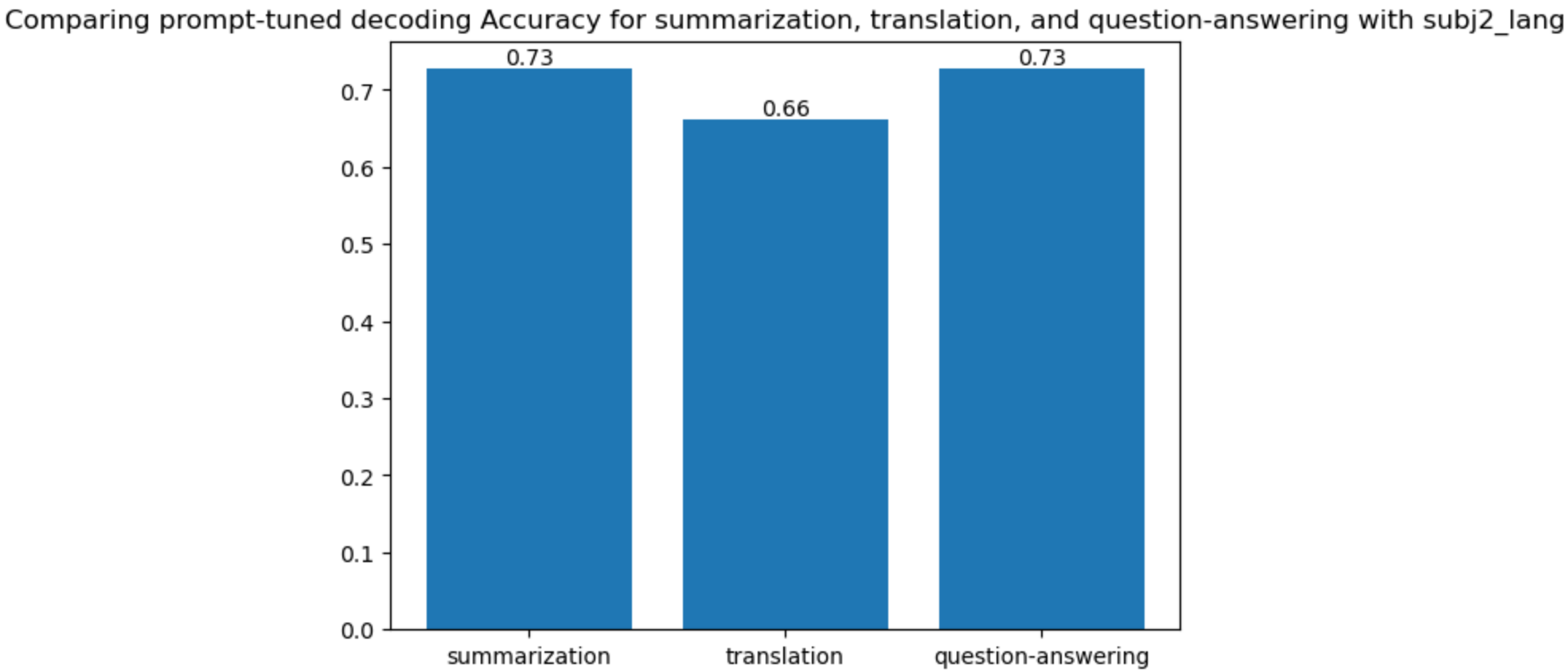

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_lang
0.75
0.71
0.75
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj2_lang
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj2_lang
0.66
0.63
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

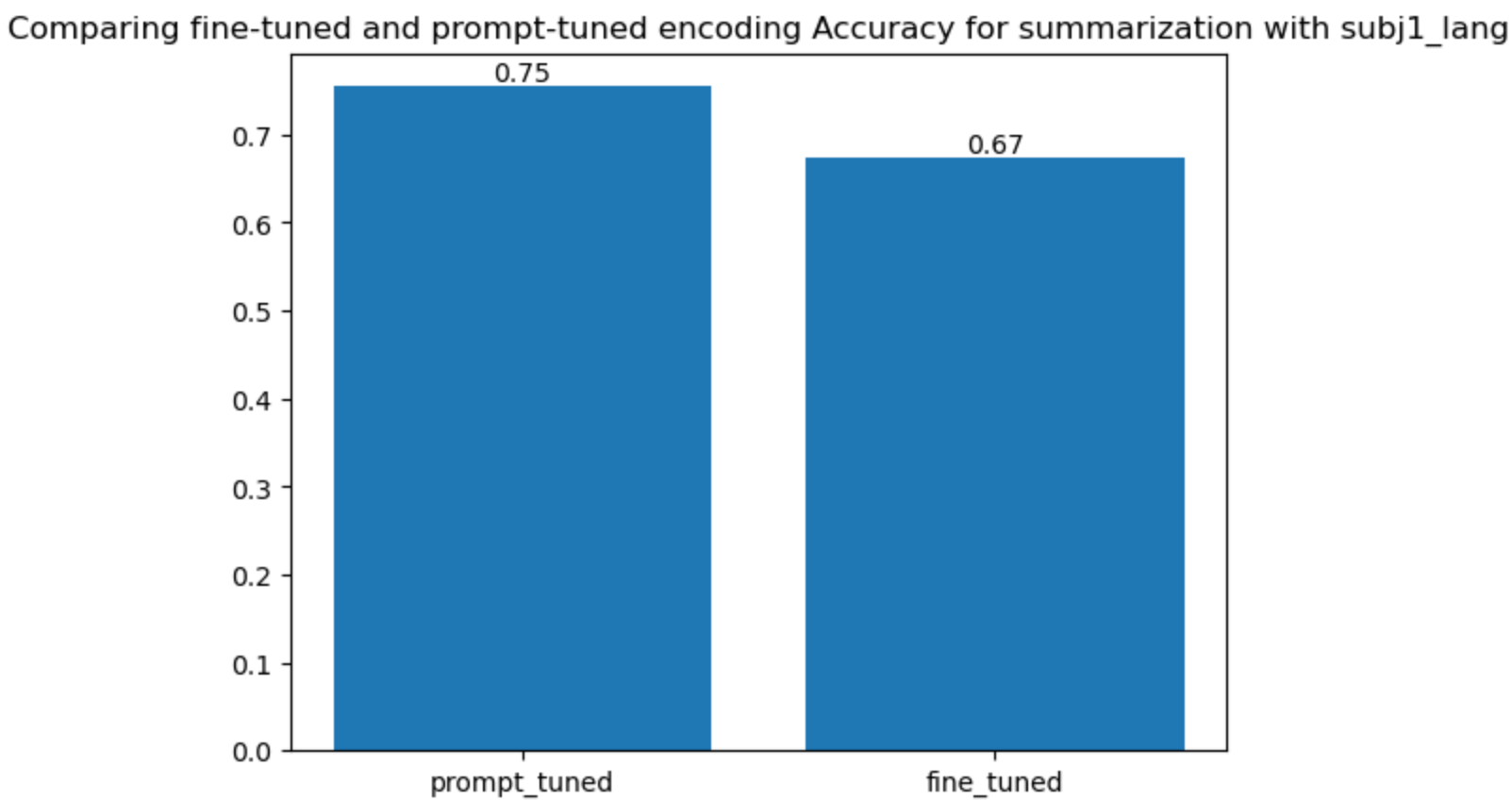

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj2_lang
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

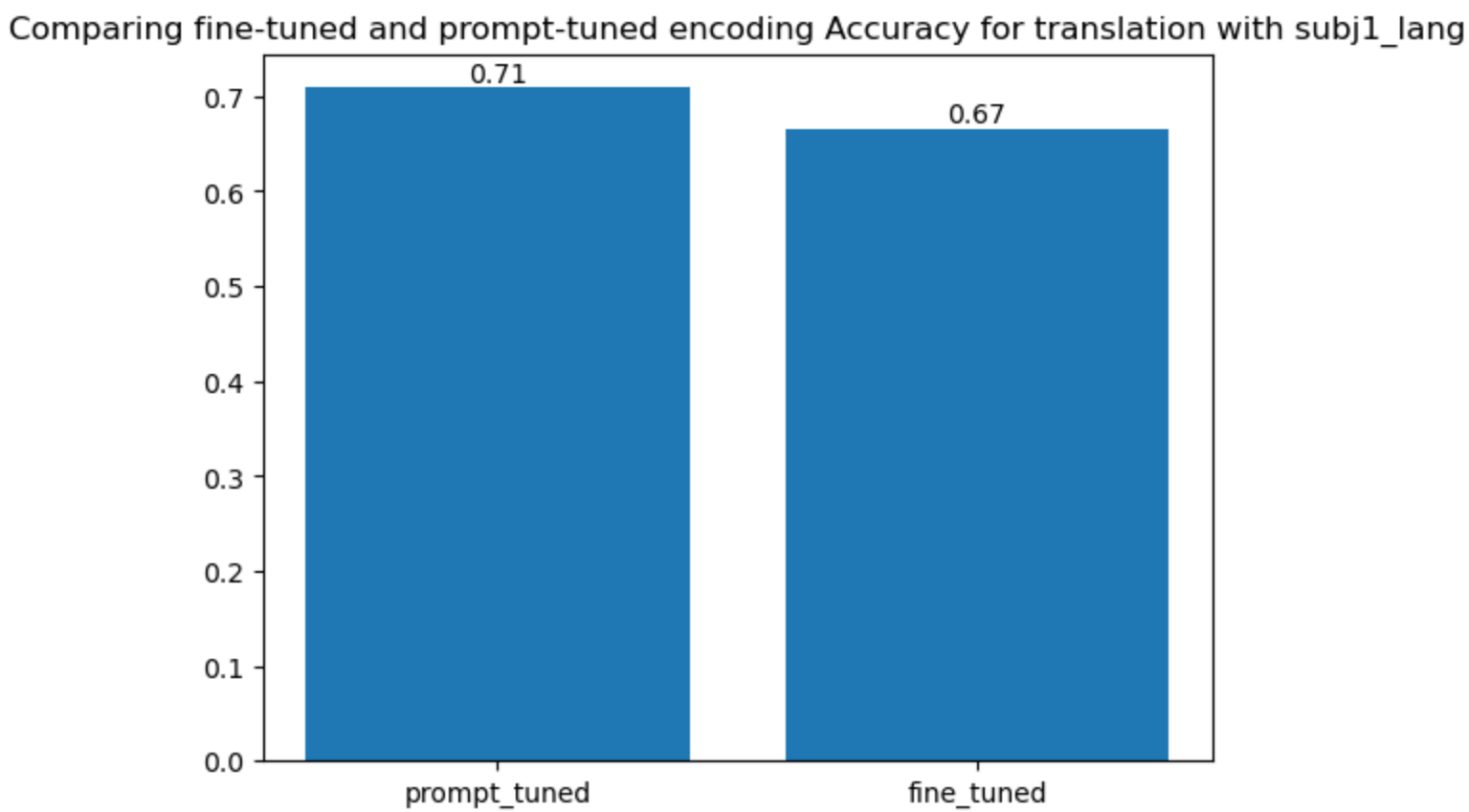

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_lang
0.63
0.63
0.63
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

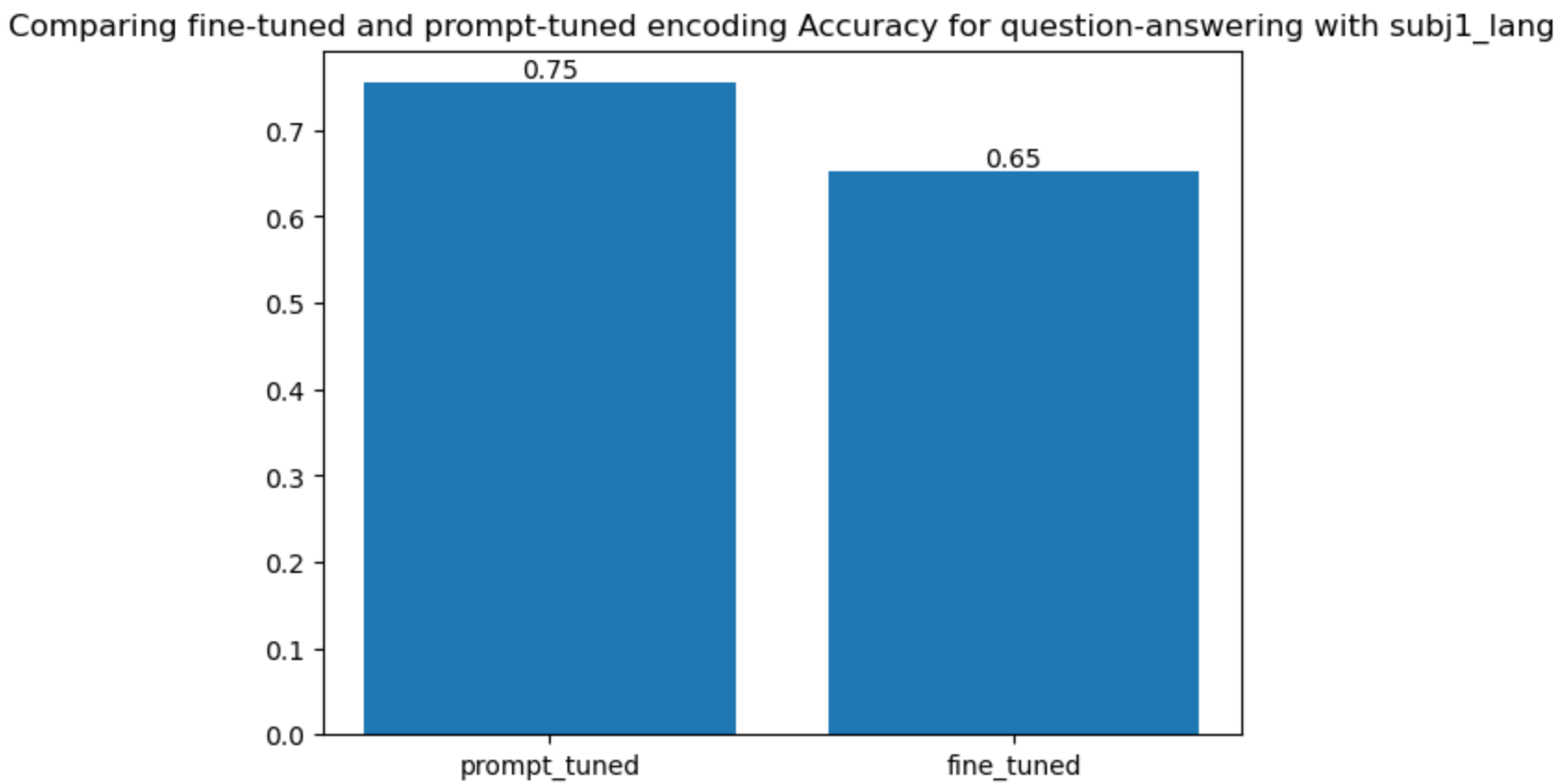

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_lang
0.73
0.66
0.73
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering

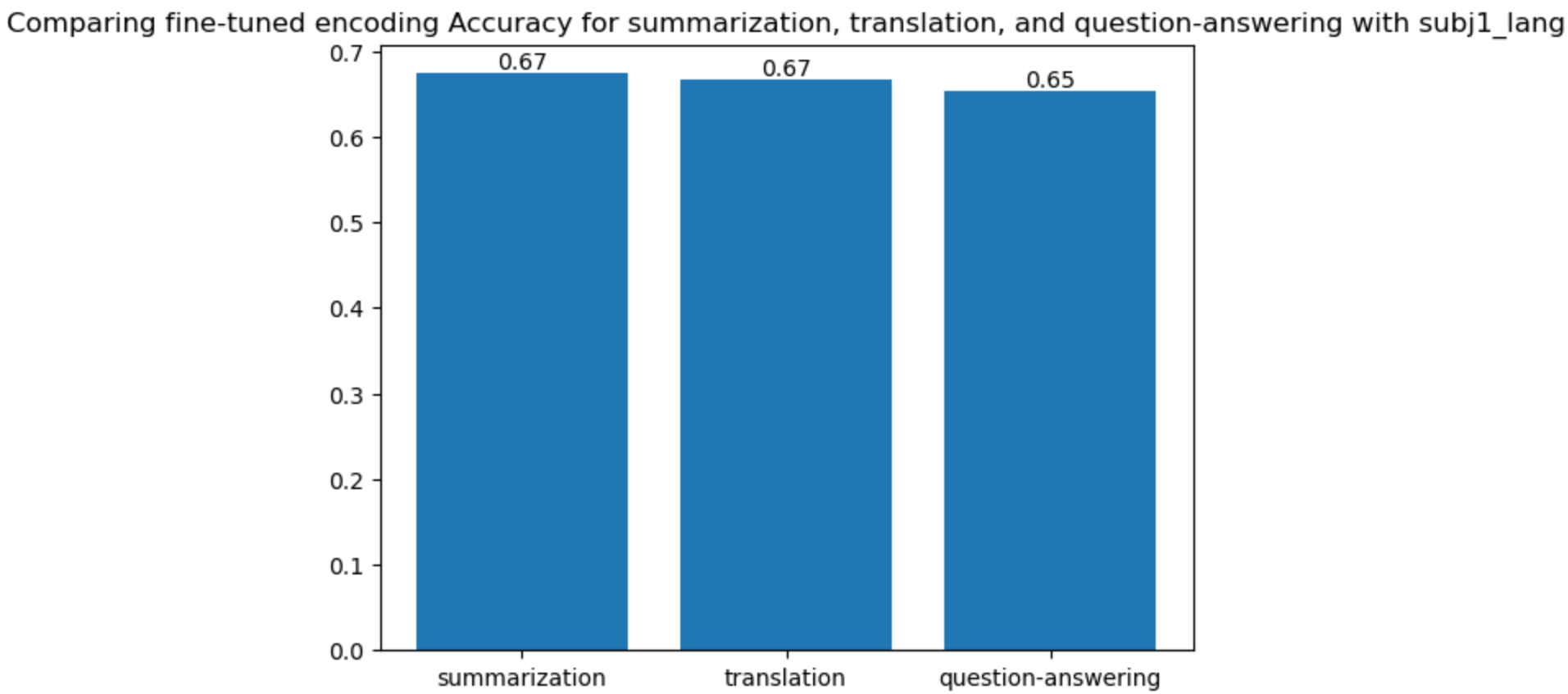
# Analysis of Prompt-tuned vs Fine-Tuned (Brain decoding)

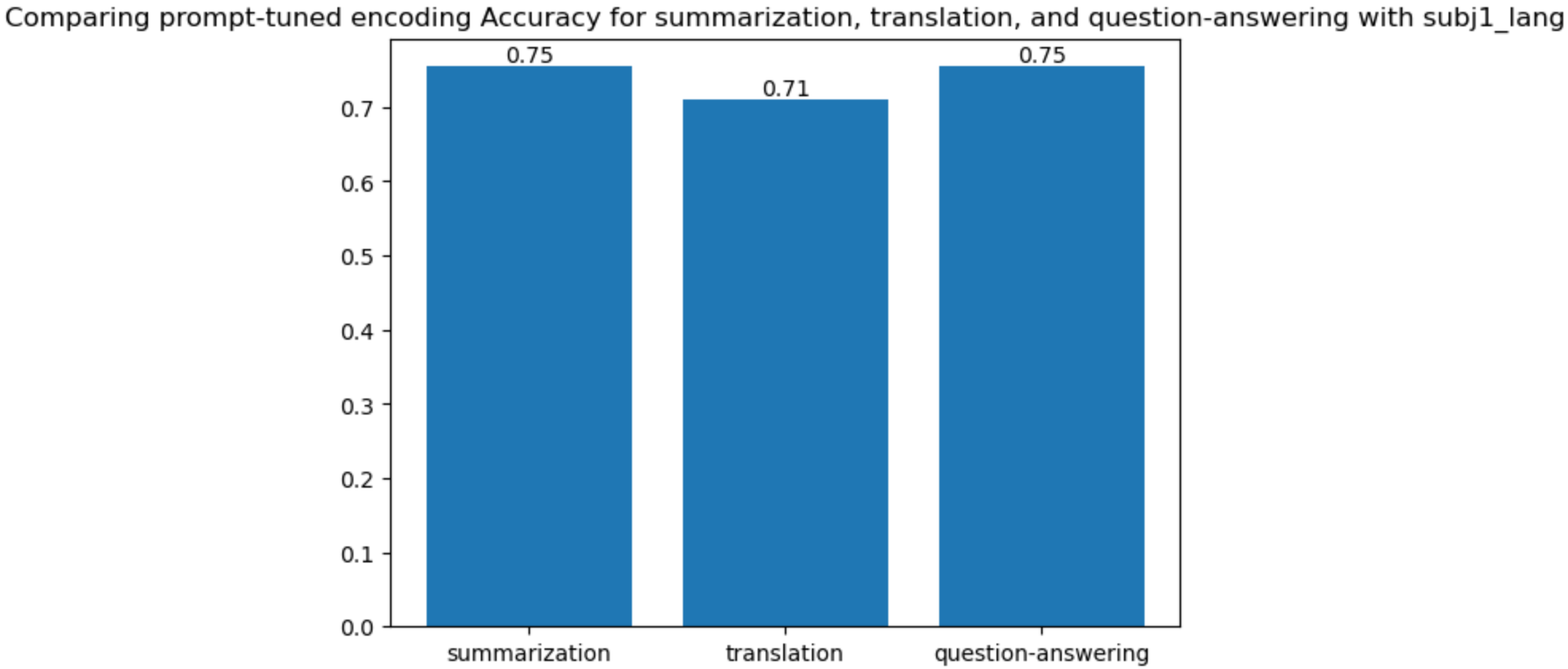
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain decoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed.
- Wheras in Fine-tuned models, all perform at an equal level, with no particular trend of one outperforming the other.
- Also the results we got are **consistent** across subjects!

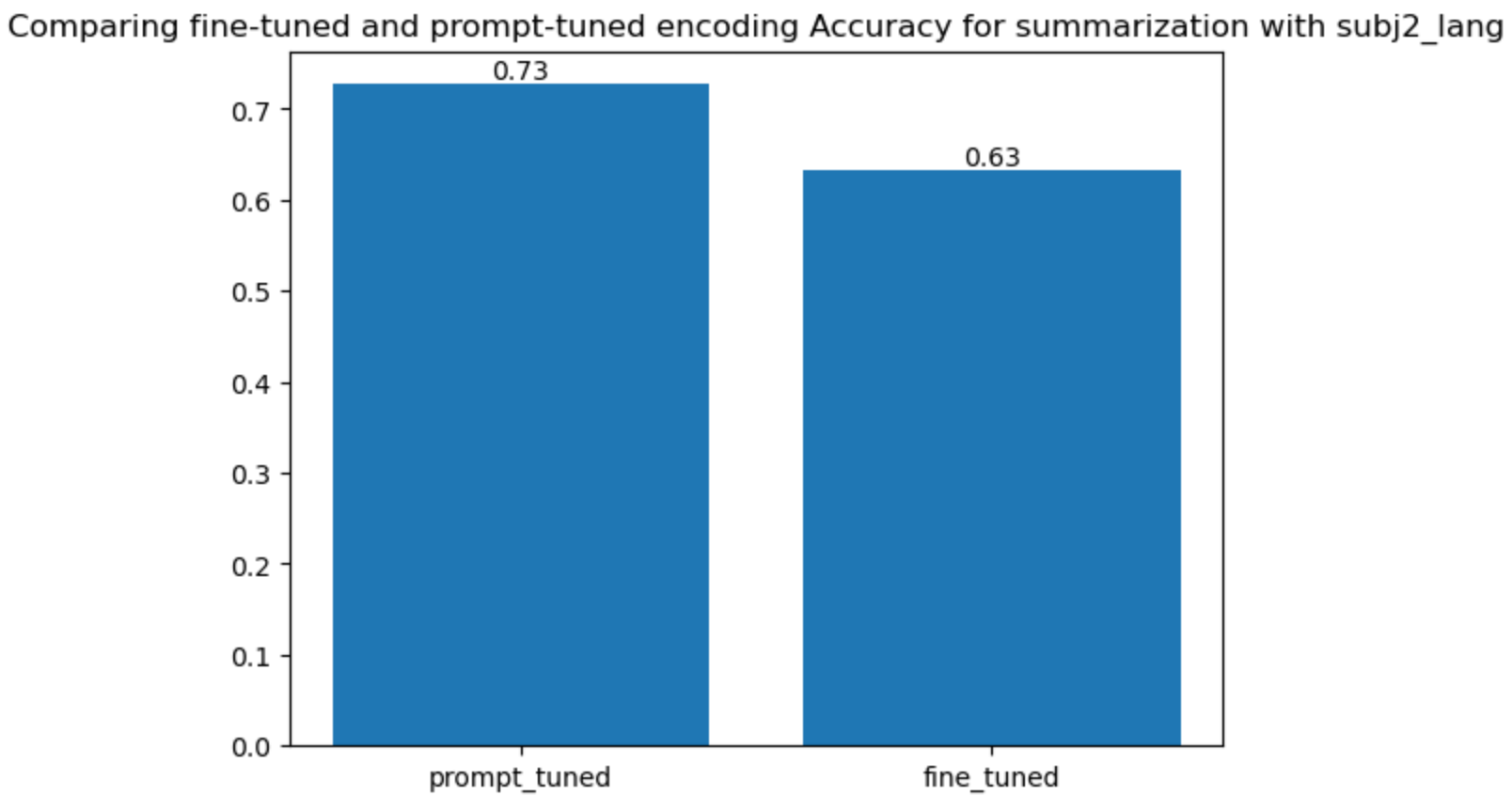

Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj1_lang
0.75
0.67
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

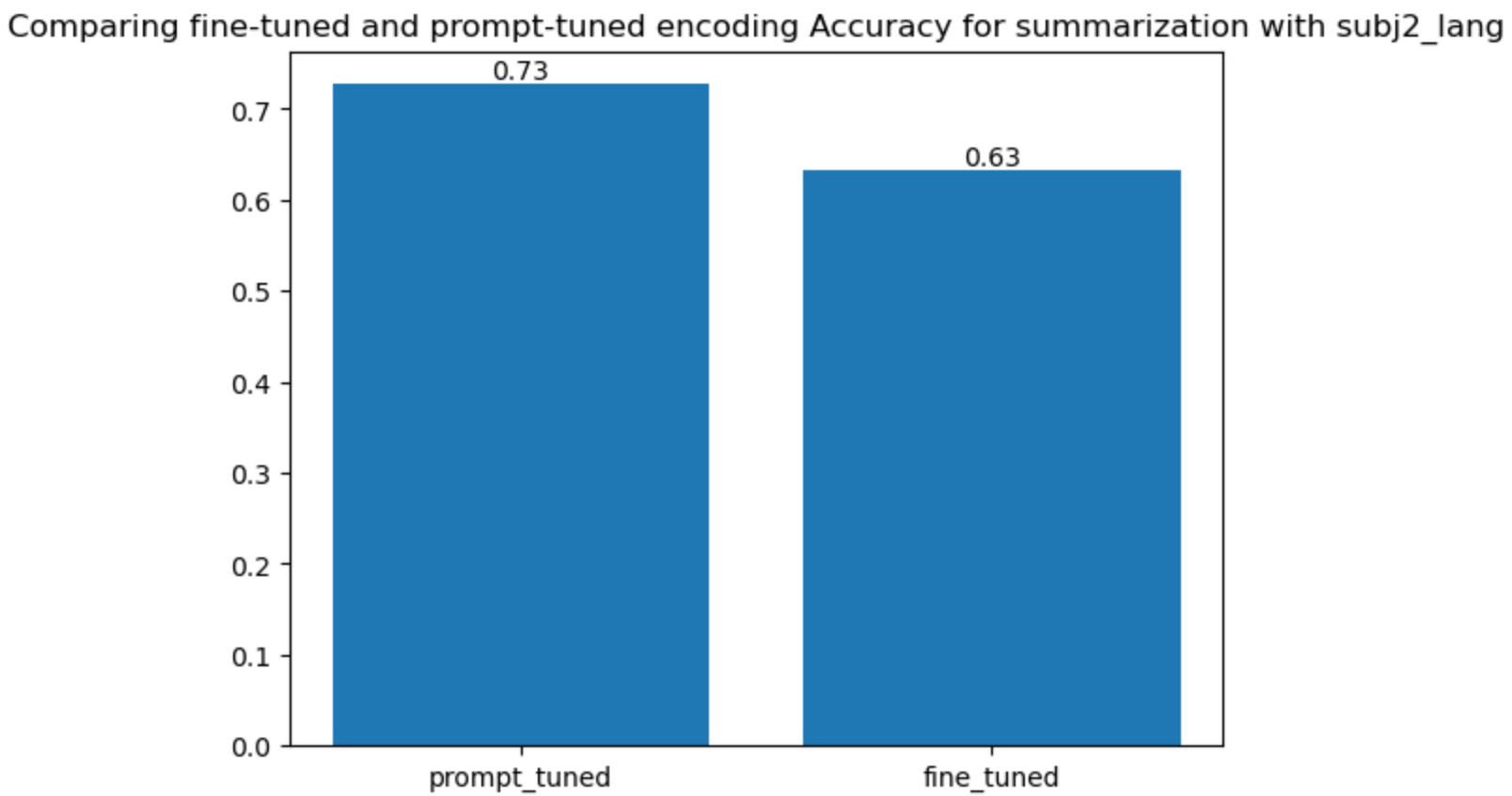

Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj1_lang
0.71
0.67
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

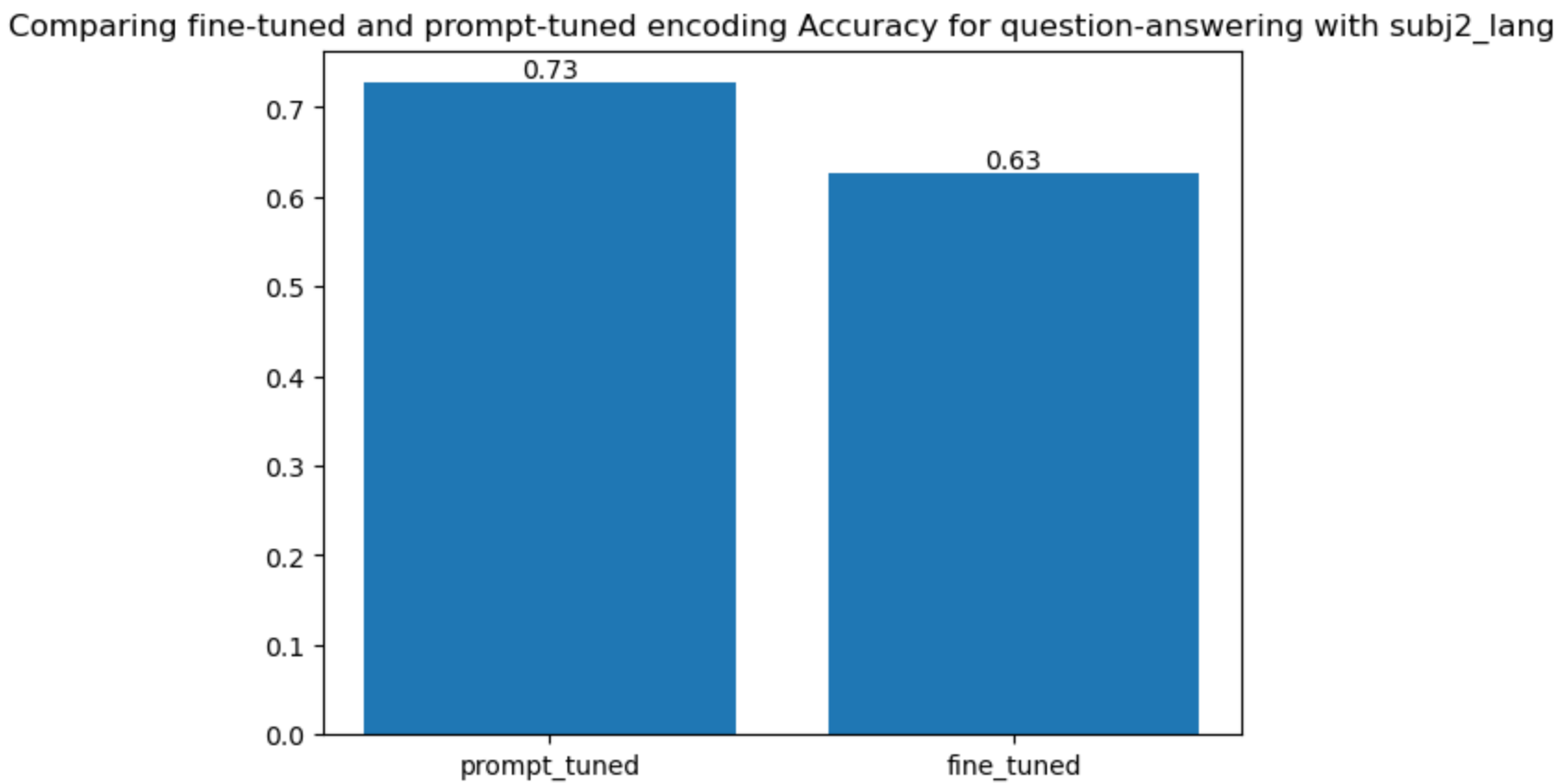

Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj1_lang
0.75
0.65
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

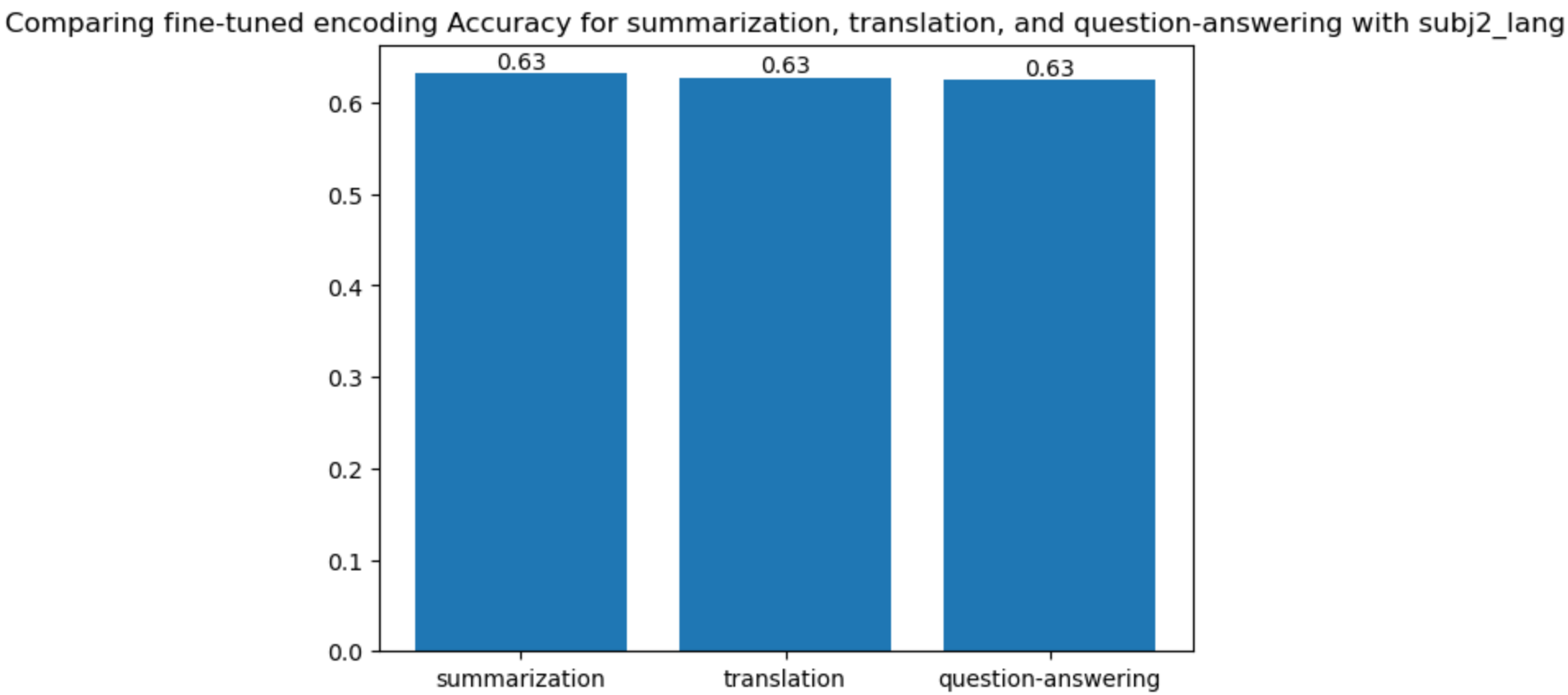

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_lang
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.67
0.67
0.65
summarization
translation
question-answering

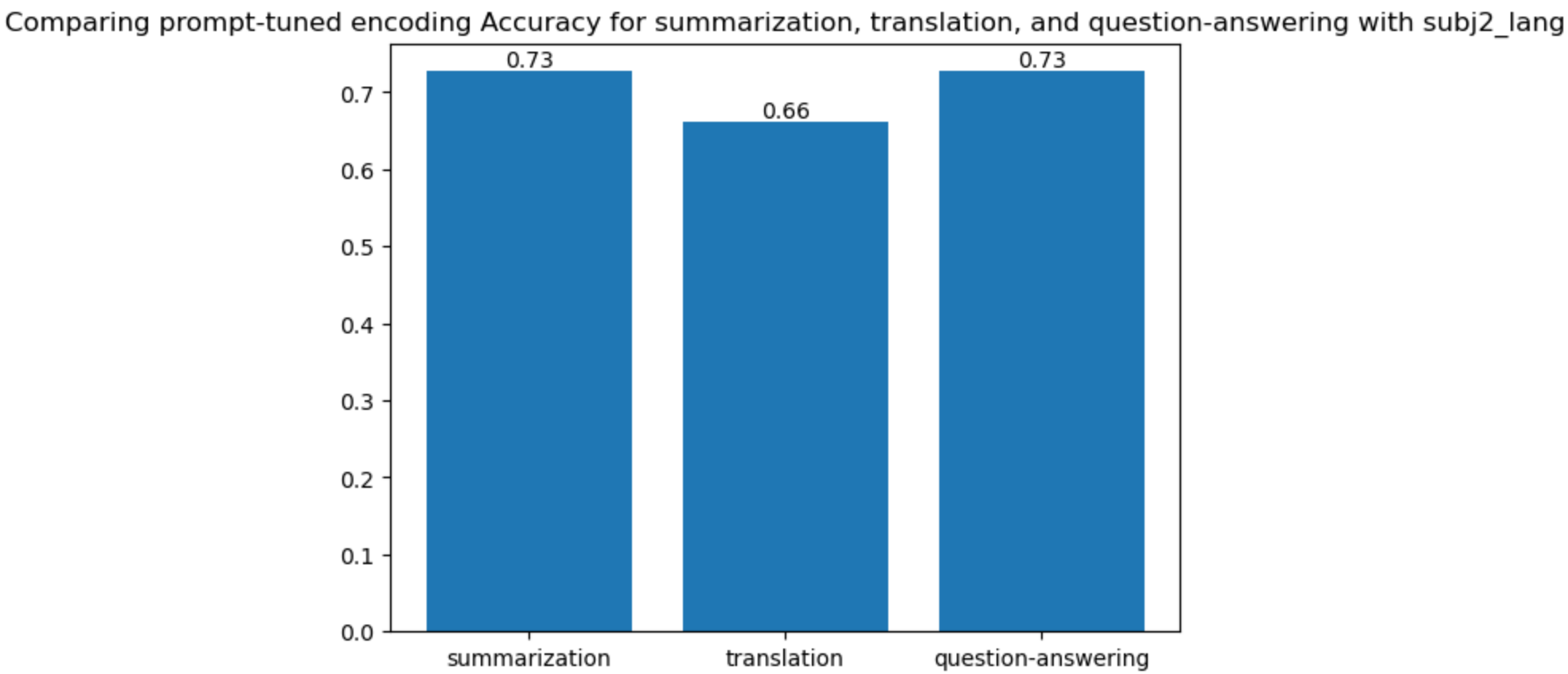

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_lang
0.75
0.71
0.75
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj2_lang
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj2_lang
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj2_lang
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

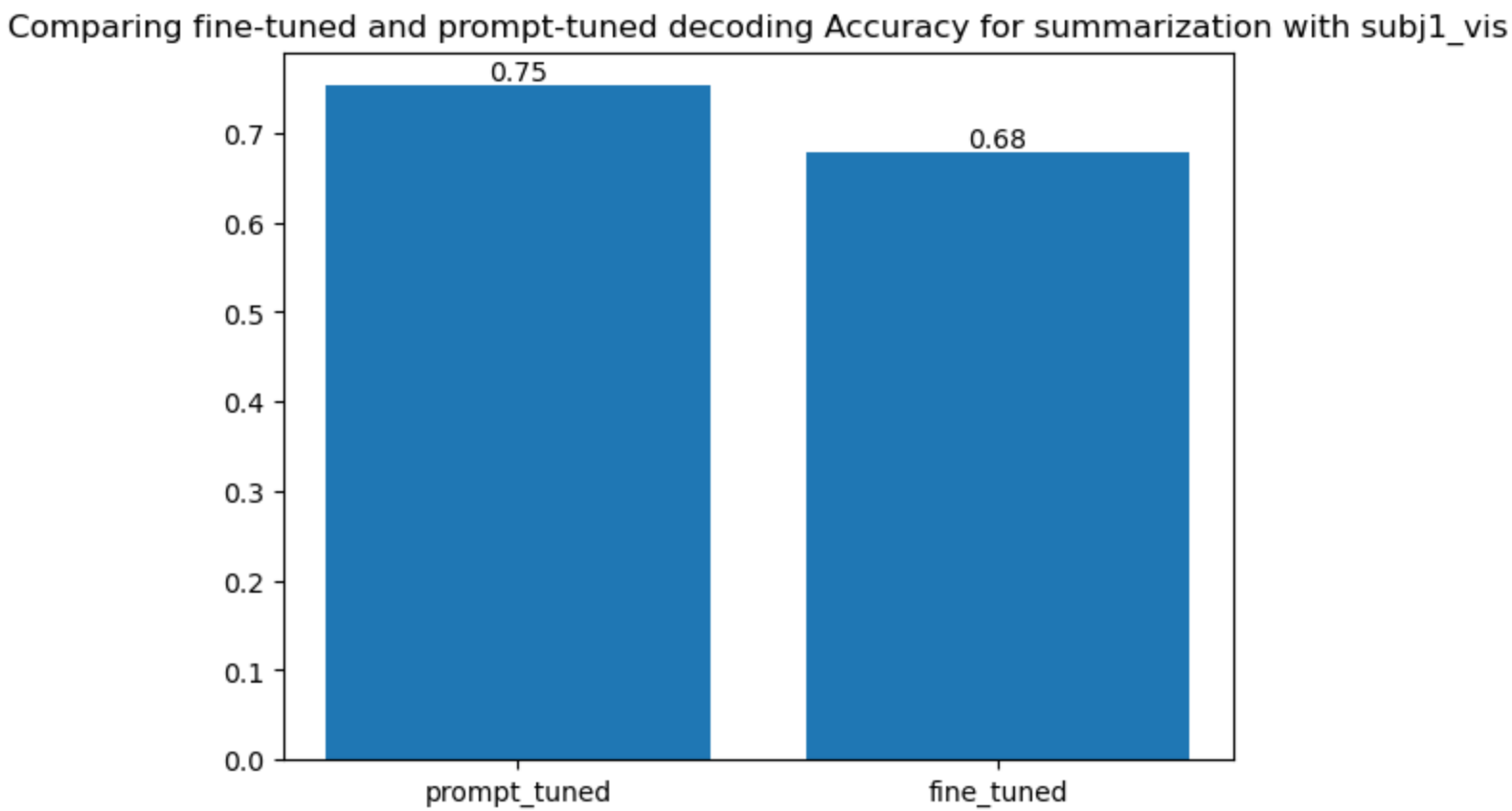

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_lang
0.63
0.63
0.63
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

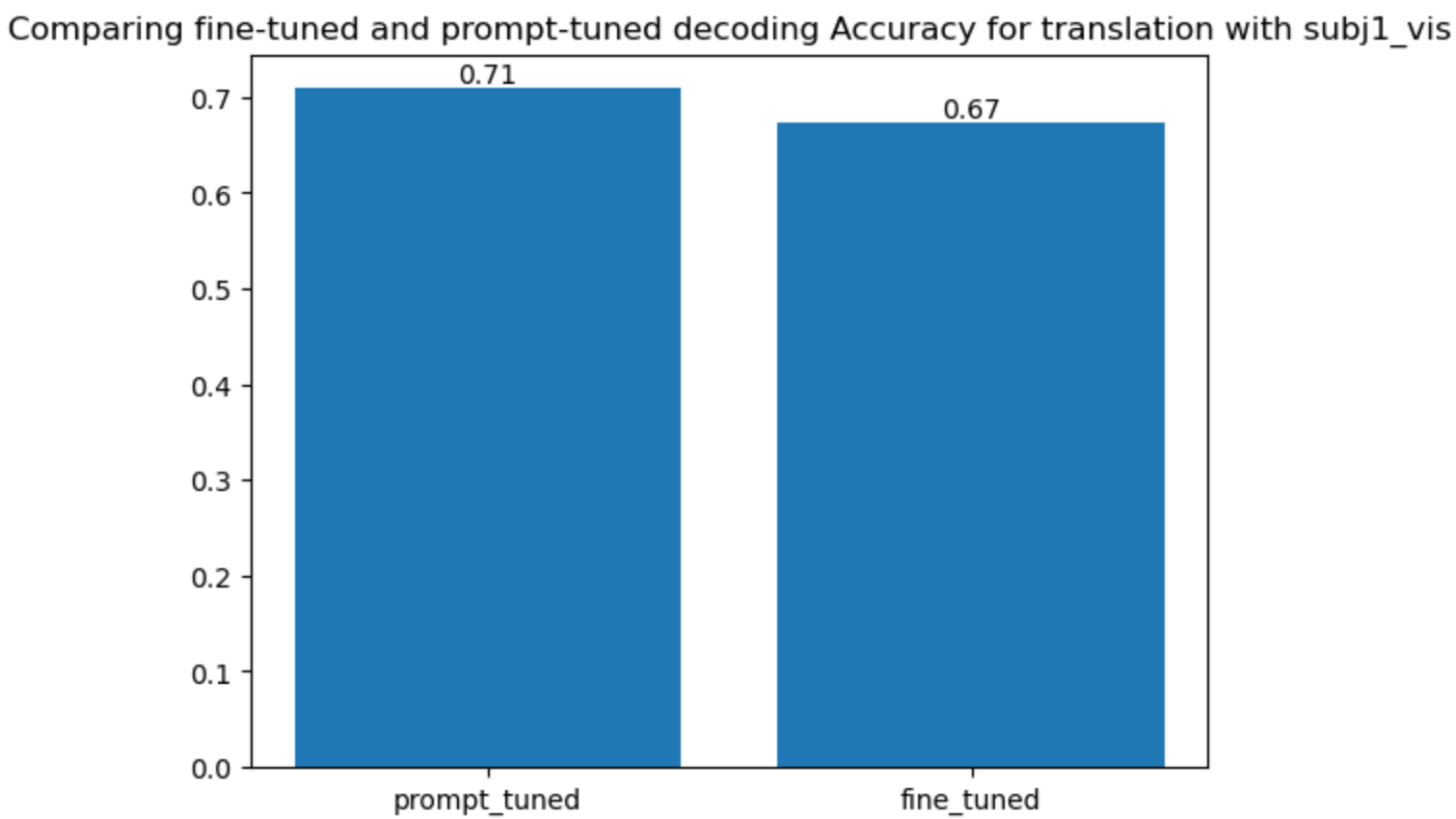

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_lang
0.73
0.66
0.73
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering

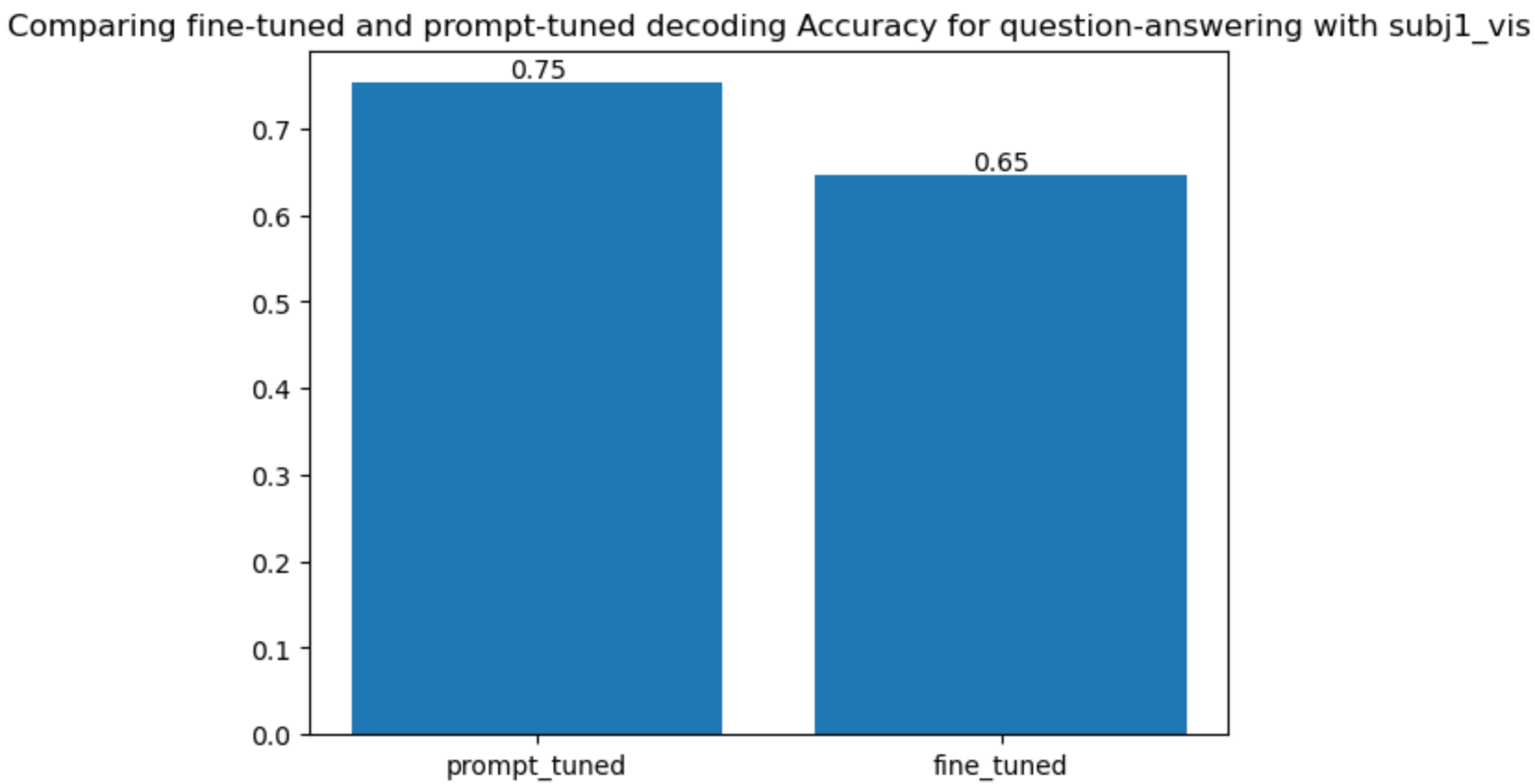
# Analysis of Prompt-tuned vs Fine-Tuned (Brain encoding)

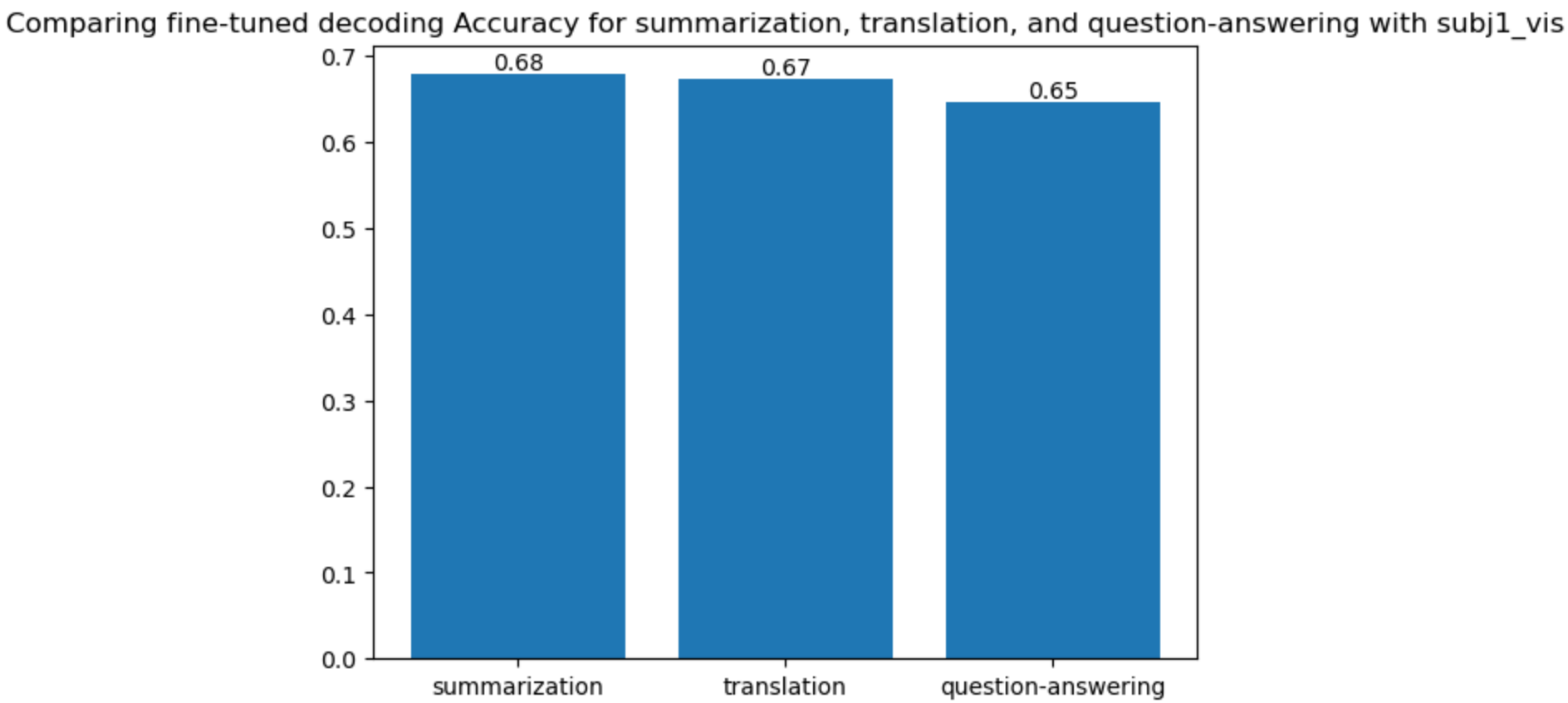
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain encoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed, this is consistent across subjects as well
- Wheras in Fine-tuned models, all models perform equally with no particular trend of out-performing the other.
- Also, the results overall are consistent across subjects.

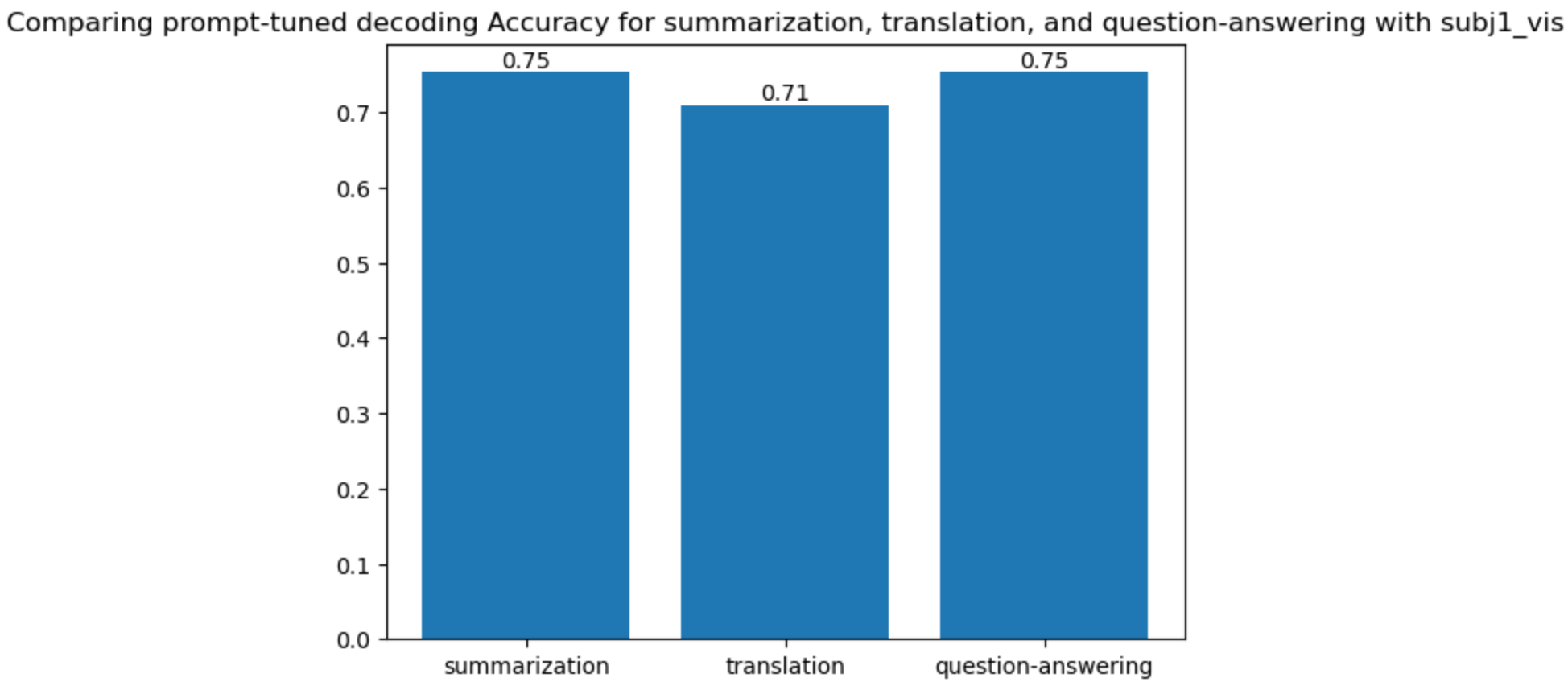
# VIS ROI

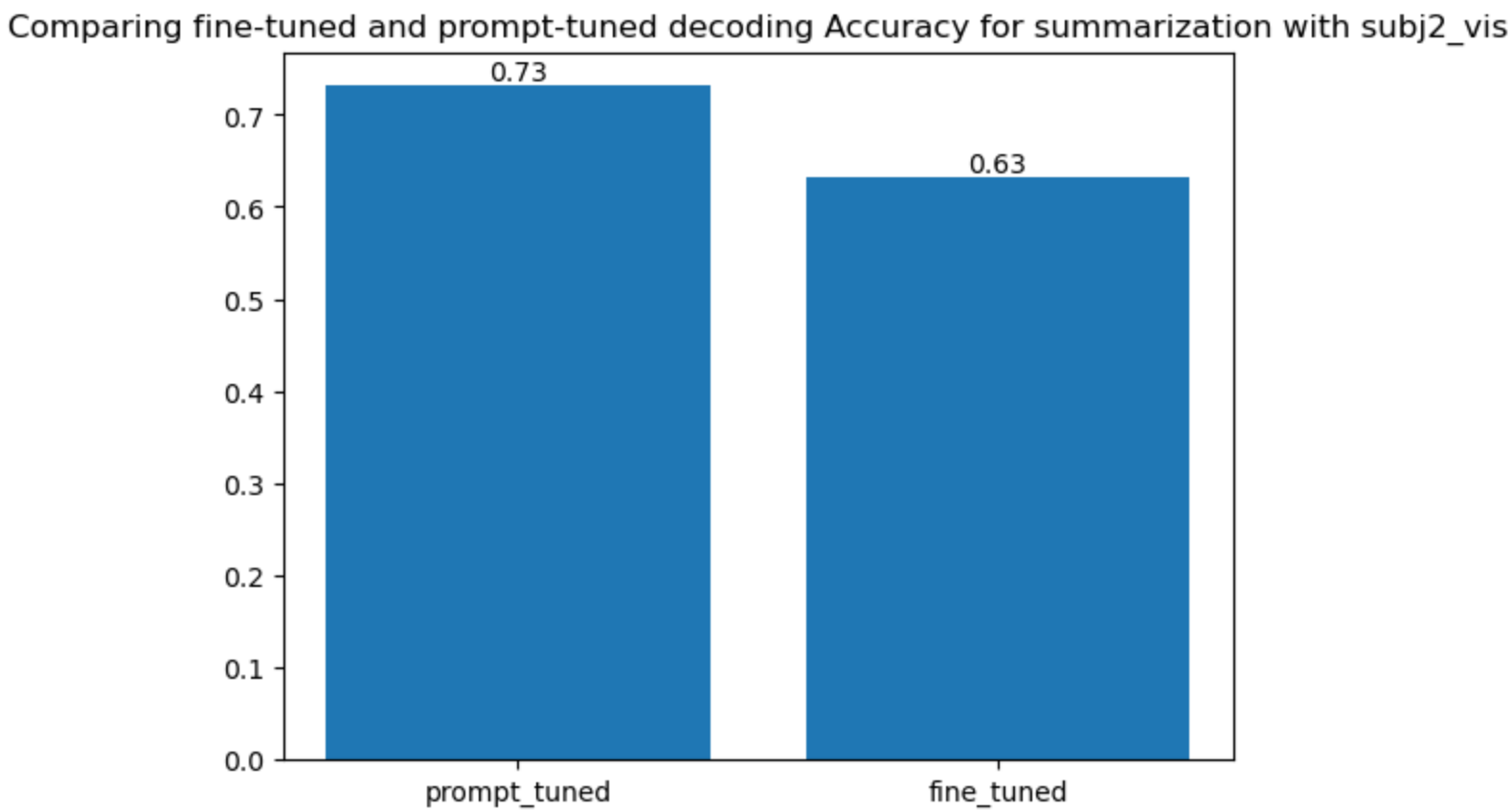

Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj1_vis
0.75
0.68
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

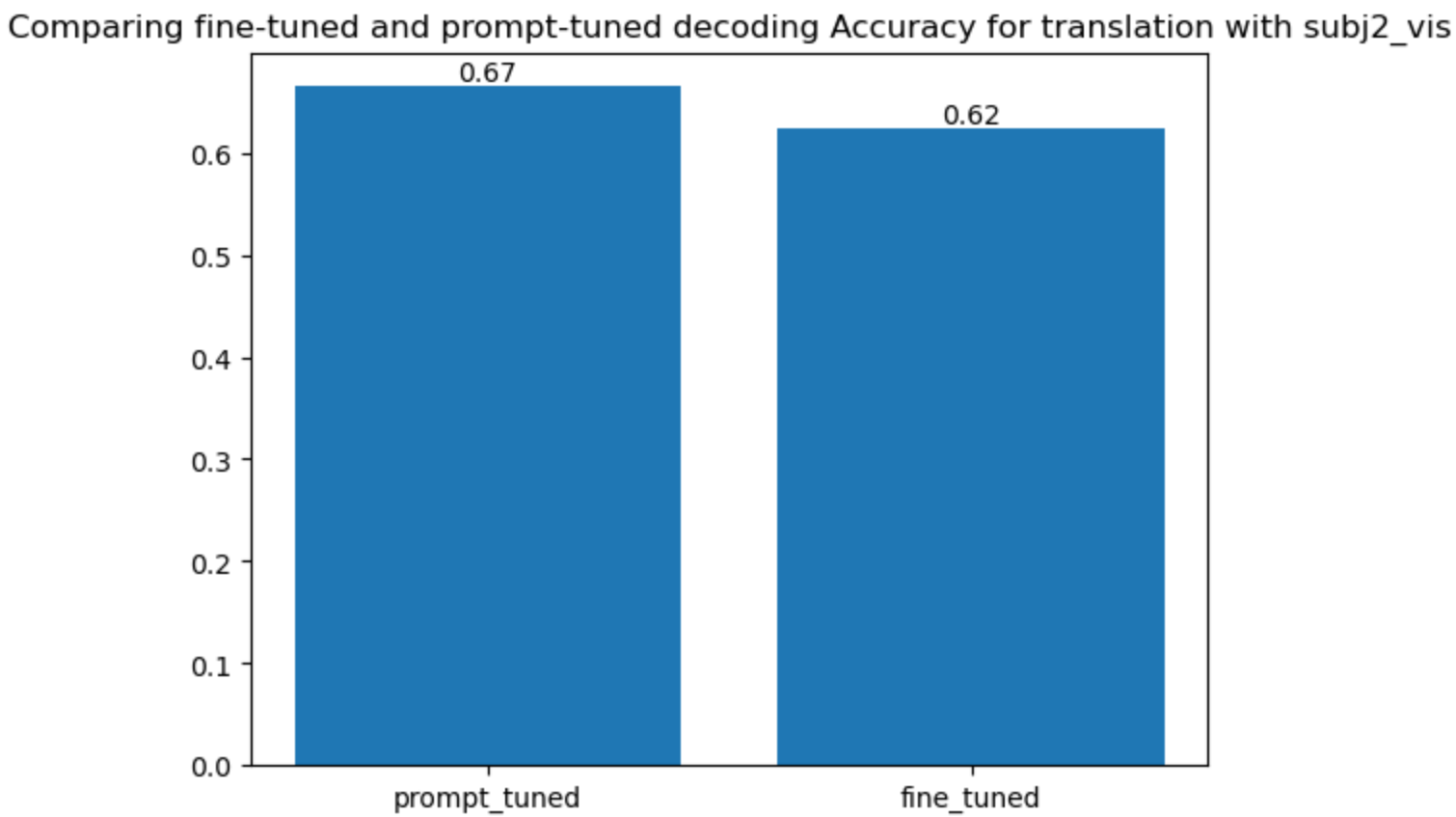

Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj1_vis
0.71
0.67
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

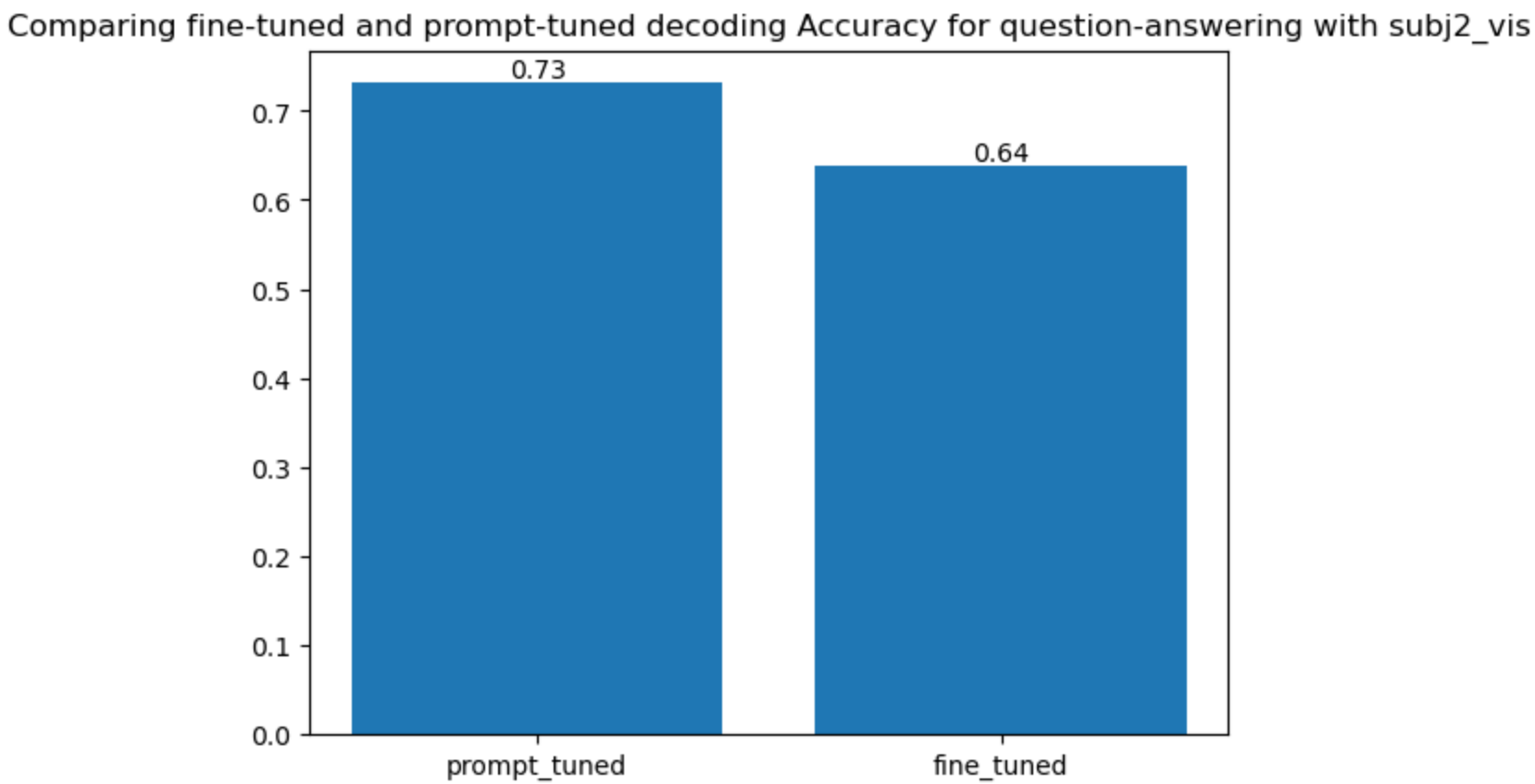

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj1_vis
0.75
0.65
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

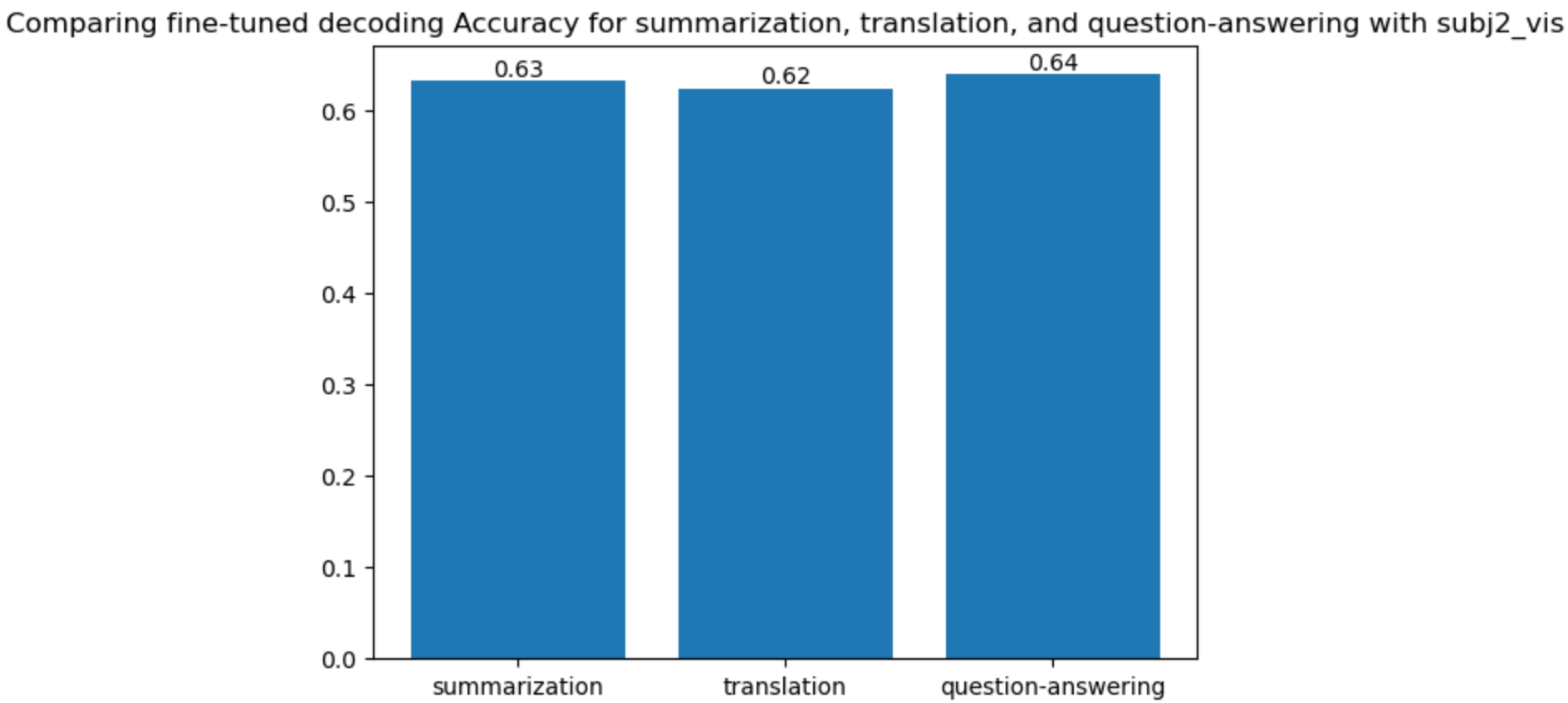

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_vis
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.68
0.67
0.65
summarization
translation
question-answering

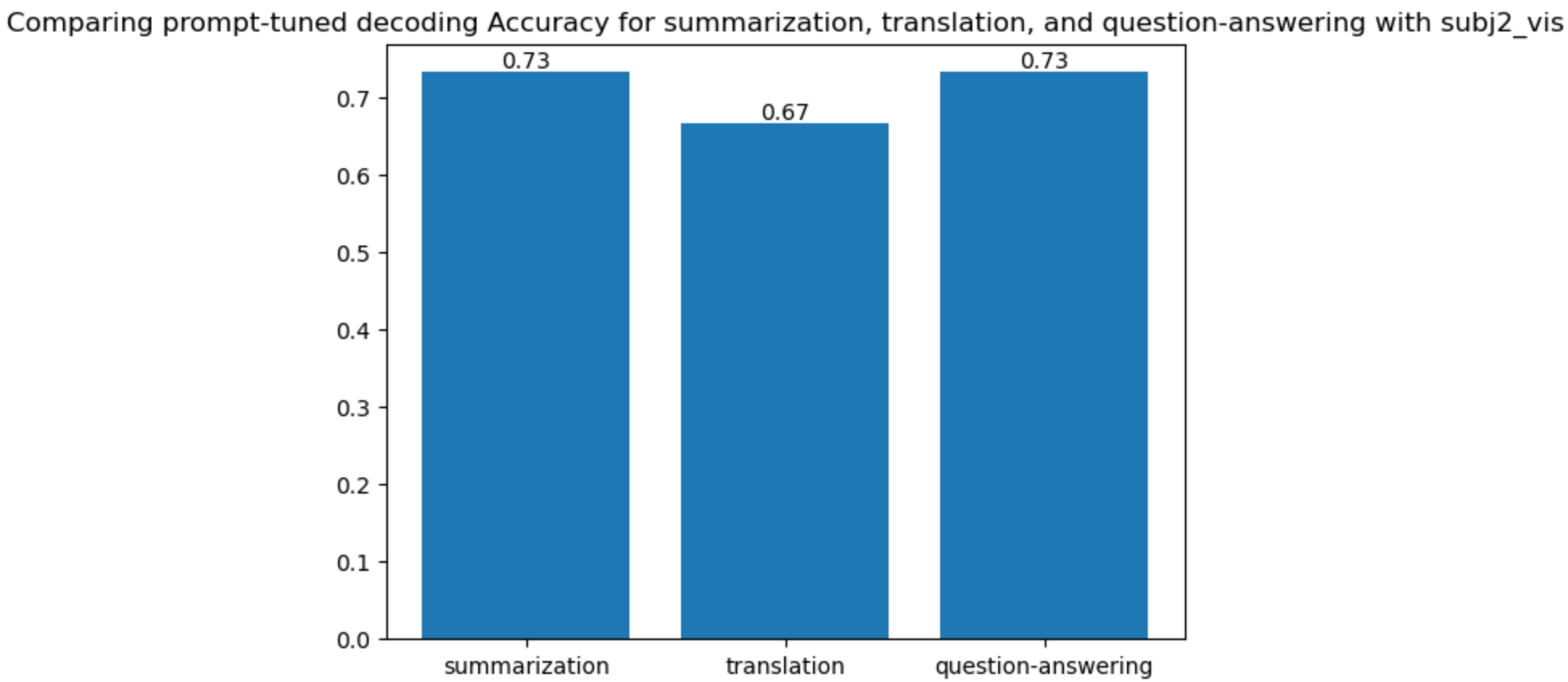

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_vis
0.75
0.71
0.75
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj2_vis
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj2_vis
0.67
0.62
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

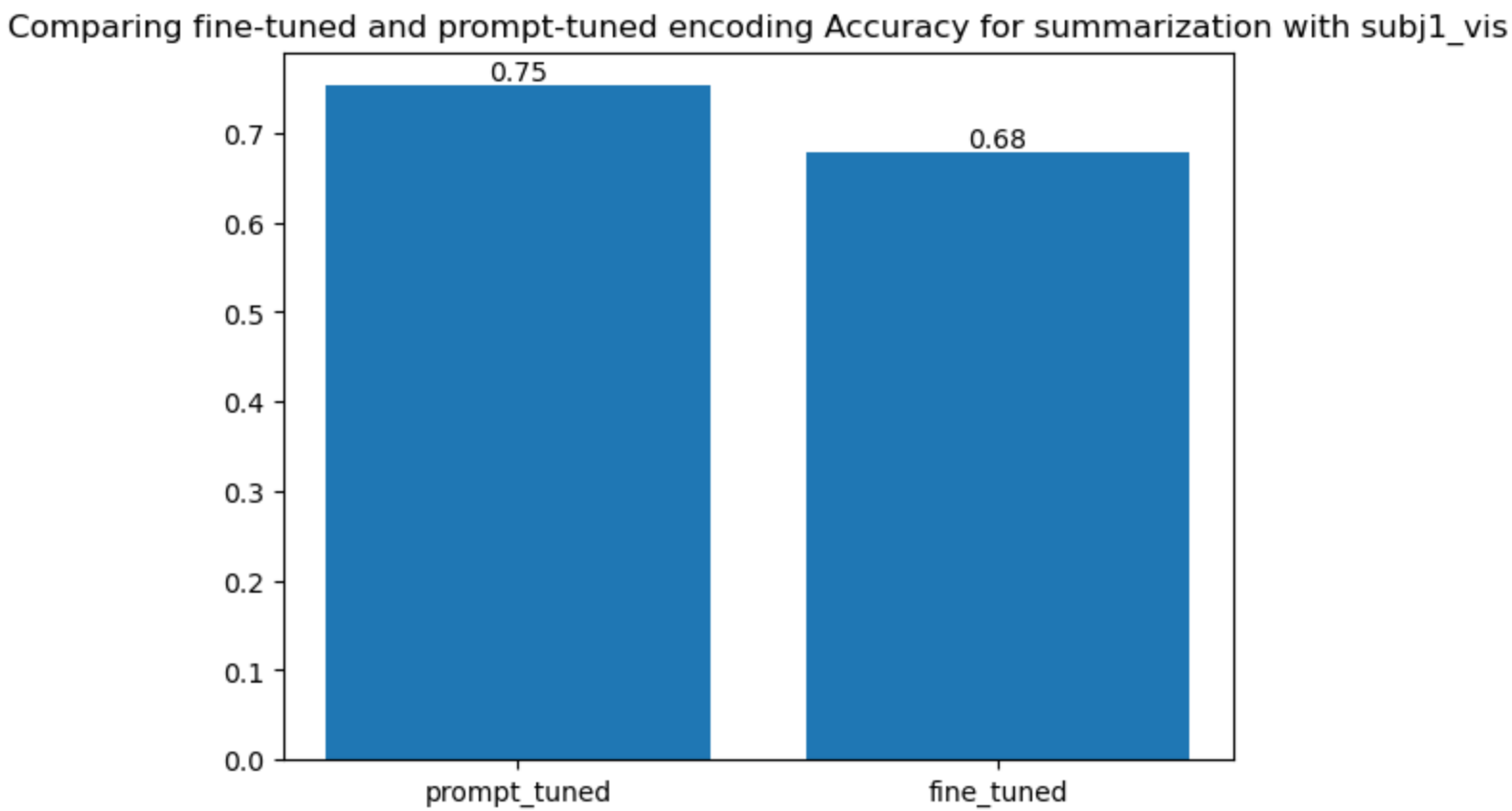

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj2_vis
0.73
0.64
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

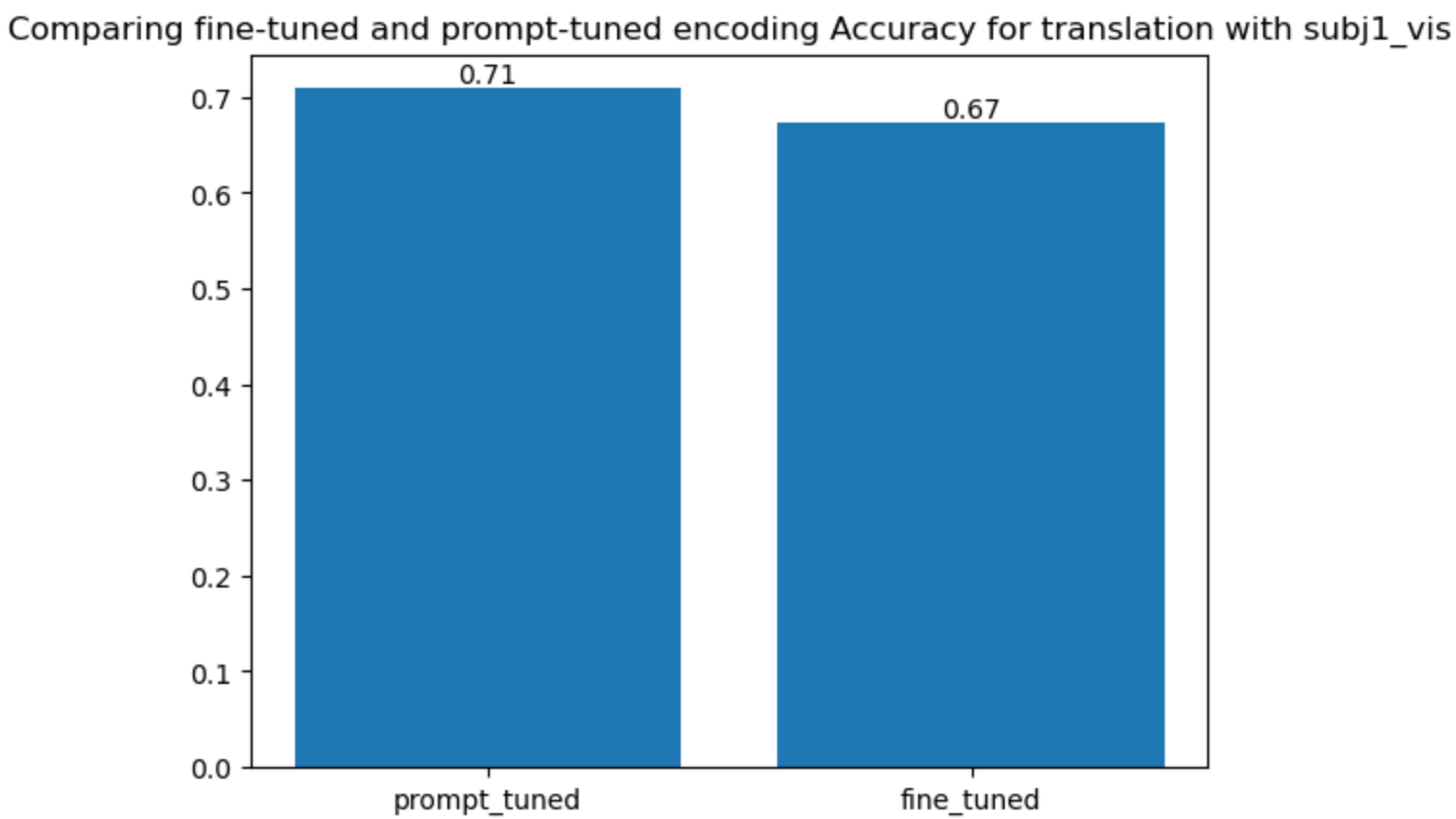

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_vis
0.63
0.62
0.64
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

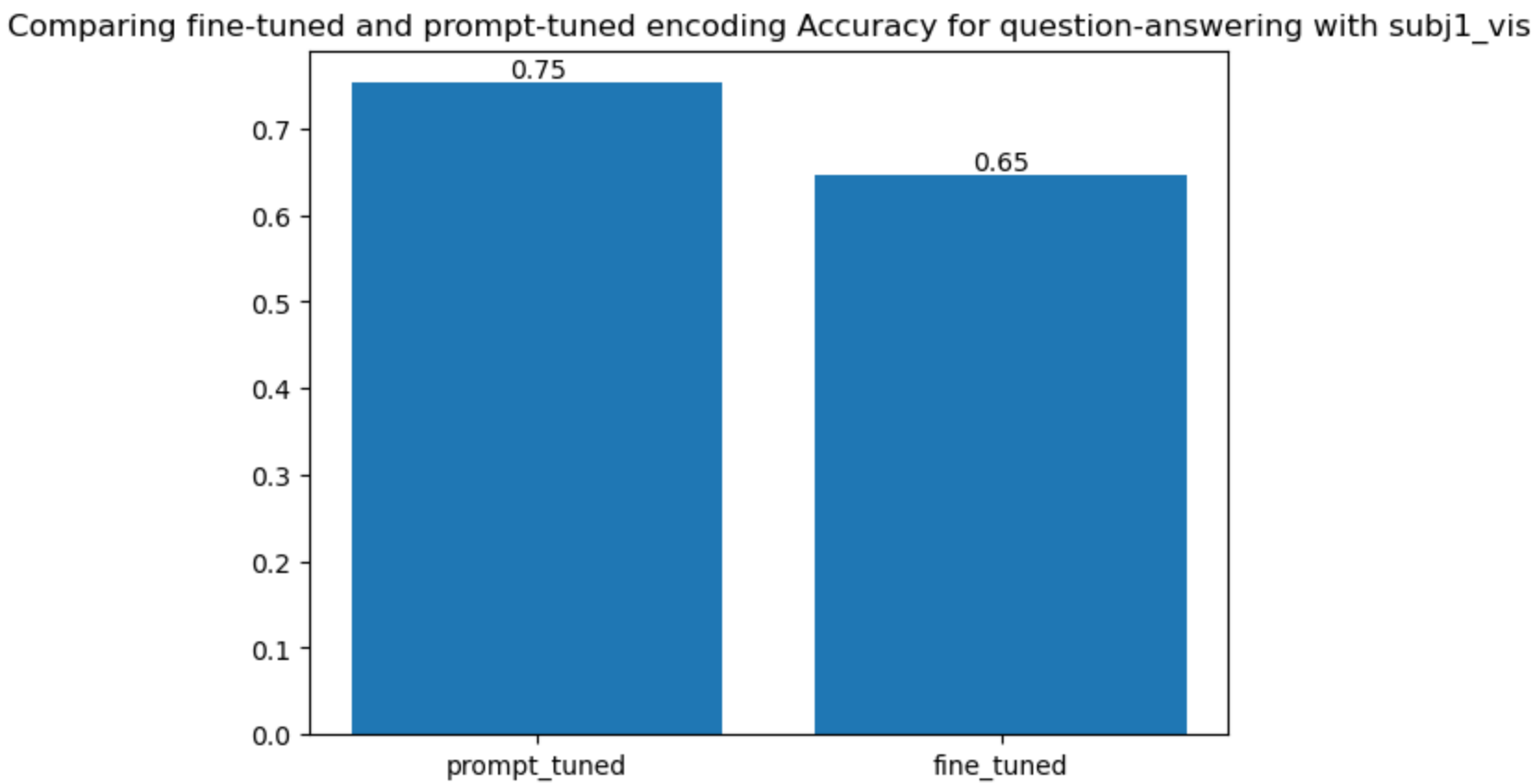

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_vis
0.73
0.67
0.73
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering

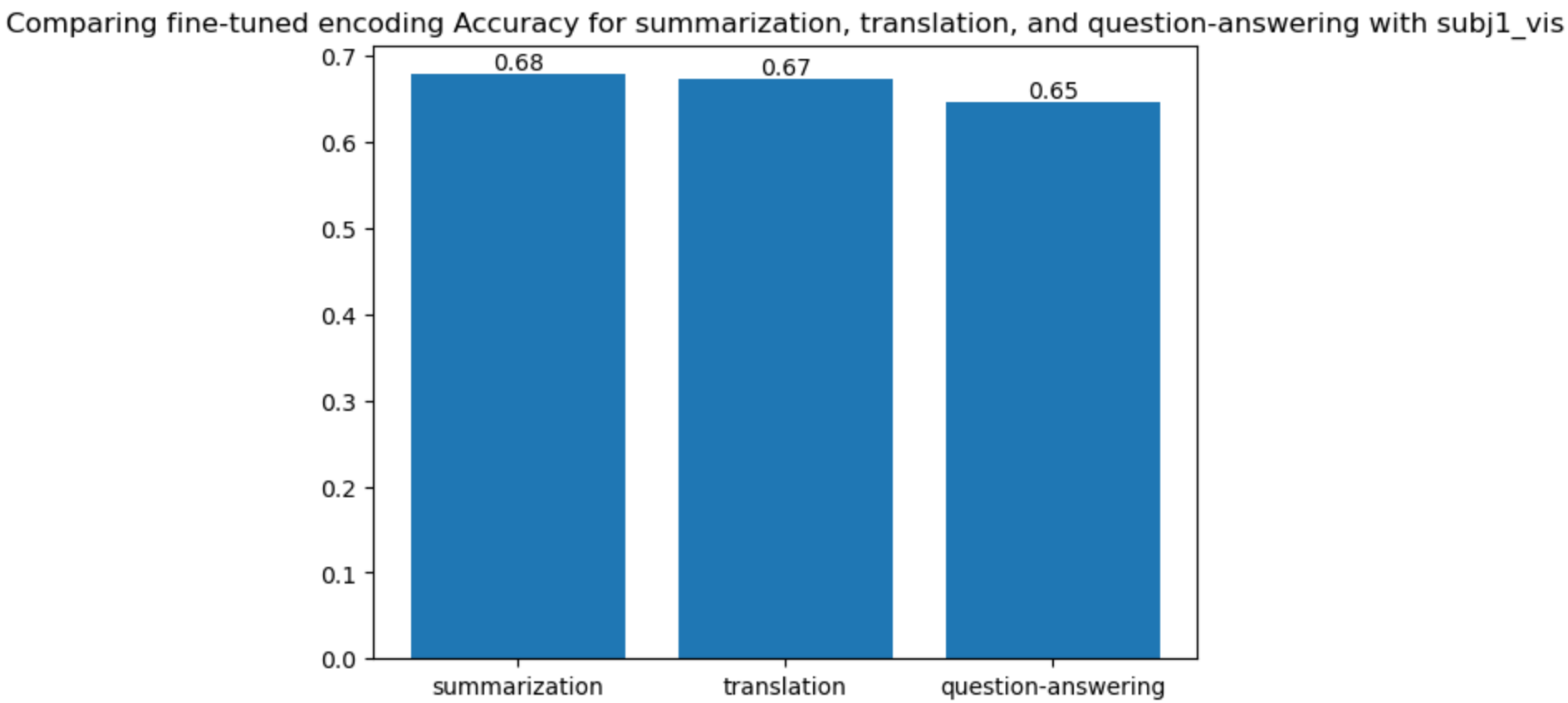
# Analysis of Prompt-tuned vs Fine-Tuned (Brain decoding)

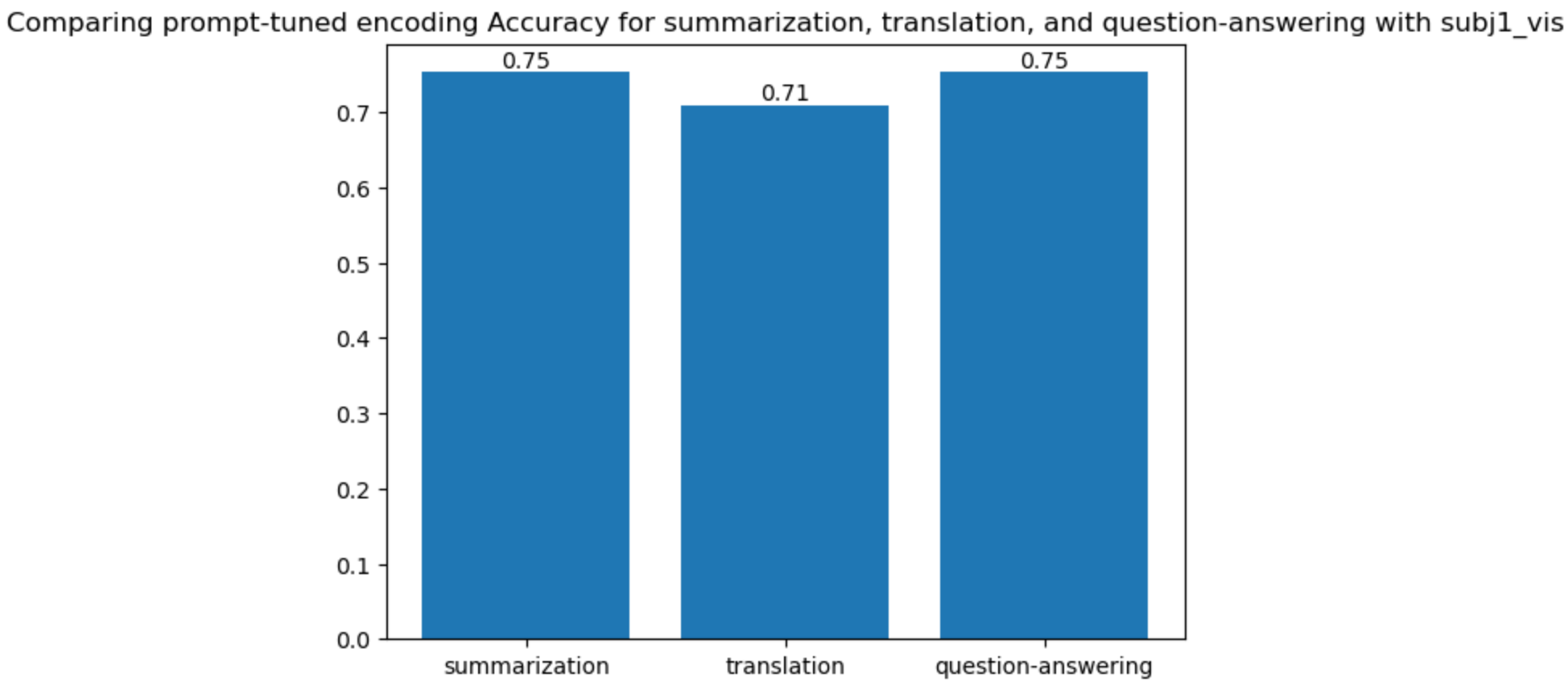
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain decoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed.
- Wheras in Fine-tuned models, all perform at an equal level, with no particular trend of one outperforming the other.
- Also the results we got are **consistent** across subjects!

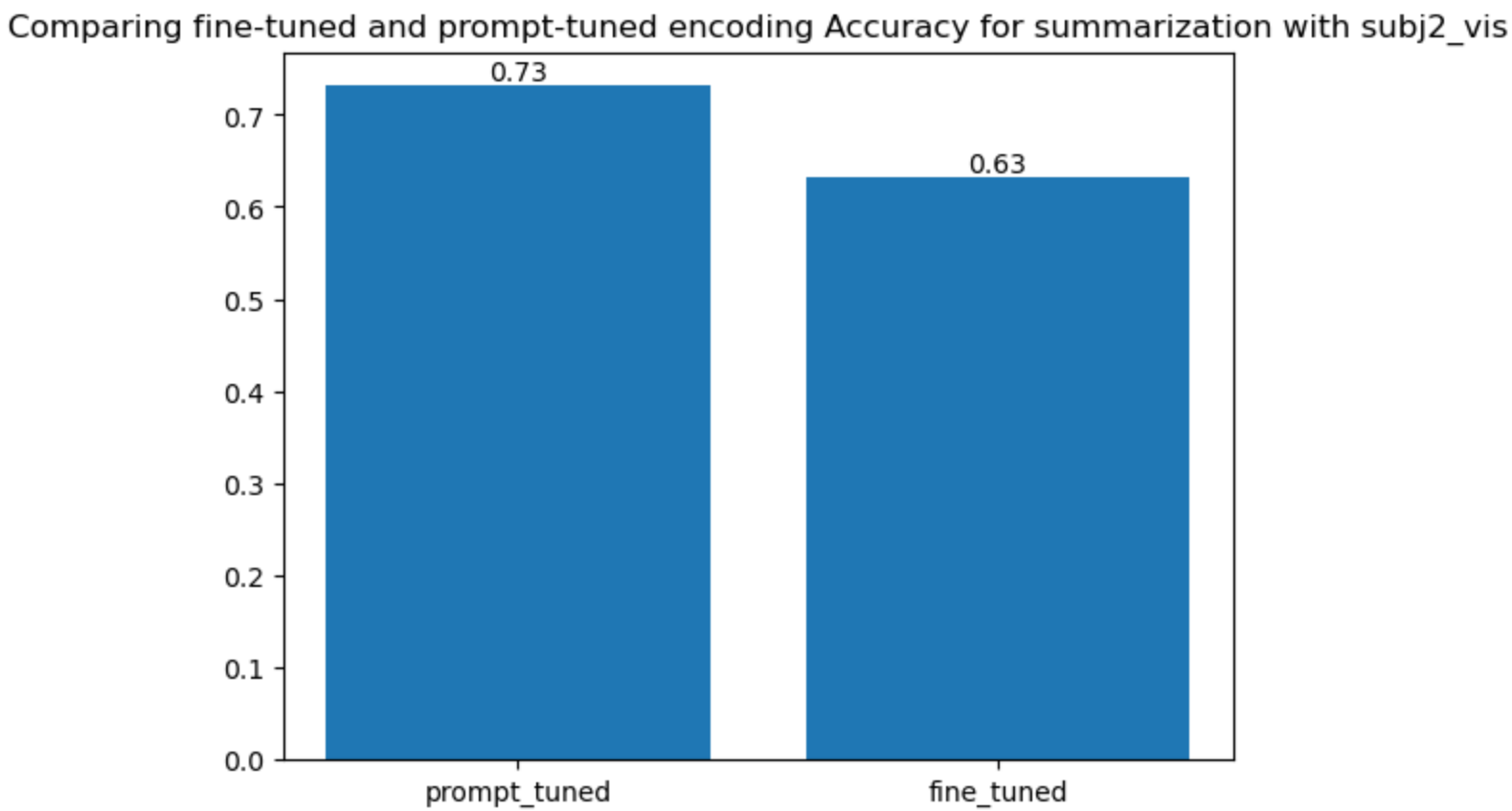

Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj1_vis
0.75
0.68
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

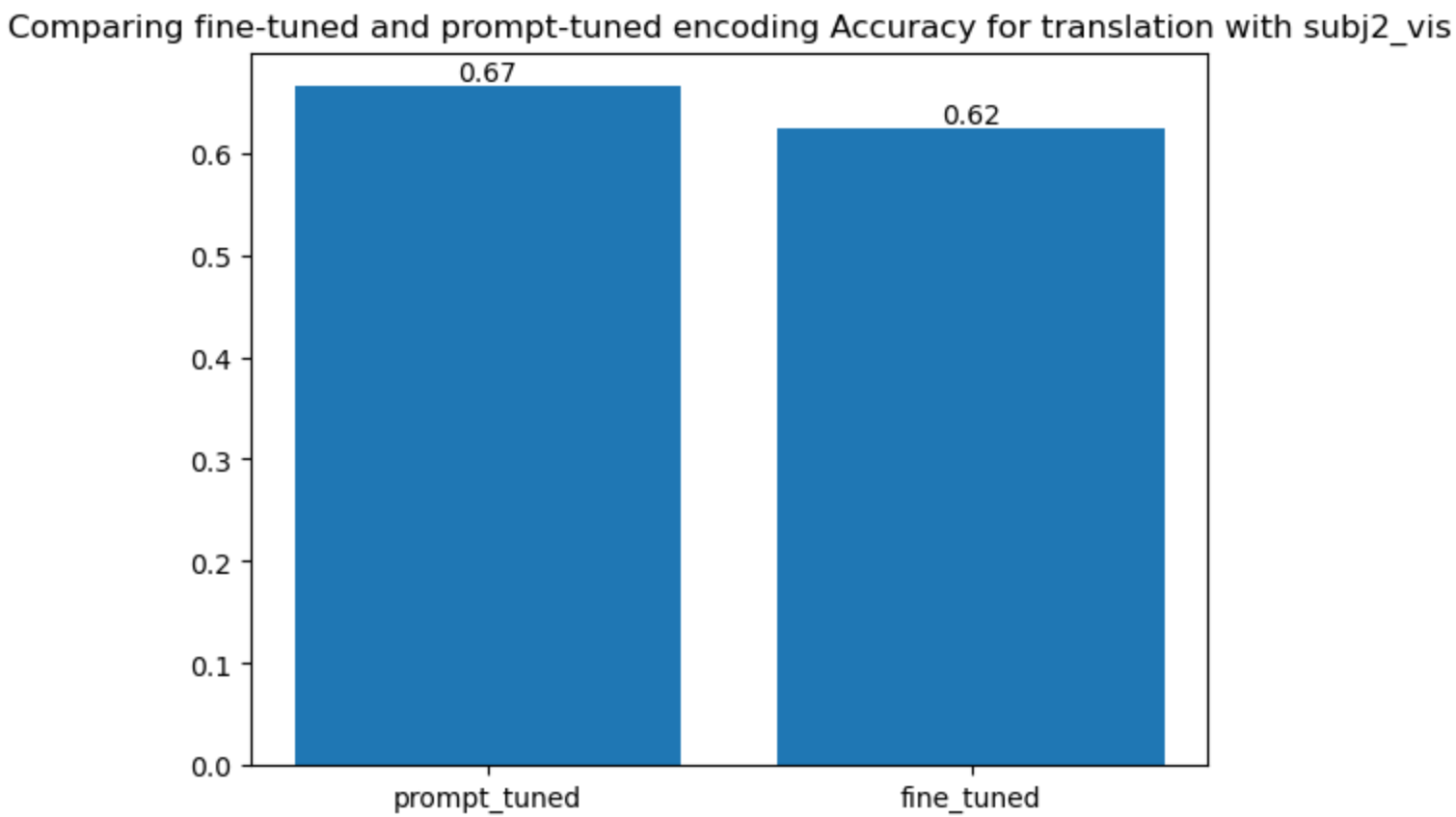

Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj1_vis
0.71
0.67
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

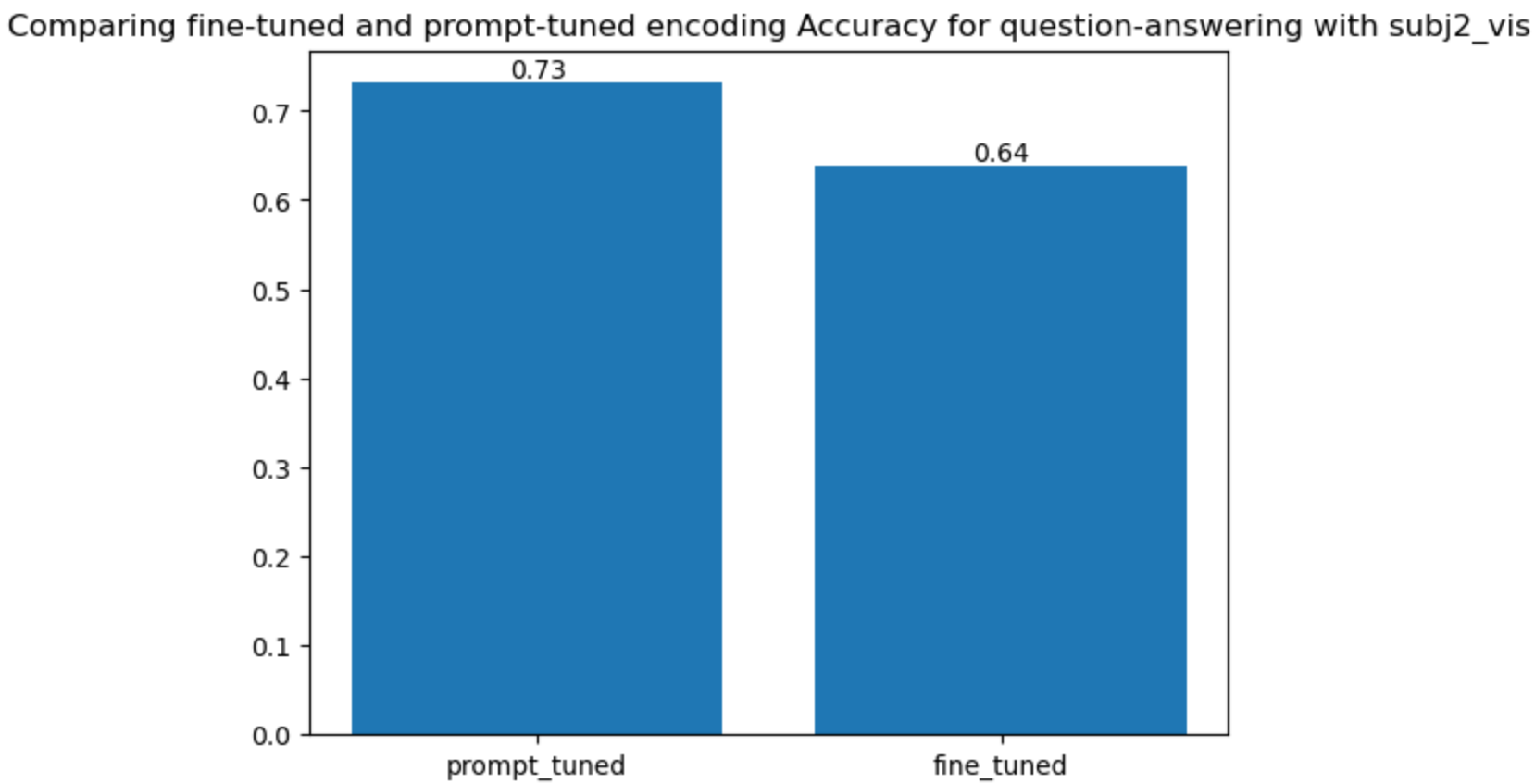

Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj1_vis
0.75
0.65
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

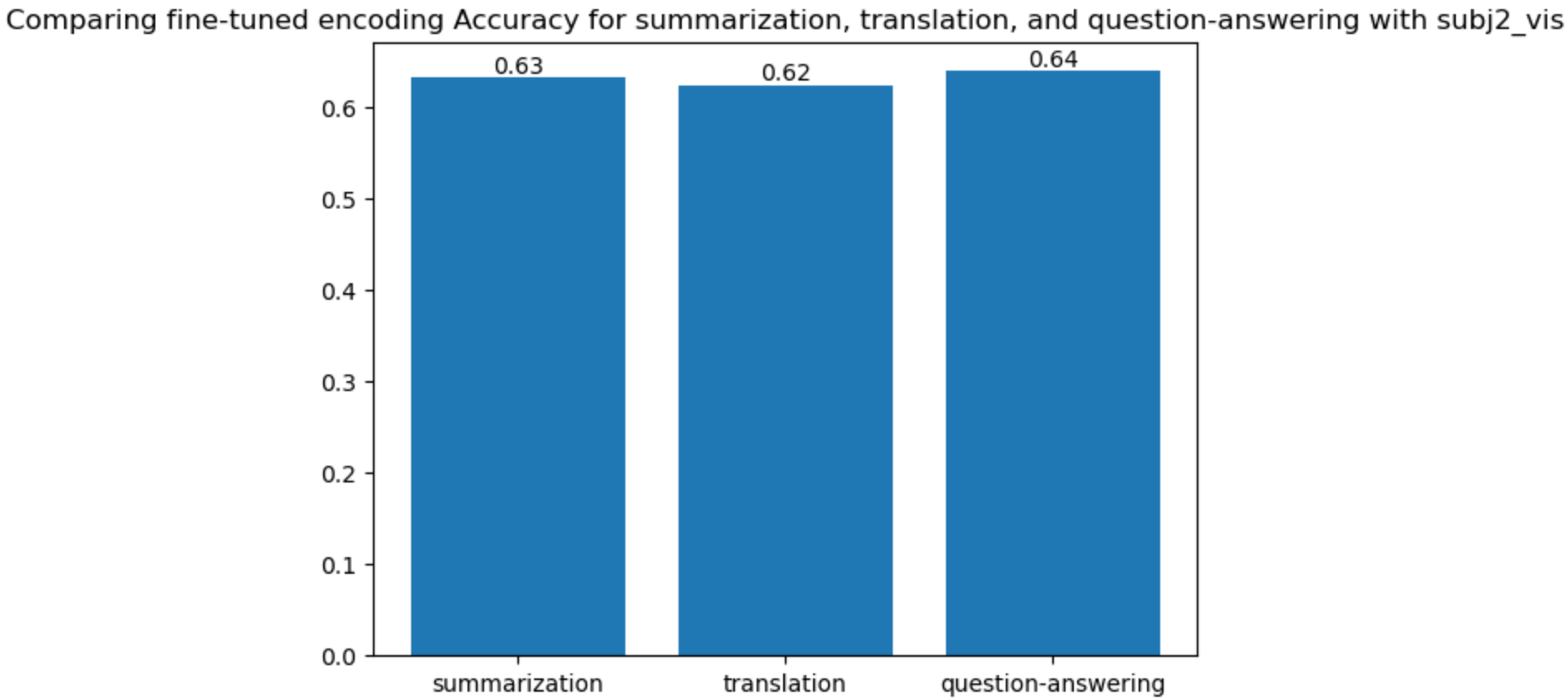

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_vis
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.68
0.67
0.65
summarization
translation
question-answering

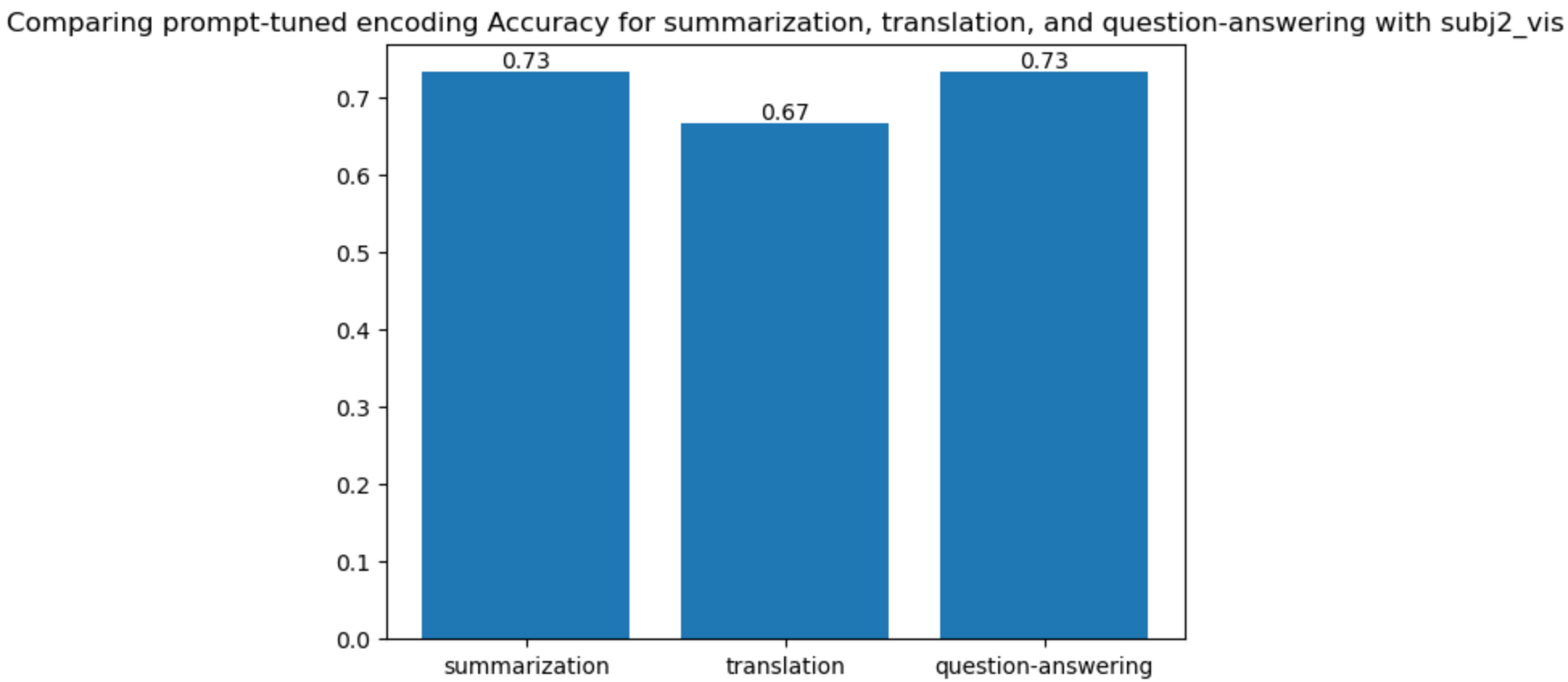

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_vis
0.75
0.71
0.75
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj2_vis
0.73
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj2_vis
0.67
0.62
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj2_vis
0.73
0.64
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned

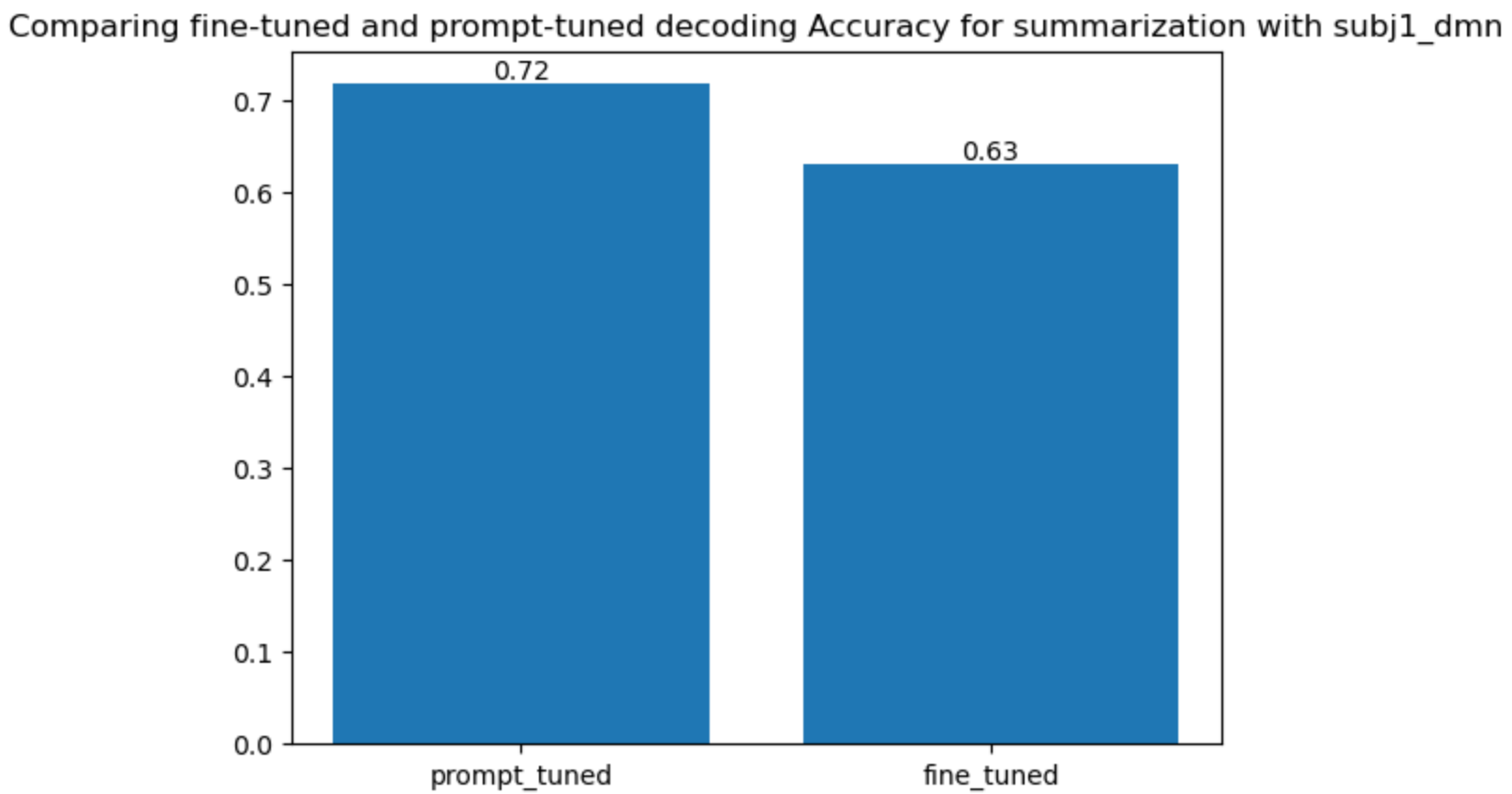

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_vis
0.63
0.62
0.64
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

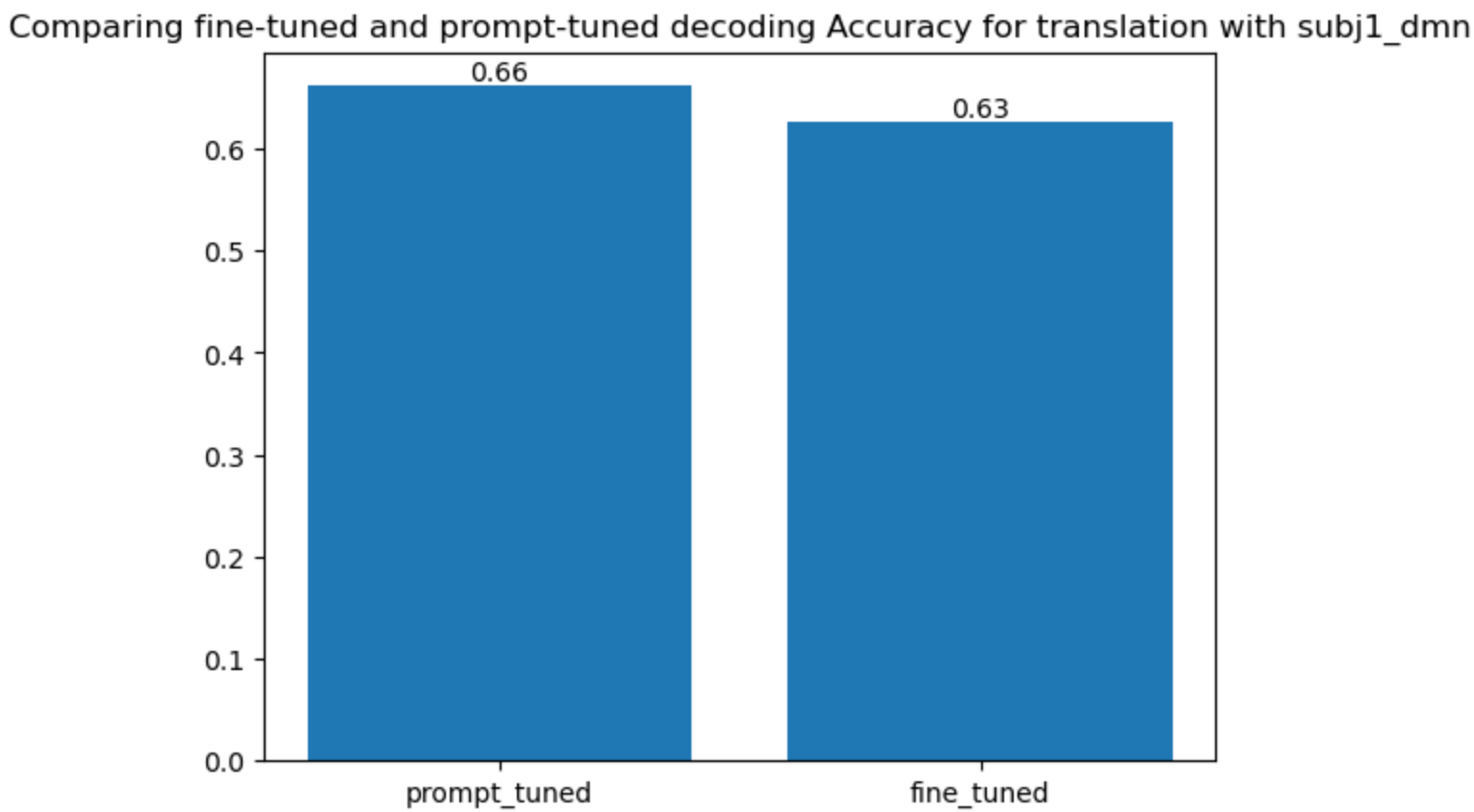

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_vis
0.73
0.67
0.73
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering

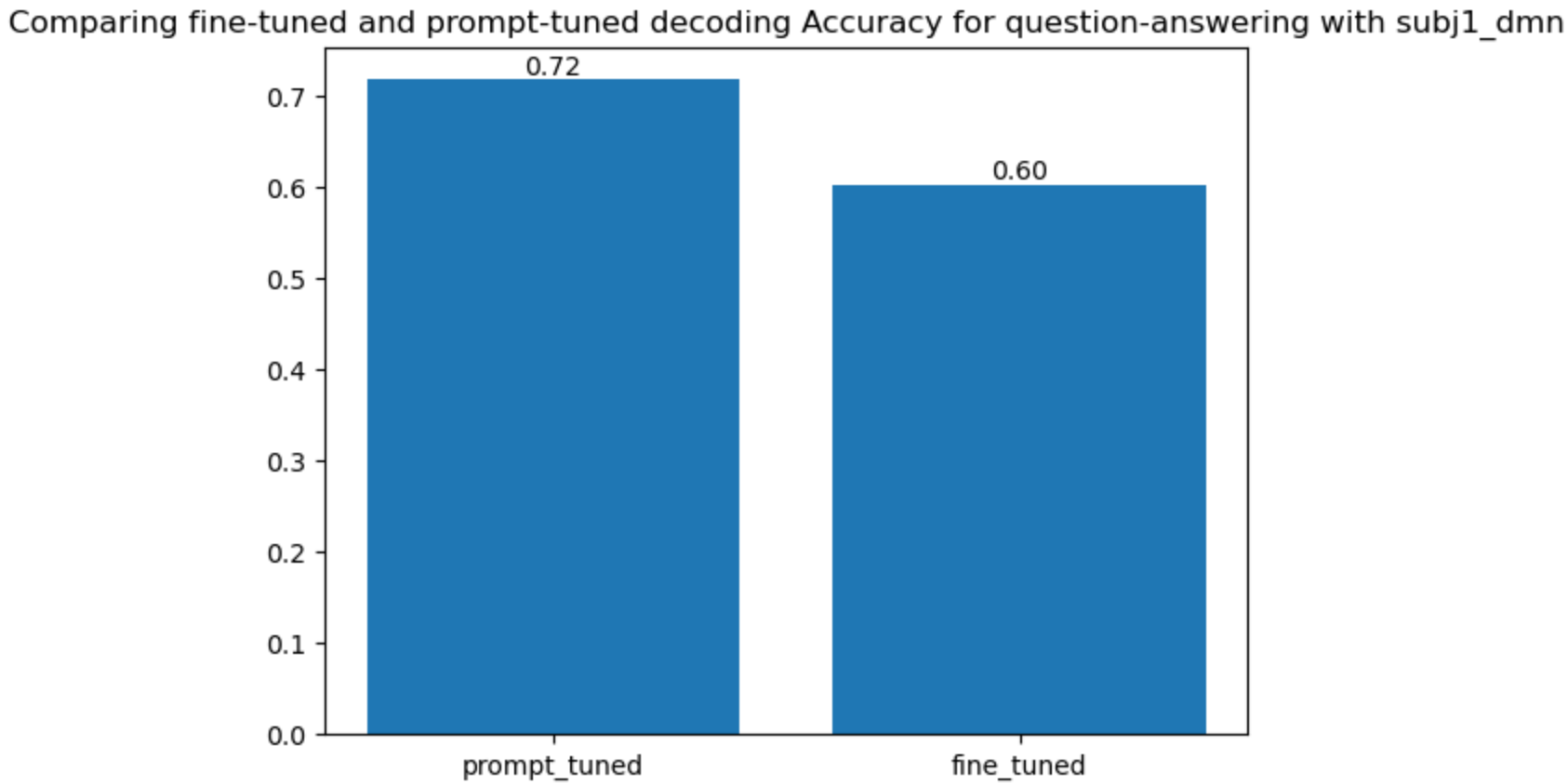
# Analysis of Prompt-tuned vs Fine-Tuned (Brain encoding)

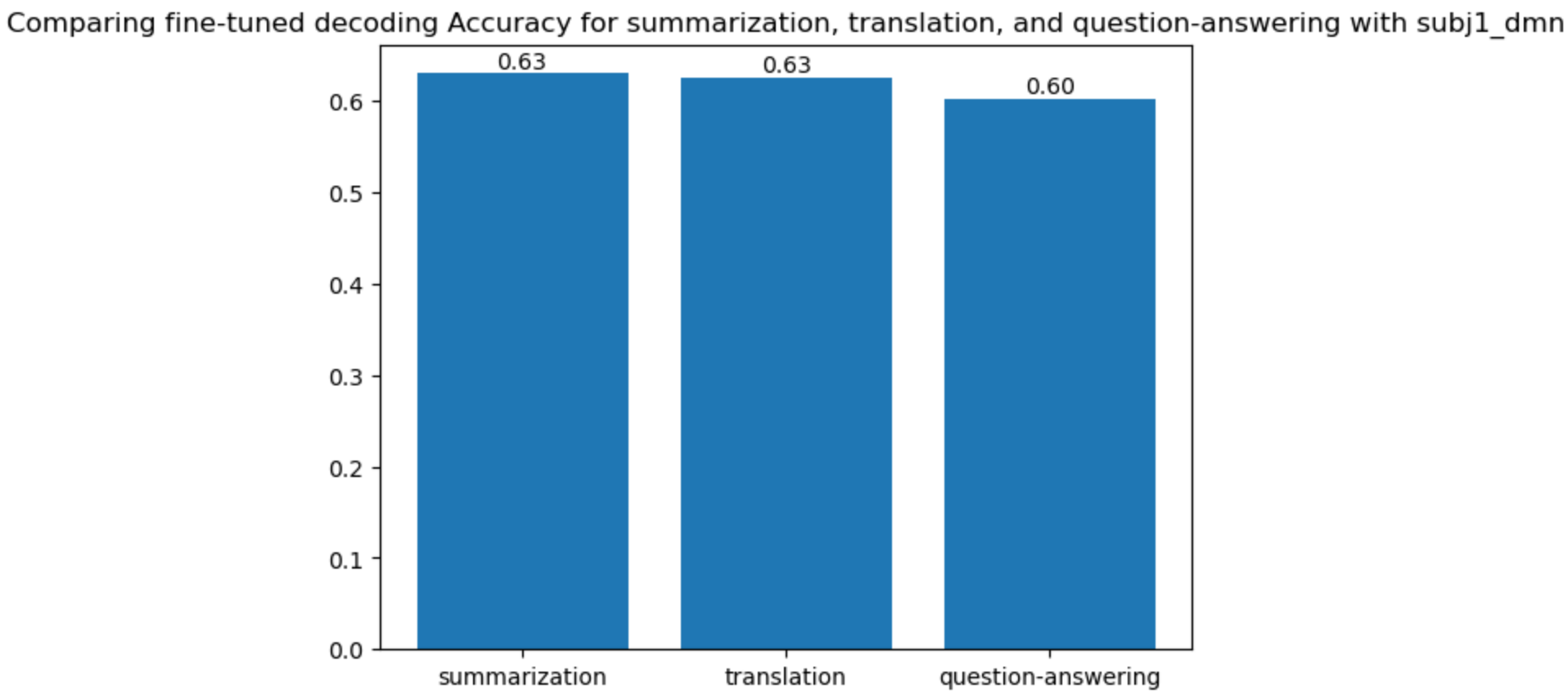
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain encoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed, this is consistent across subjects as well
- Wheras in Fine-tuned models, all models perform equally with no particular trend of out-performing the other, in one subject Sum model outperformed, in another, the QA model outperformed
- Since it is the Vision ROI, which focuses on object detection, not much of a direct correlation can be drawn.
- Also, the results overall are consistent across subjects.

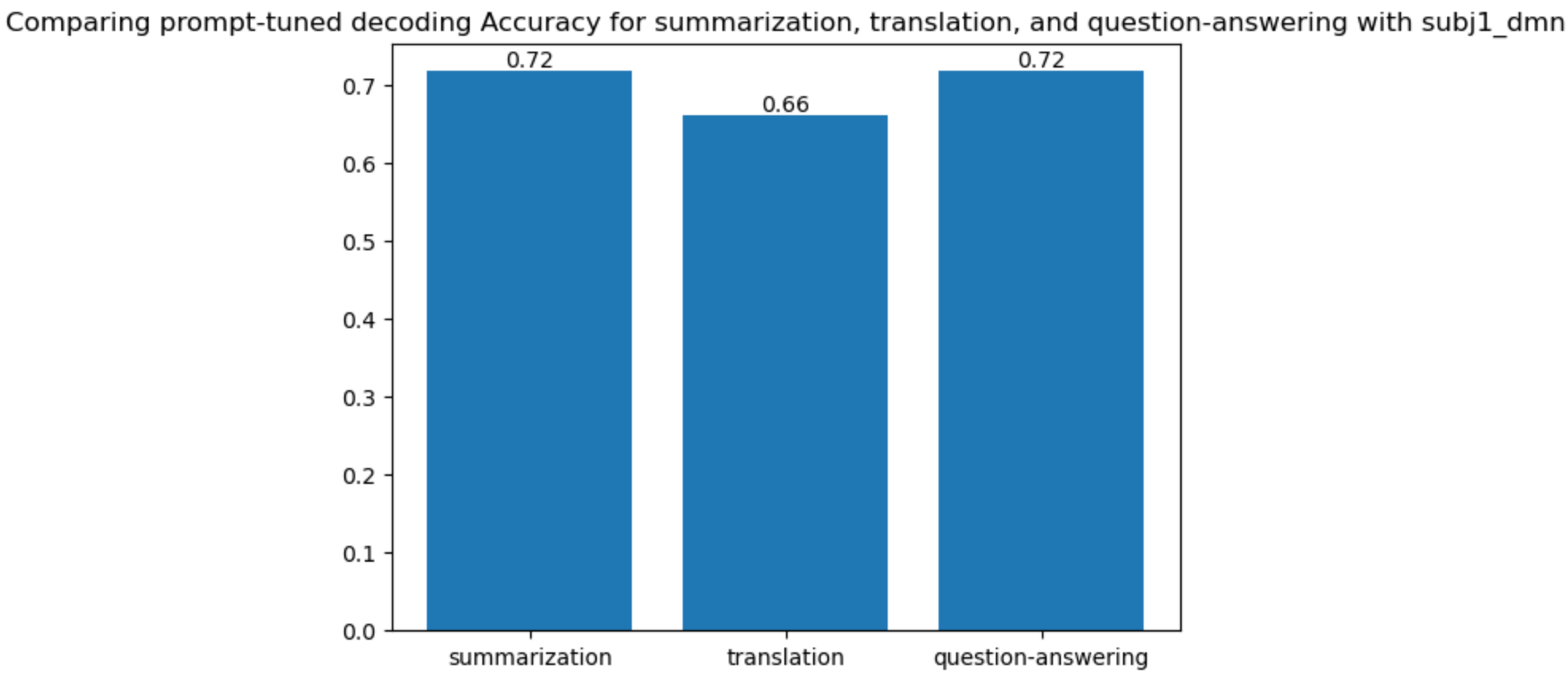
# DMN ROI

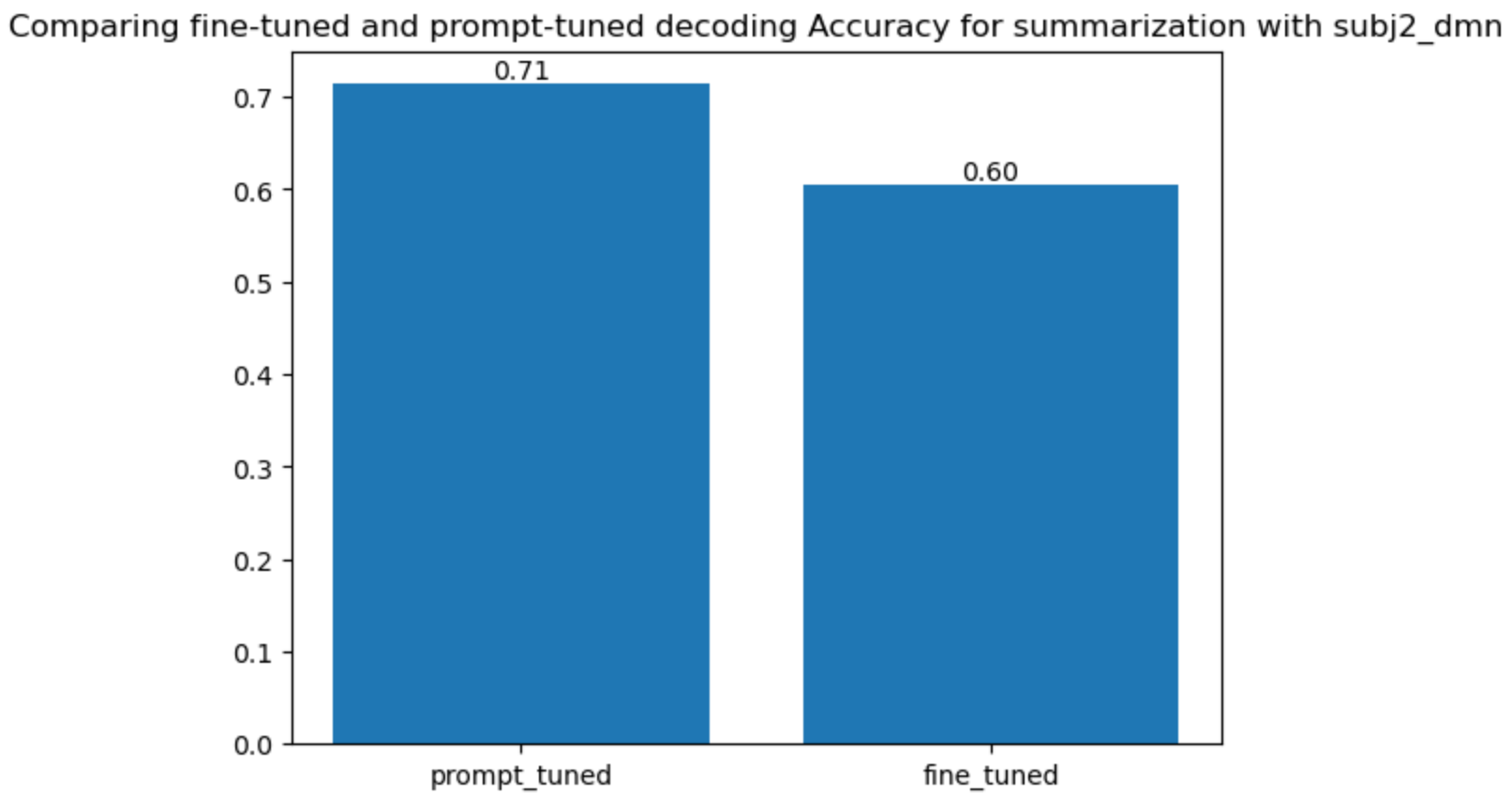

Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj1_dmn
0.72
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

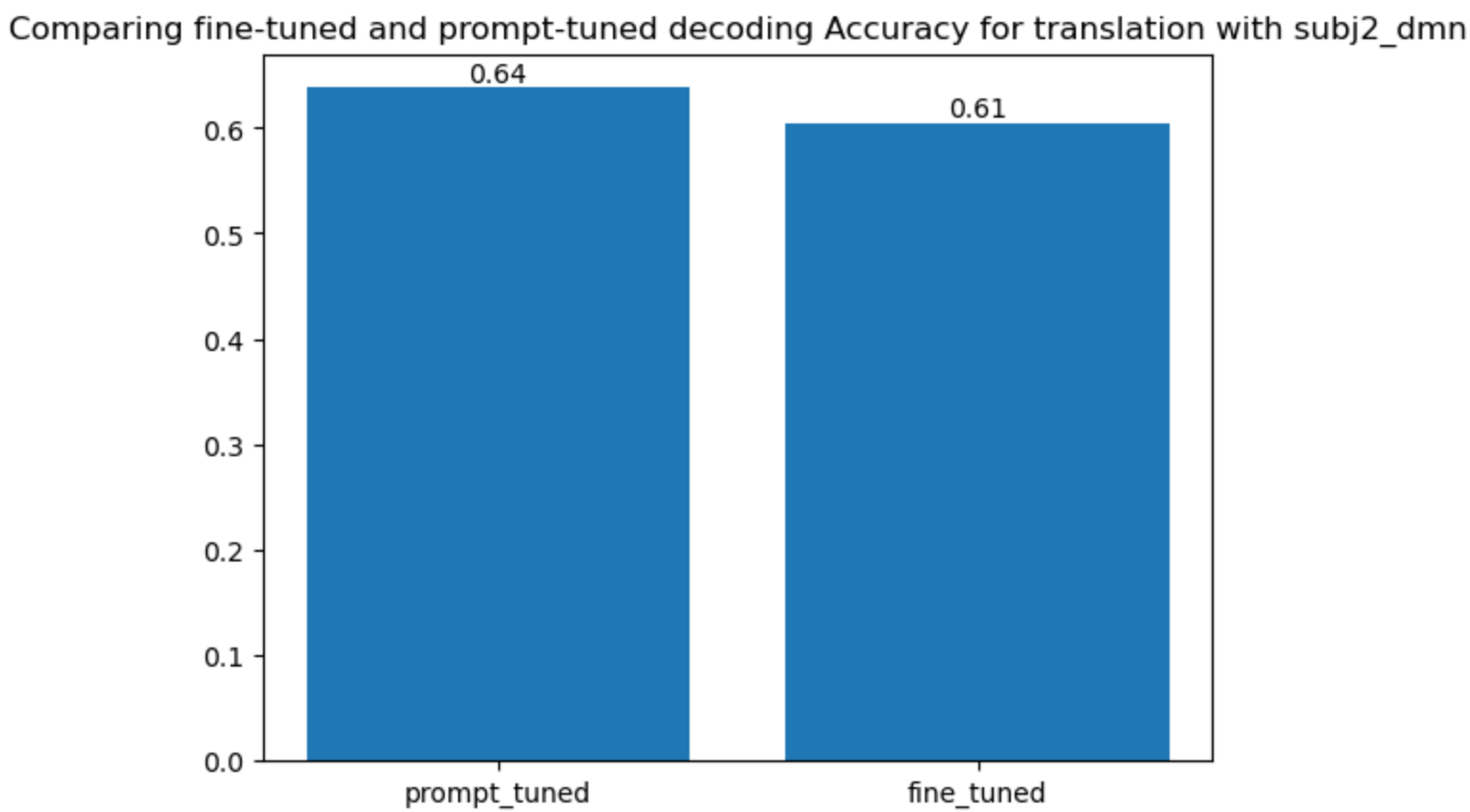

Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj1_dmn
0.66
0.63
0.0
0.1
0.2
0.3
0.4
0.5
0.6
prompt_tuned
fine_tuned

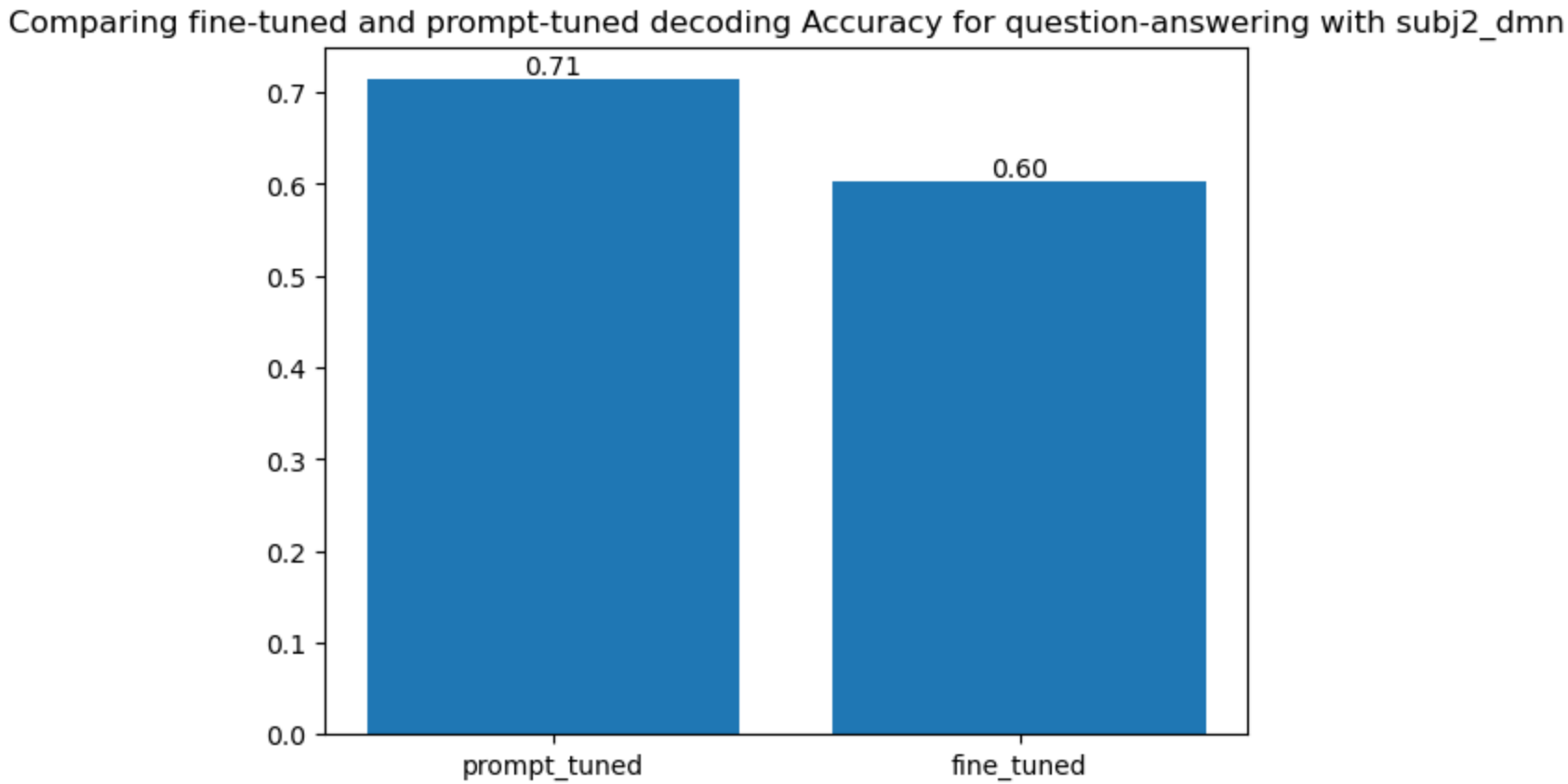

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj1_dmn
0.72
0.60
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

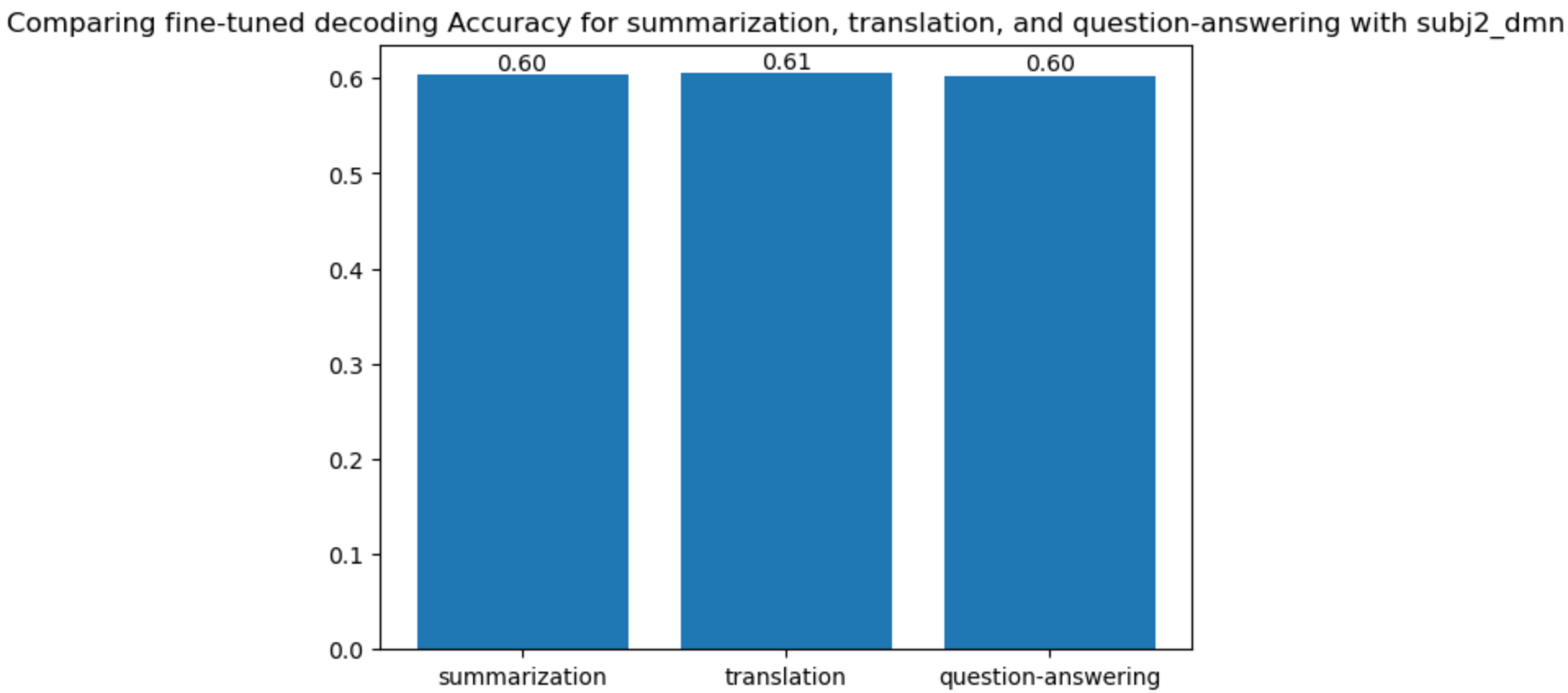

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_dmn
0.63
0.63
0.60
0.0
0.1
0.2
0.3
0.4
0.5
0.6
summarization
translation
question-answering

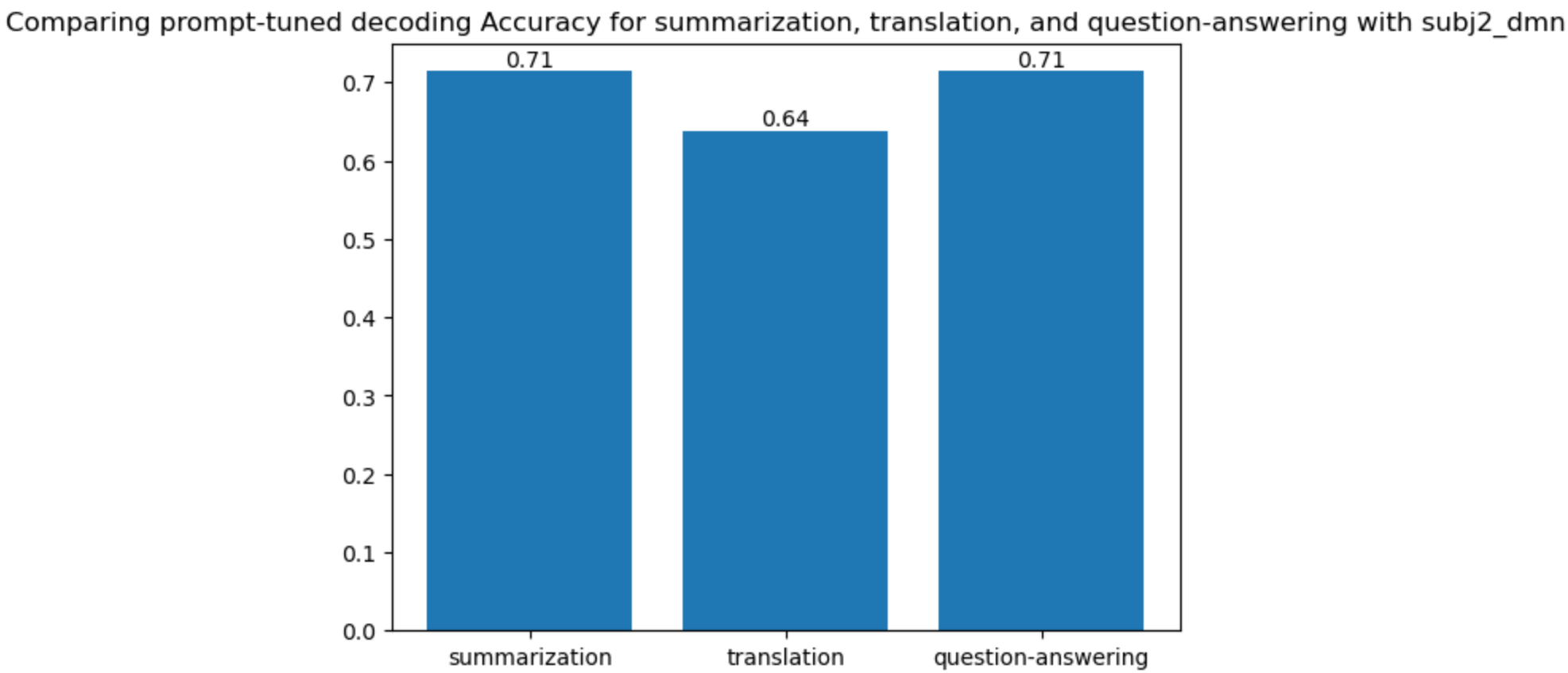

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj1_dmn
0.72
0.66
0.72
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned decoding Accuracy for summarization with subj2_dmn
0.71
0.60
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for translation with subj2_dmn
0.64
0.61
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

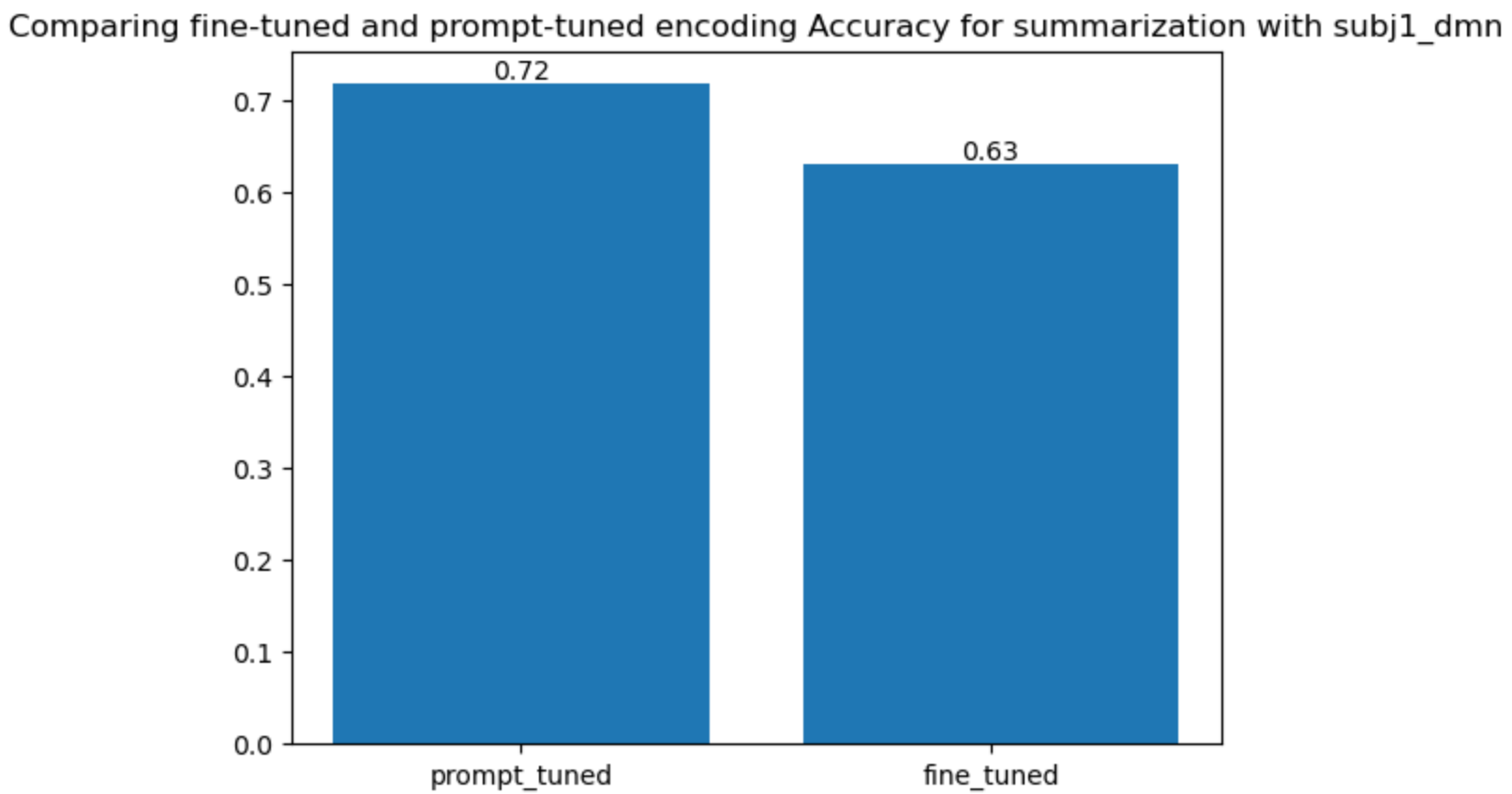

Comparing fine-tuned and prompt-tuned decoding Accuracy for question-answering with subj2_dmn
0.71
0.60
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

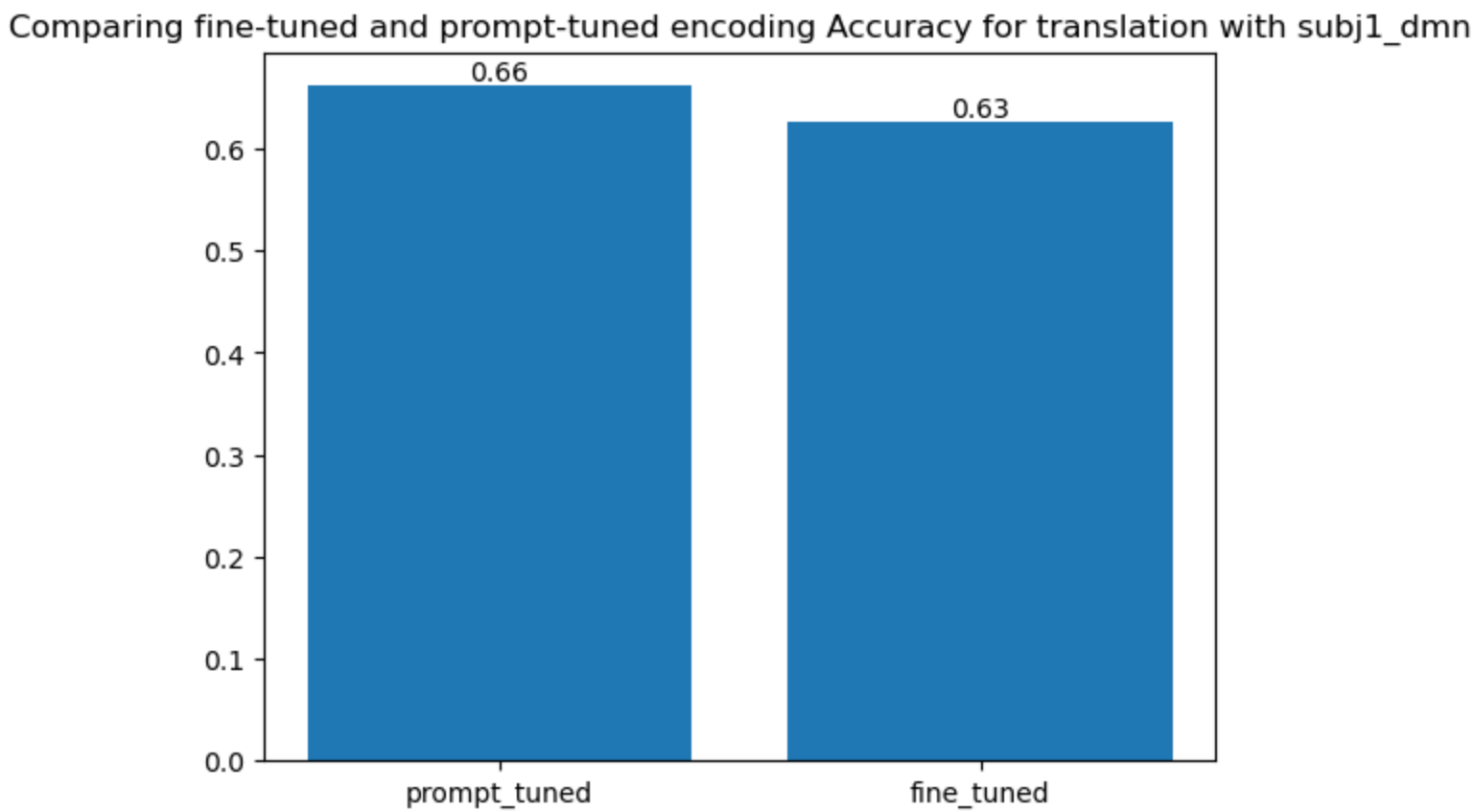

Comparing fine-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_dmn
0.60
0.61
0.60
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

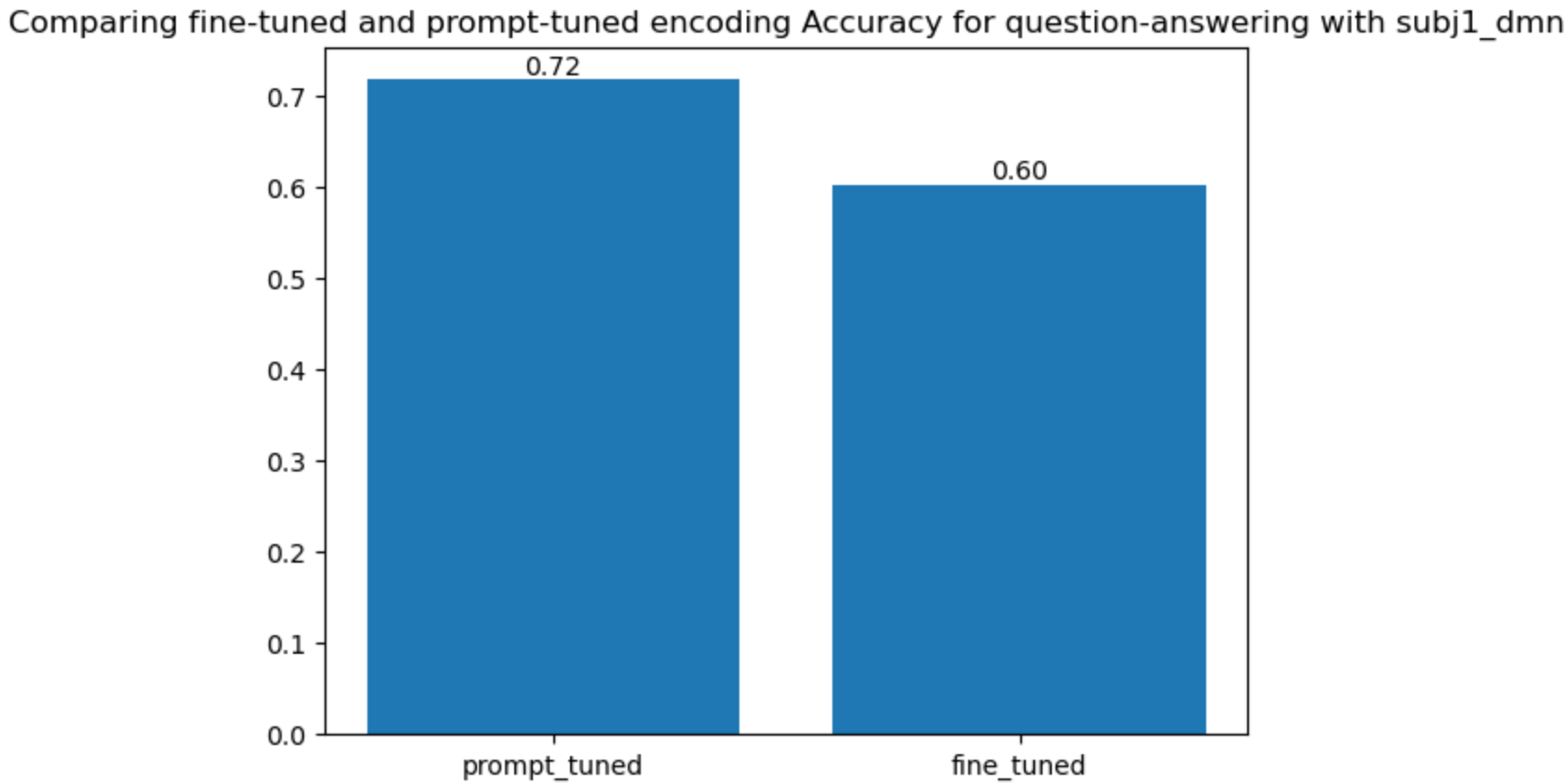

Comparing prompt-tuned decoding Accuracy for summarization, translation, and question-answering with subj2_dmn
0.71
0.64
0.71
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

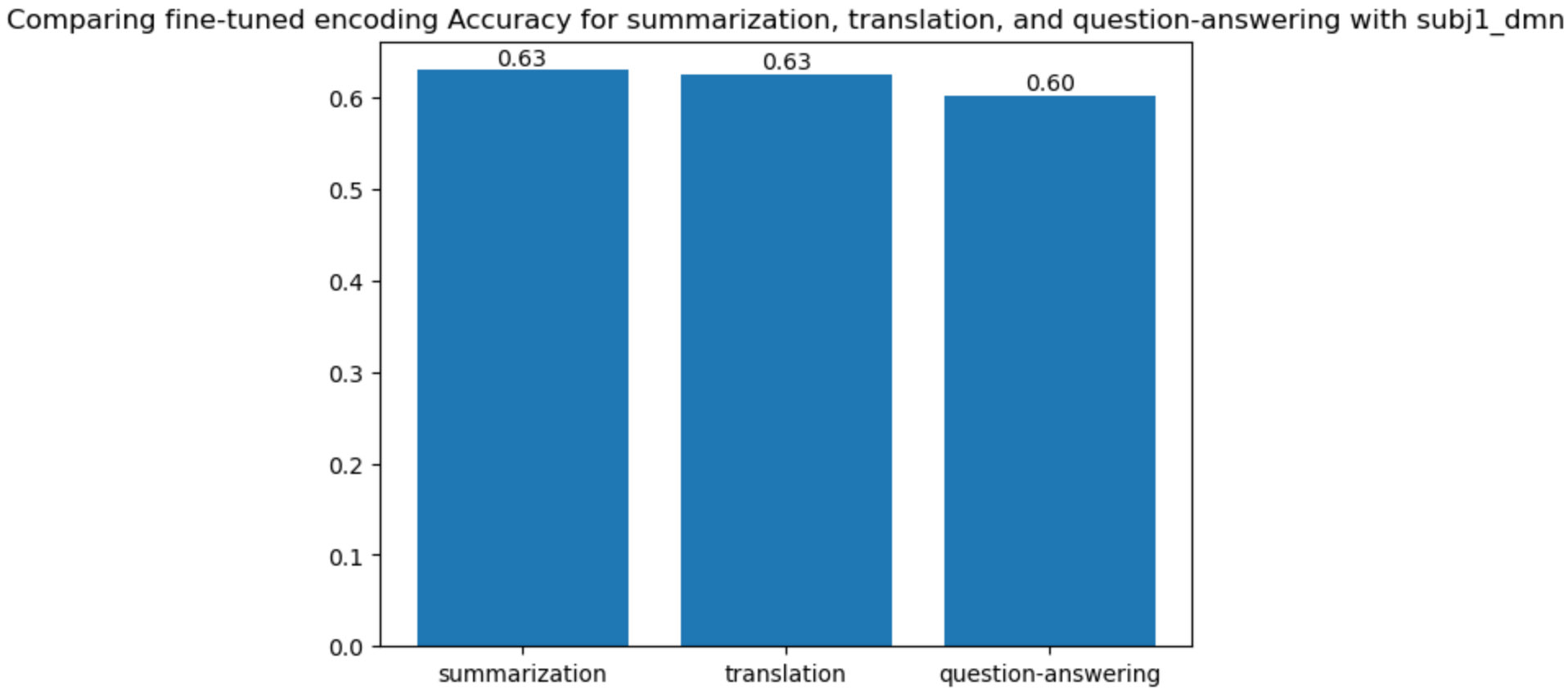
# Analysis of Prompt-tuned vs Fine-Tuned (Brain decoding)

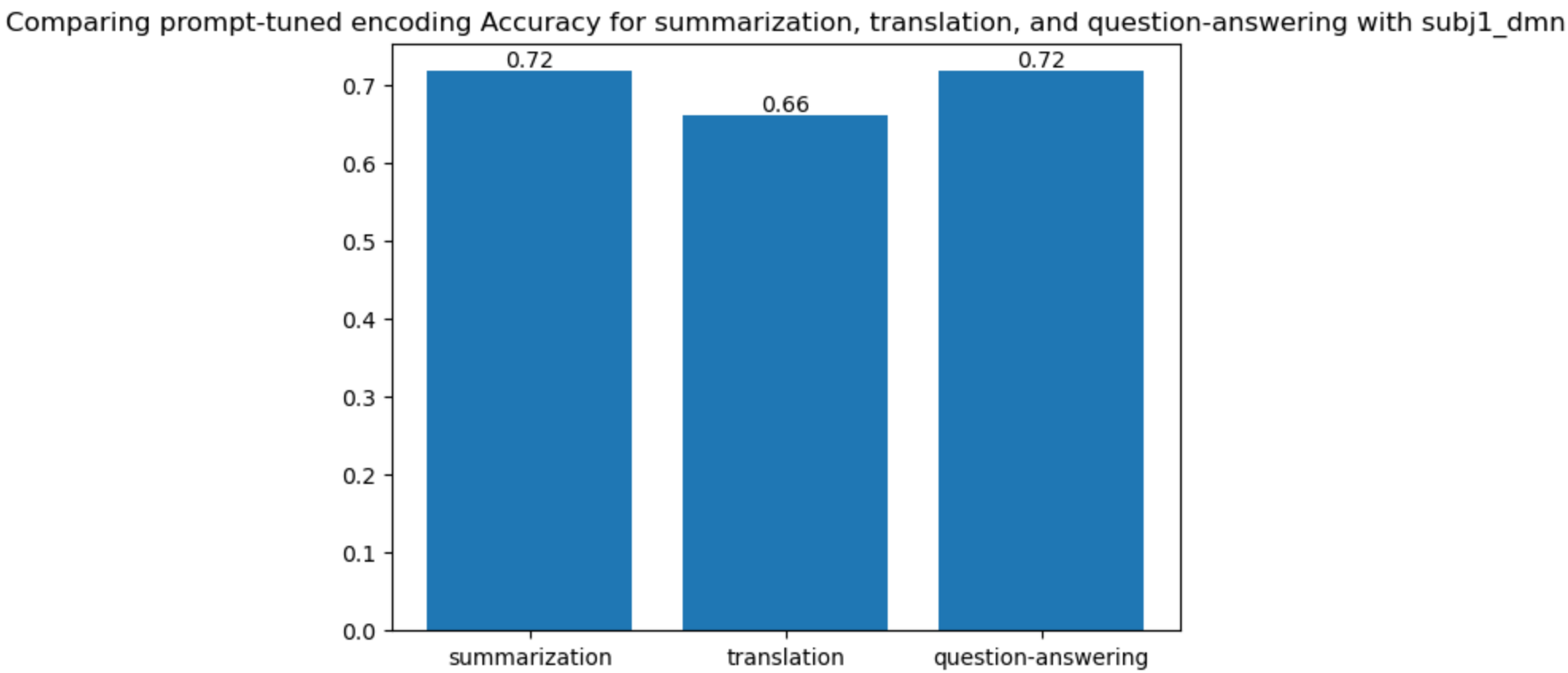
- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain decoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed.
- Wheras in Fine-tuned models, the MT model outperforms both the Sum and QA, not by a big margin tho!
- Also the results we got are **consistent** across subjects!

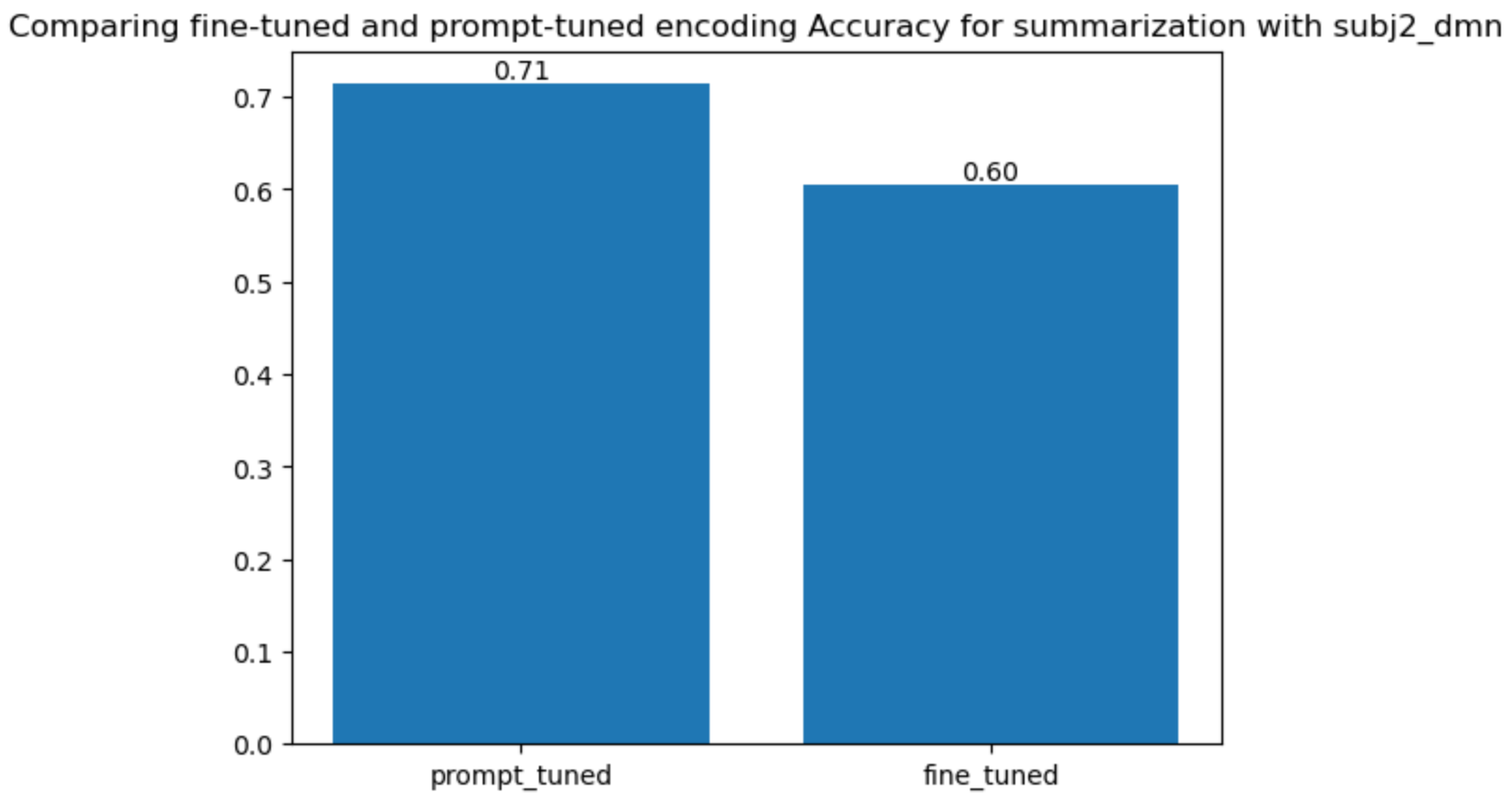

Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj1_dmn
0.72
0.63
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

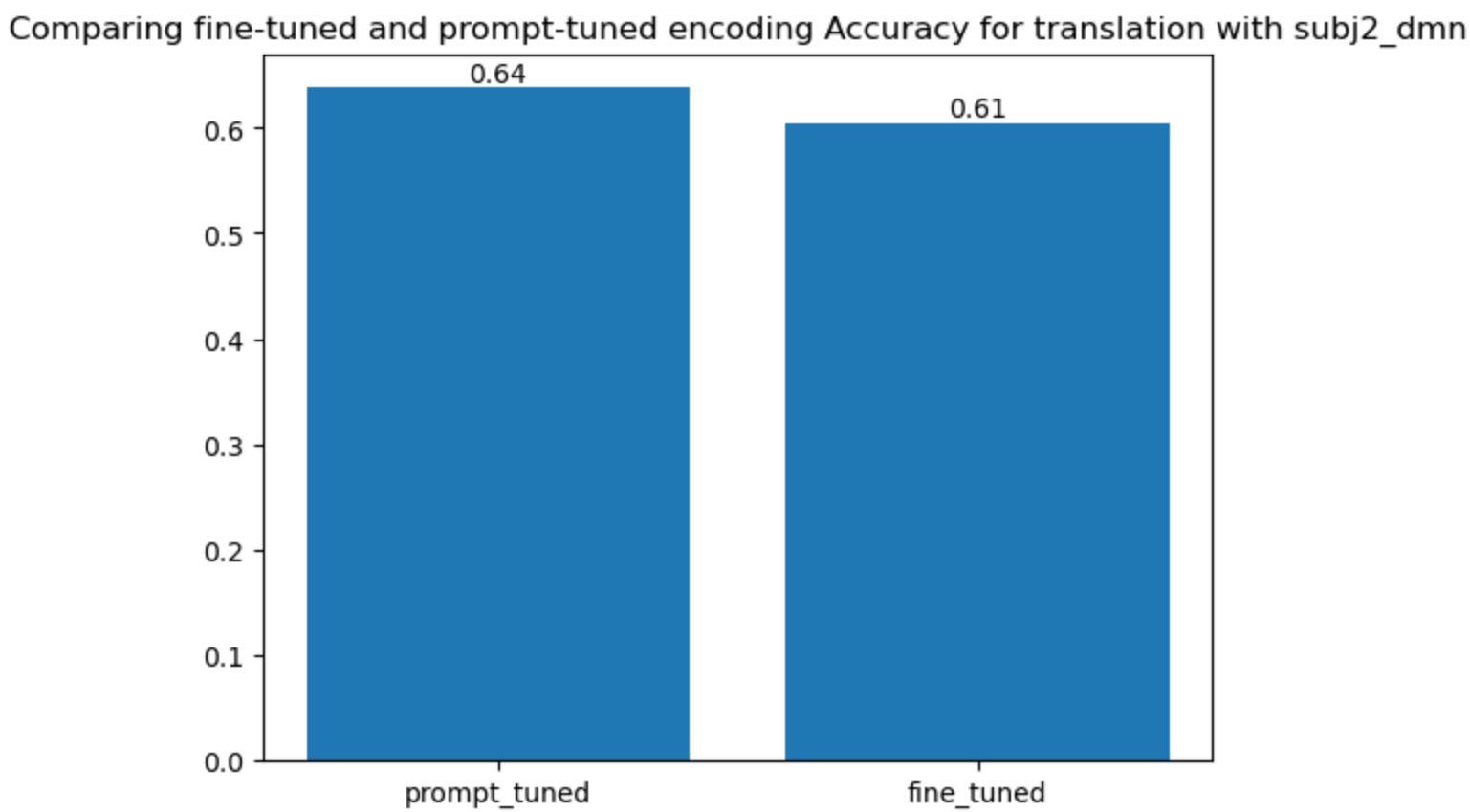

Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj1_dmn
0.66
0.63
0.0
0.1
0.2
0.3
0.4
0.5
0.6
prompt_tuned
fine_tuned

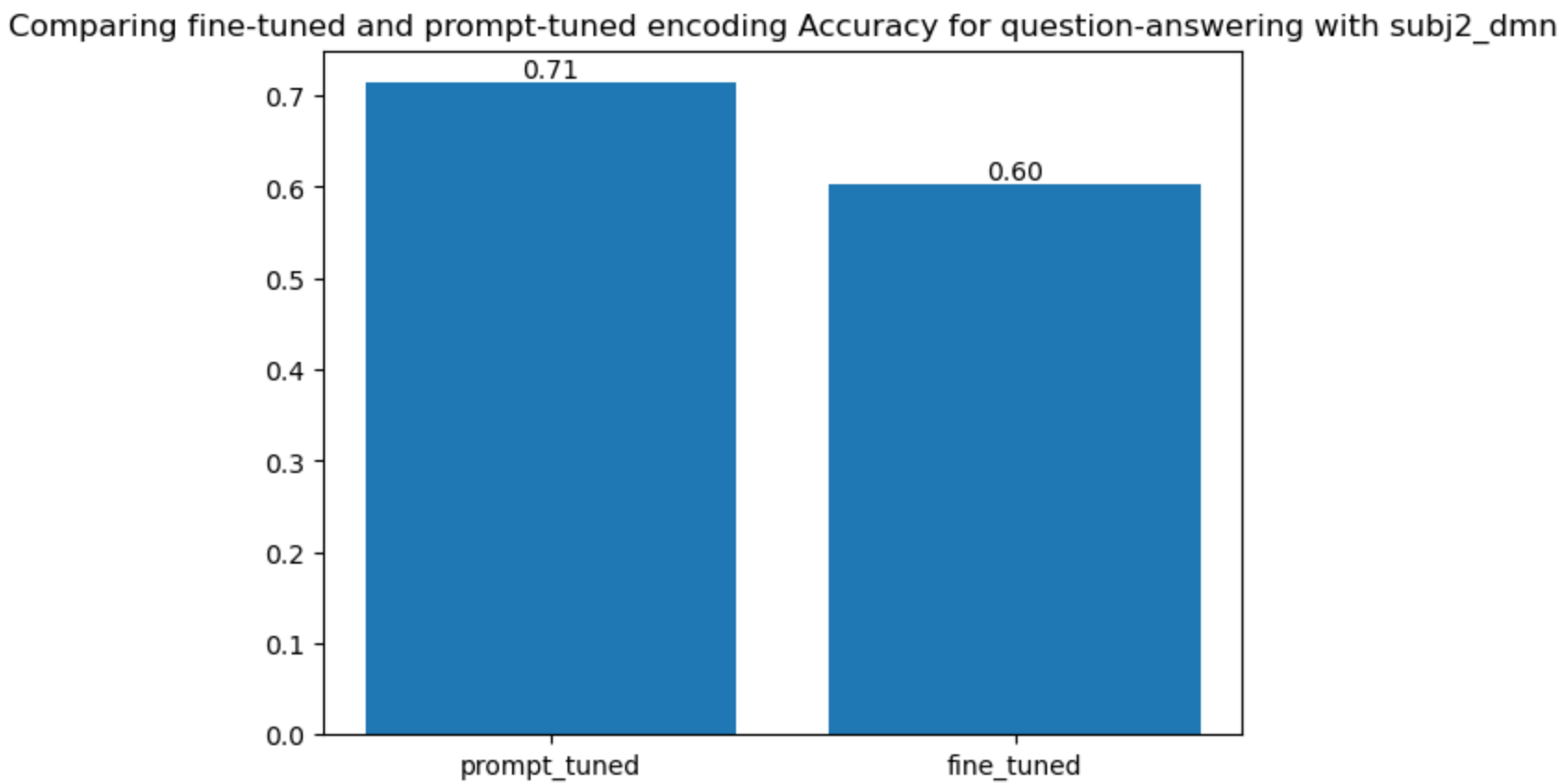

Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj1_dmn
0.72
0.60
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned

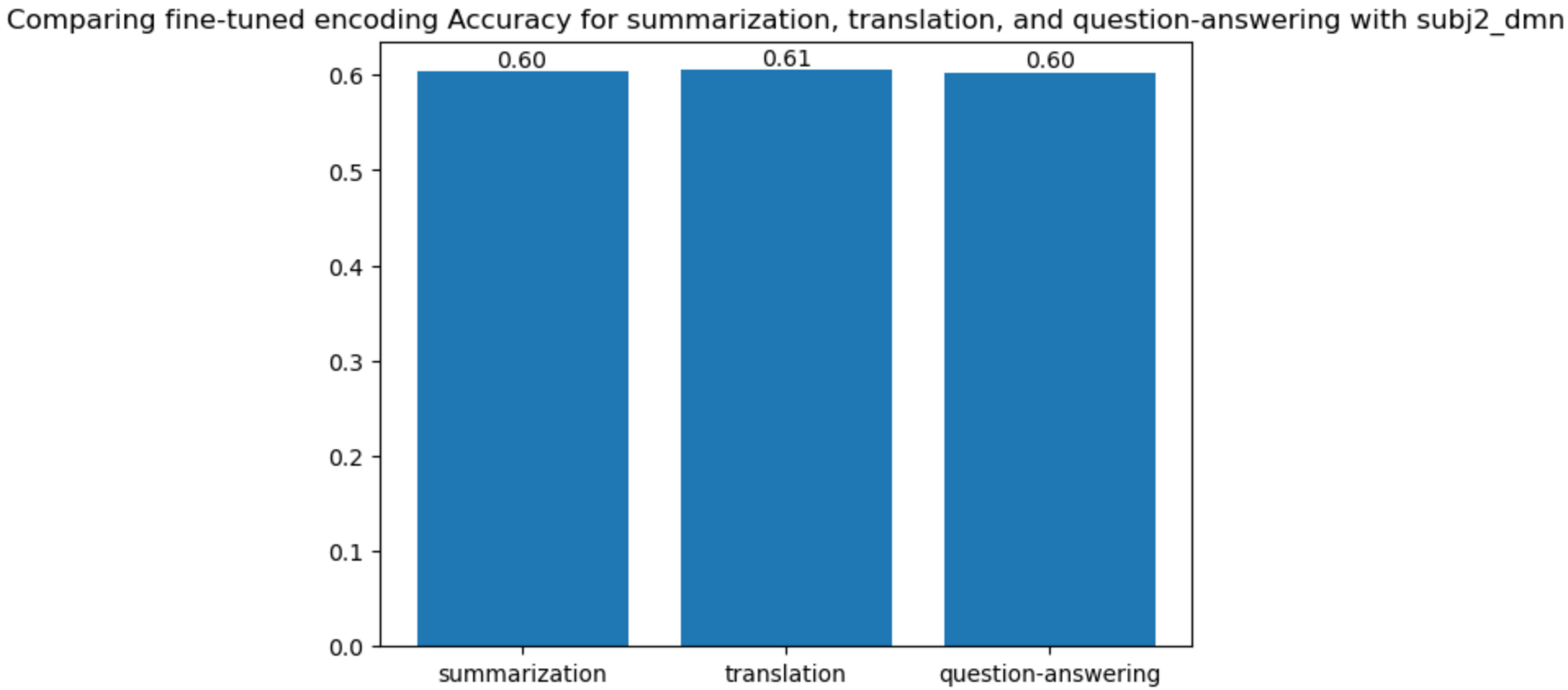

Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_dmn
0.63
0.63
0.60
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

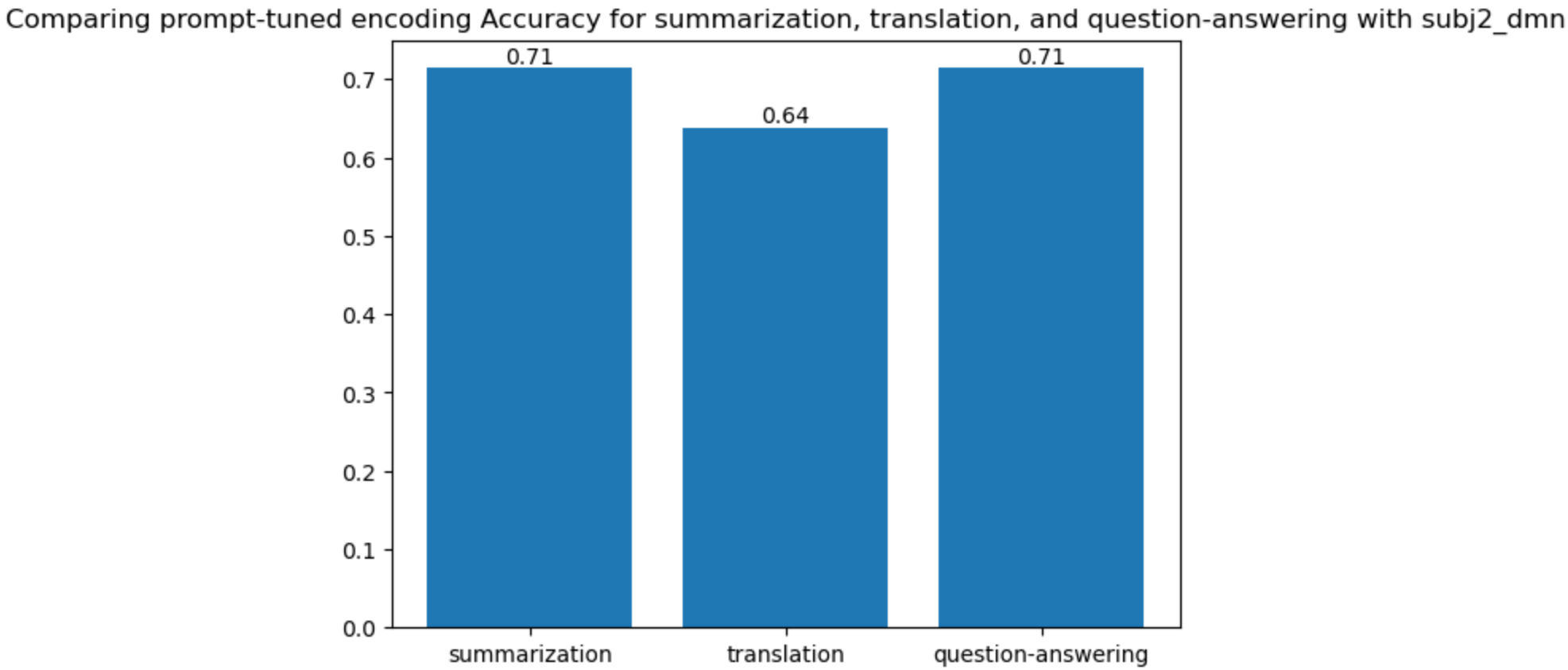

Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj1_dmn
0.72
0.66
0.72
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
summarization
translation
question-answering


Comparing fine-tuned and prompt-tuned encoding Accuracy for summarization with subj2_dmn
0.71
0.60
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for translation with subj2_dmn
0.64
0.61
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for question-answering with subj2_dmn
0.71
0.60
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
prompt_tuned
fine_tuned


Comparing fine-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_dmn
0.60
0.61
0.60
0.0
0.1
0.2
0.3
0.4
0.5
0.6
summarization
translation
question-answering


Comparing prompt-tuned encoding Accuracy for summarization, translation, and question-answering with subj2_dmn
0.71
0.64
0.71
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
summarization
translation
question-answering

# Analysis of Prompt-tuned vs Fine-Tuned (Brain encoding)

- Consistently Prompt-tuned GPT-2 better accounts for brain representations in **brain encoding** than Fine-tuned GPT-2.
- In Prompt-tuned models, the Sum and the QA models perform **equally well** wheras the MT model gets outperformed, this is consistent across subjects as well
- Wheras in Fine-tuned models, the MT model outperforms the other 2, but not by a big margin.
- Also, the results overall are consistent across subjects.

# VISION

Novel contribution

# Visual brain encoding and decoding

- Given an image, getting the brain representations is encoding and vice-versa is decoding.
- Using BOLD5000 dataset, we have images and the corresponding brain representations.
- Two questions:
  - Which model to fine-tune and prompt-tune on, now for vision?
  - What are some datasets we choose, to fine-tune or prompt-tune them?

# Using Vit Base (Google)

- There is not a discussion of decoder only models for vision especially for brain encoding and decoding.
- It is a BERT-like model, so we can get the image features out of this ViT Base model.
- The original paper uses BERT for text and uses classification tasks because generation can't be possible using encoder models.
- Going by the same analogy, two classification tasks, one of hand-written and one of object detection have been used.
- The rest reasons for GPT-2 follow here as well!

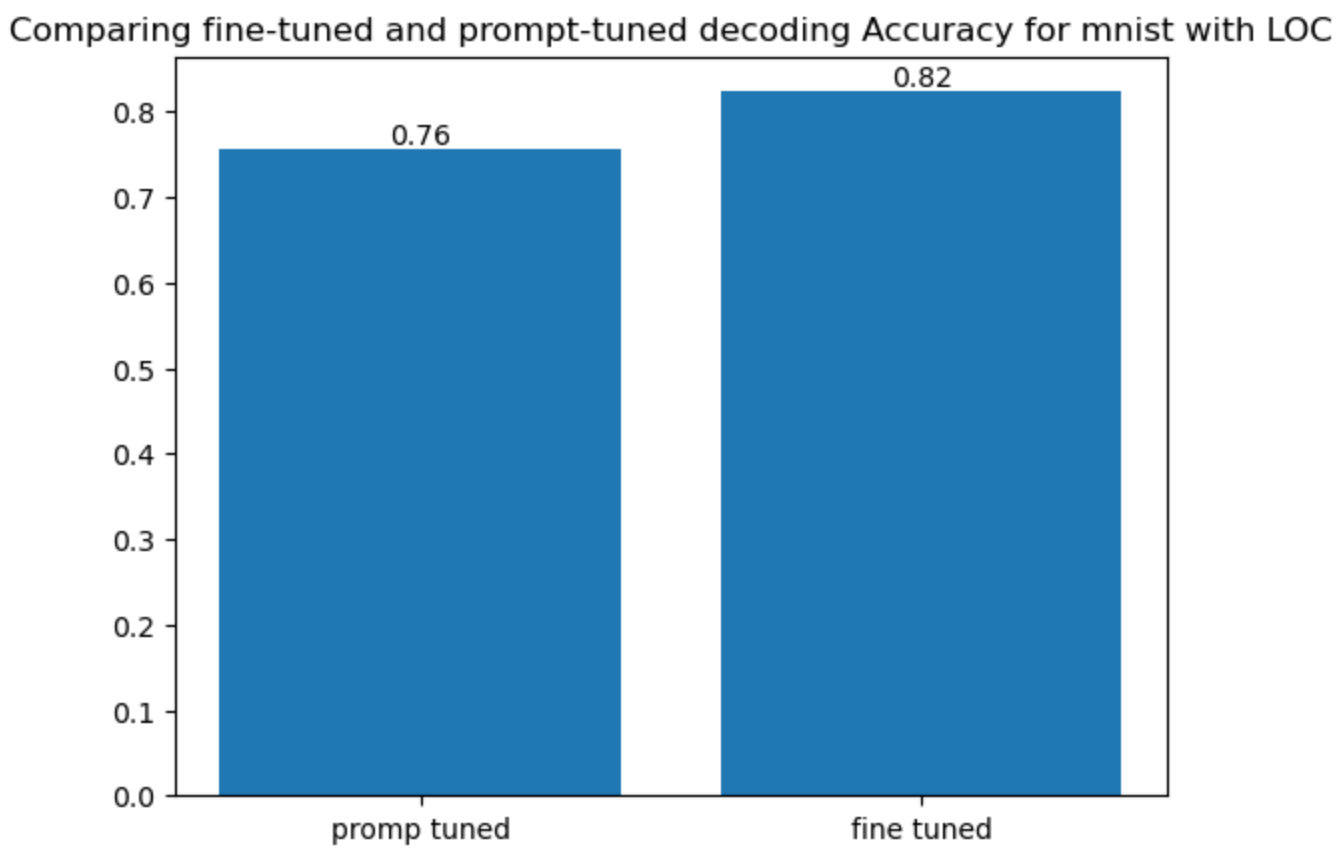
# Datasets chosen

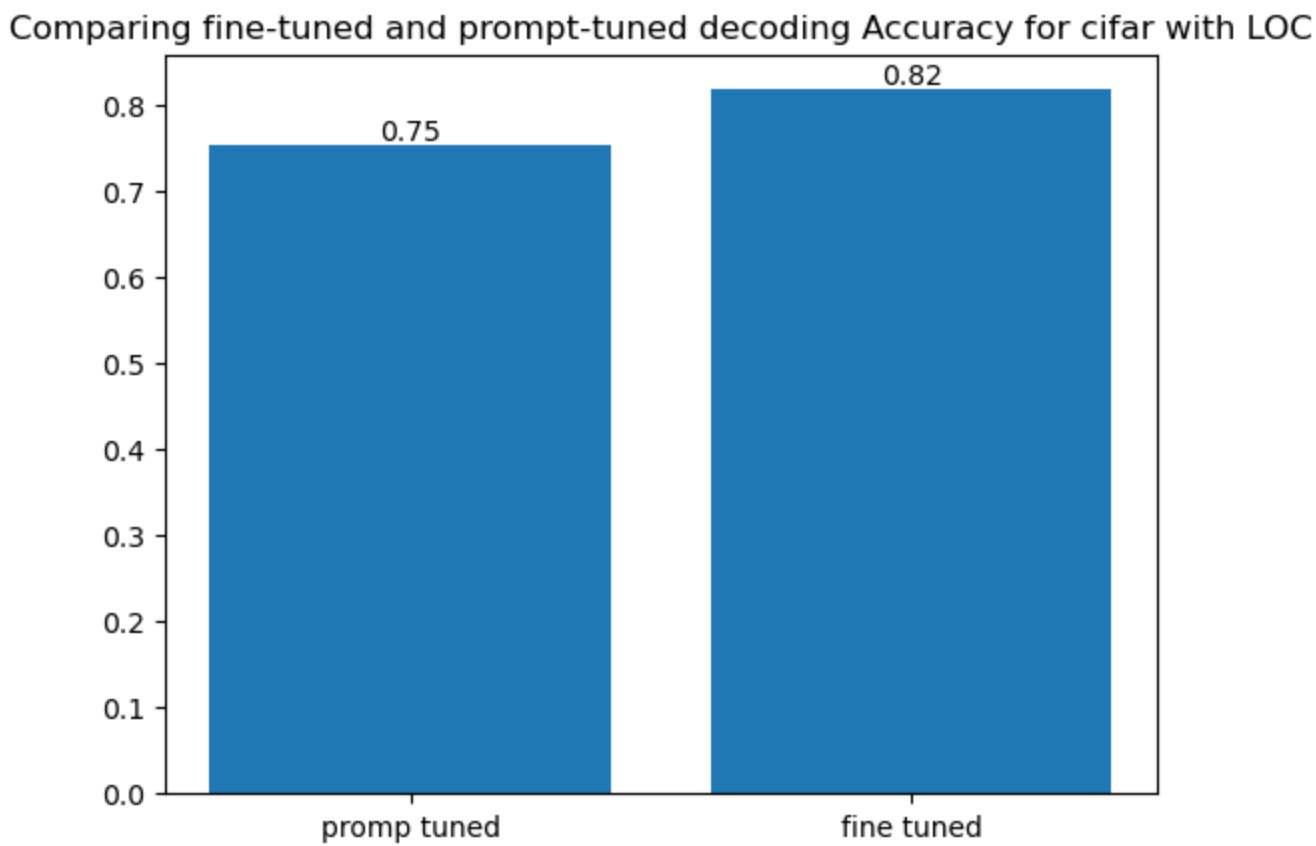
- MNIST, CIFAR-10
- Why?
  - Much of it is mentioned in the previous slide
  - To bring out a contrast between object detection (CIFAR-10) and plain number detection (MNIST)

# Methodology

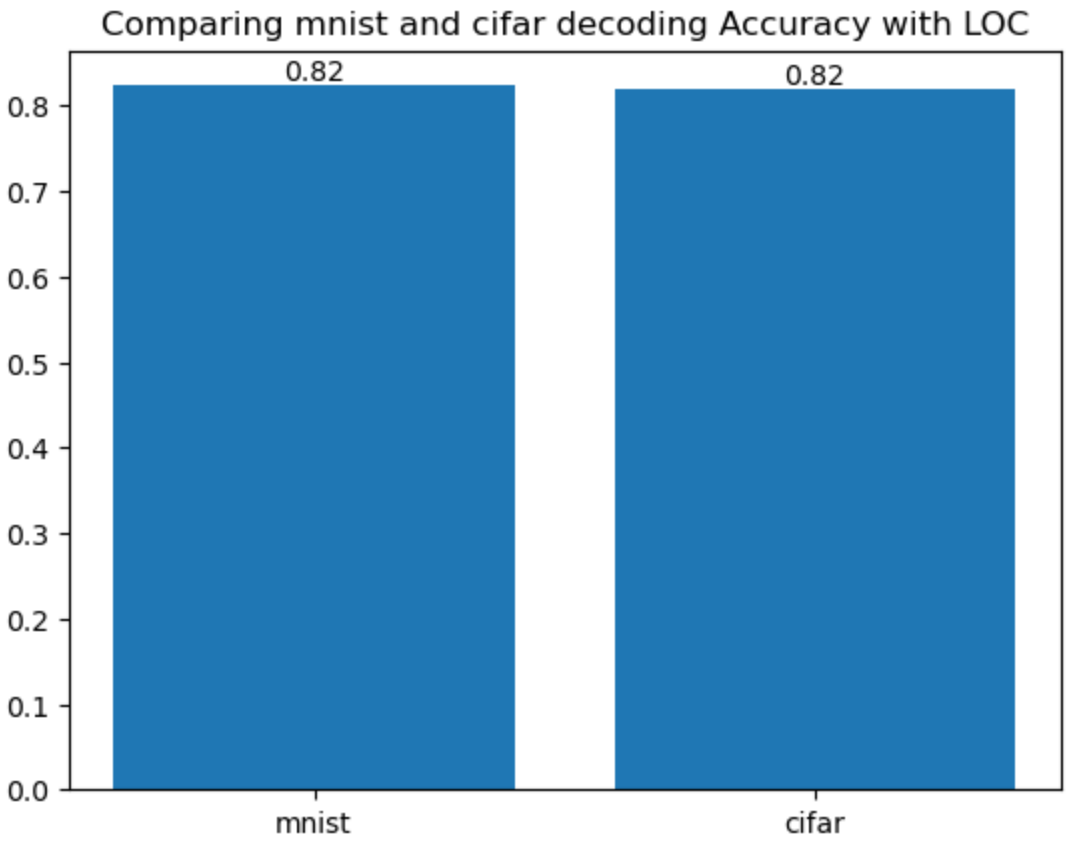
- We fine-tune and prompt-tune both using ViT Base on both the MNIST and CIFAR-10 datasets.
- Train Neural encoders and decoders to account for correlation between brain representations and image representations.
- Do the same task ROI wise and do an overall analysis.
- We are using the standard **2v2 accuracy metric** for all our analysis.

# LOC ROI

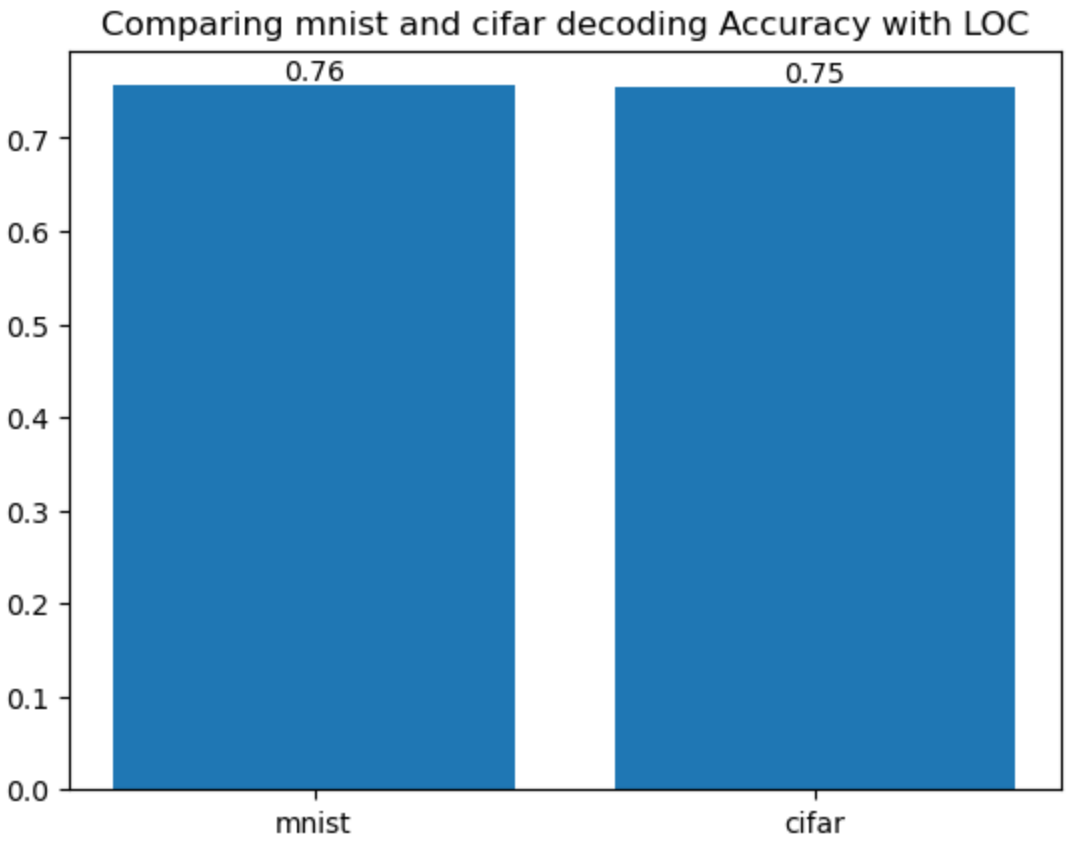

Comparing fine-tuned and prompt-tuned decoding Accuracy for mnist with LOC
0.82
0.76
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for cifar with LOC
0.82
0.75
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned

# Fine-tuning

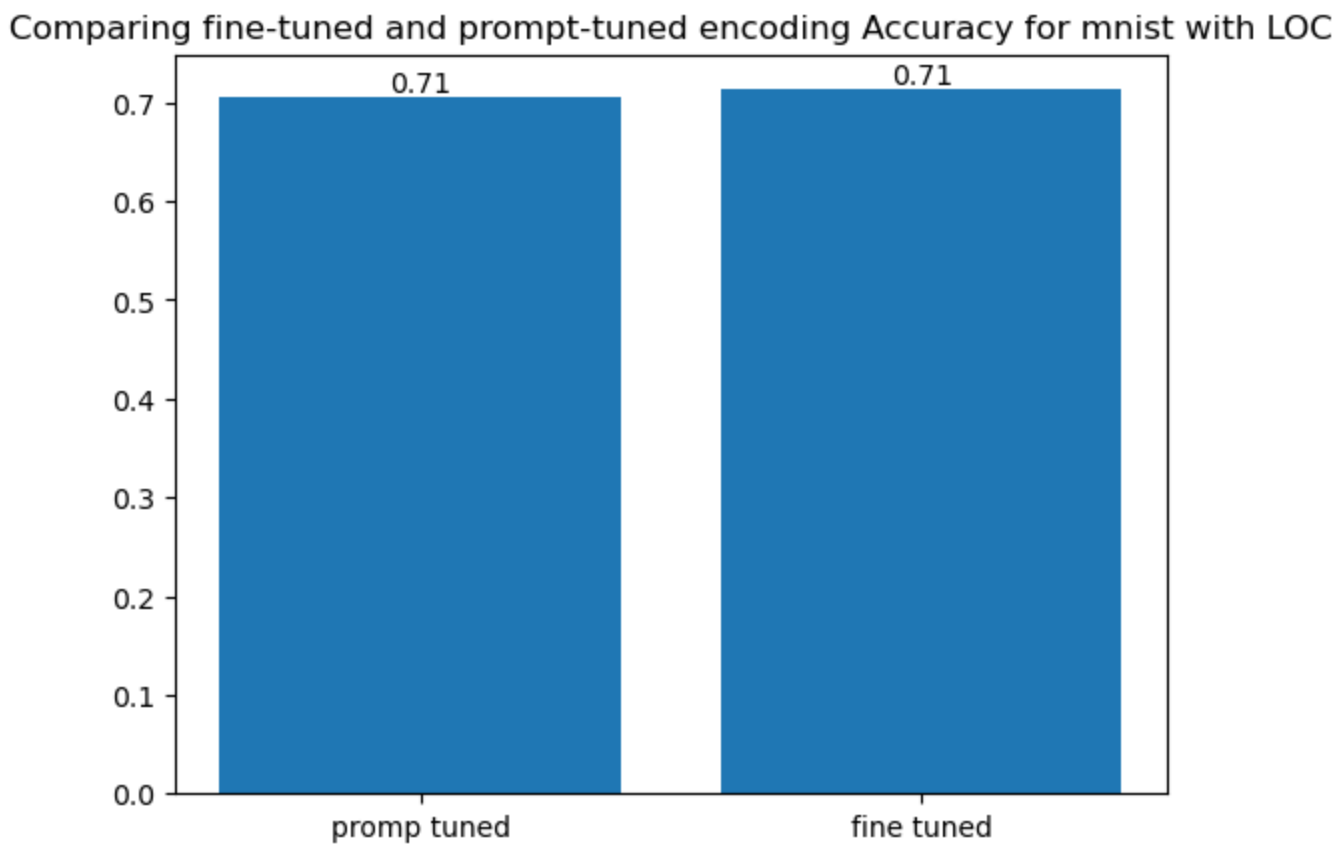

Comparing mnist and cifar decoding Accuracy with LOC
0.82
0.82
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mnist
cifar

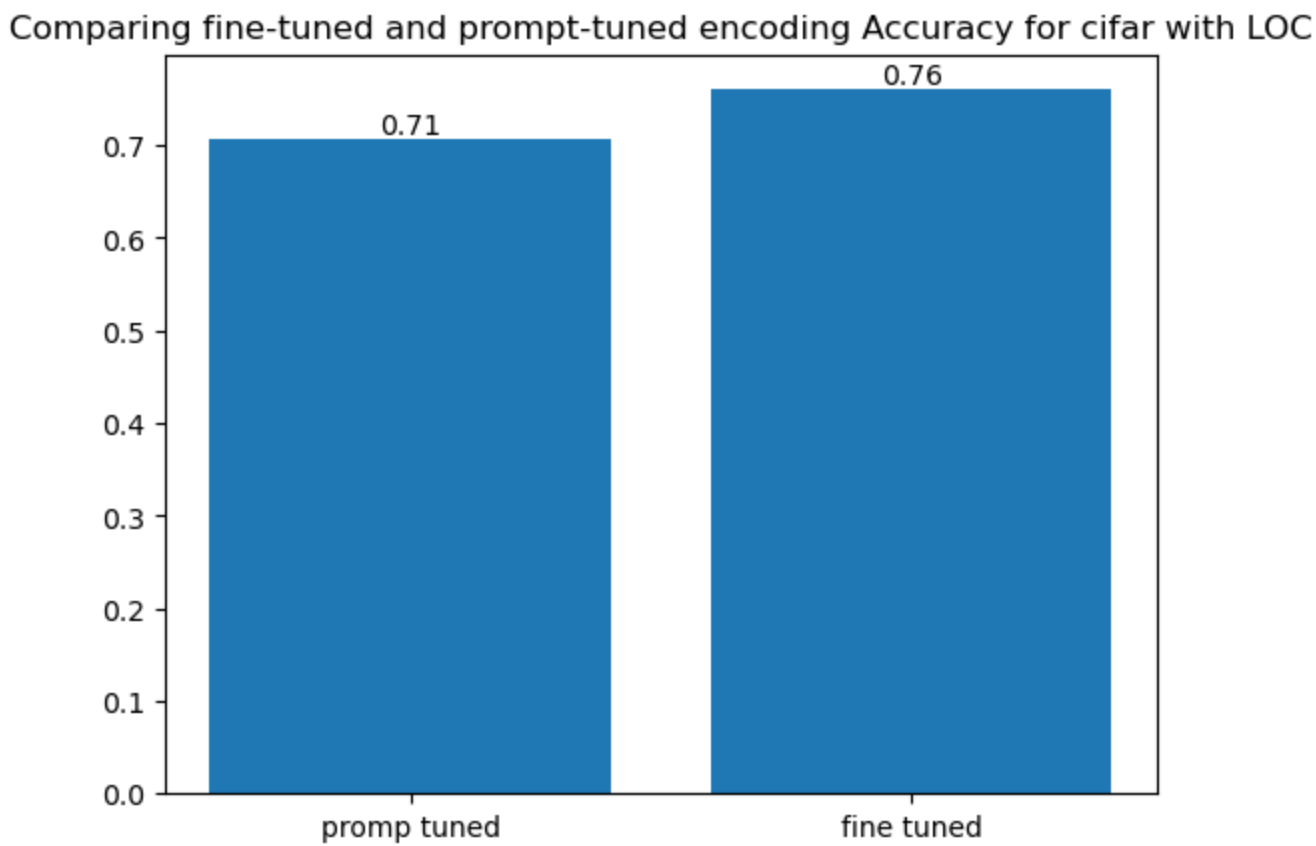
# Prompt-tuning



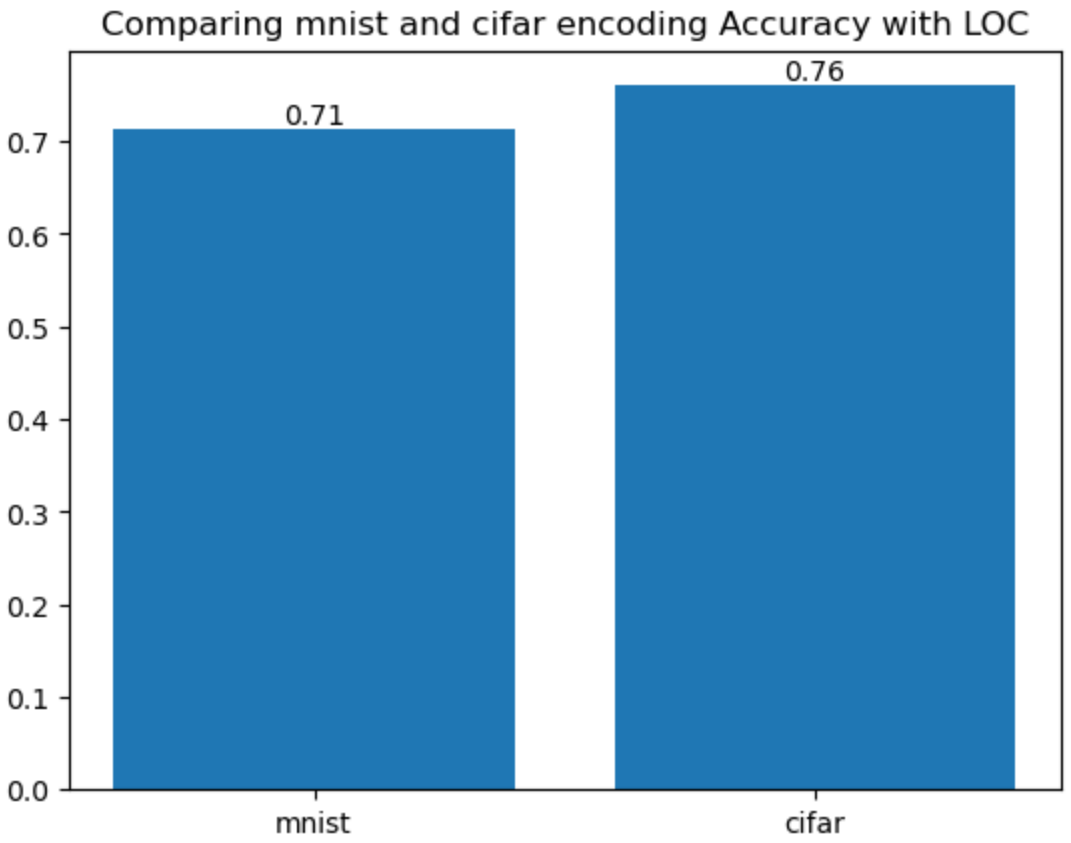
# Prompt-tuning vs Fine-tuning (Brain Decoding)

- To our surprise, with a clear trend, **fine-tuning models outperform prompt-tuning models.**
- The prompt-tuning, whose reason is unknown but gives the replicate trend of the text data.
- Also both the prompt-tuned models, one on MNIST and CIFAR give the same accuracy, indicating not much of difference in comprehension of written text and objects.

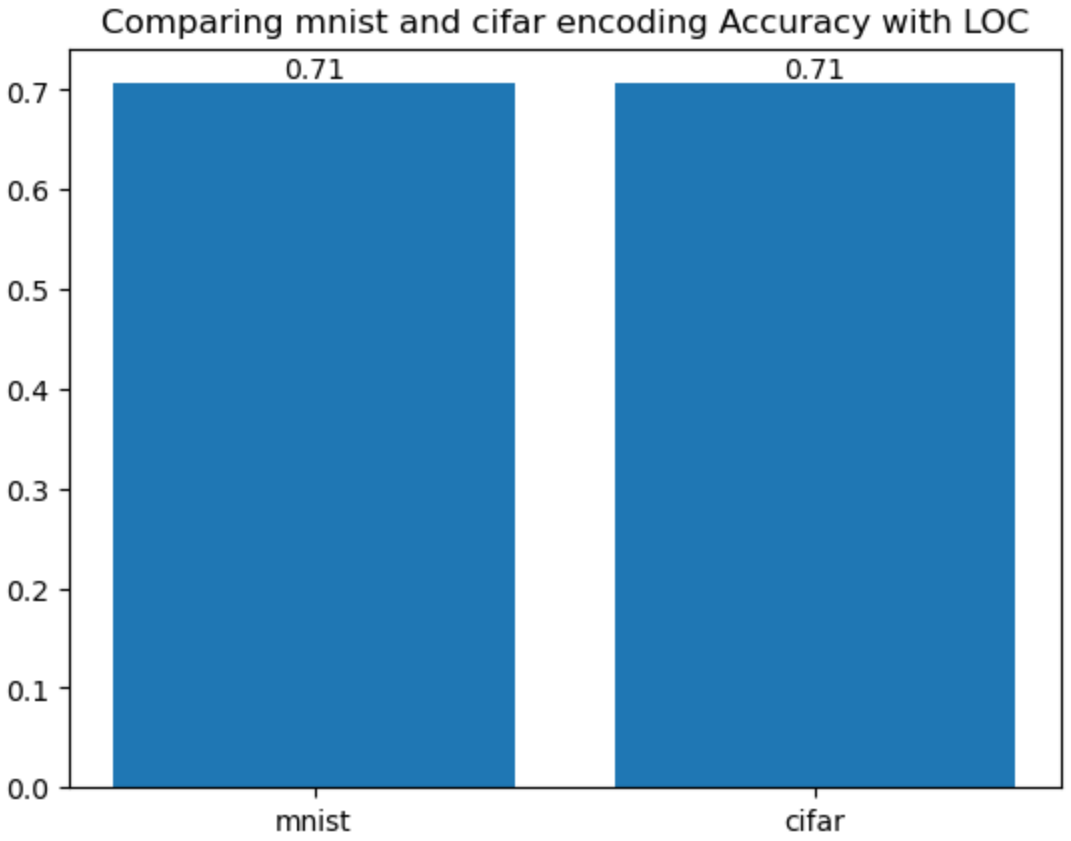

Comparing fine-tuned and prompt-tuned encoding Accuracy for mnist with LOC
0.71
0.71
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for cifar with LOC
0.71
0.76
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
promp tuned
fine tuned

# Fine-Tuning


Comparing mnist and cifar encoding Accuracy with LOC
0.71
0.76
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
mnist
cifar

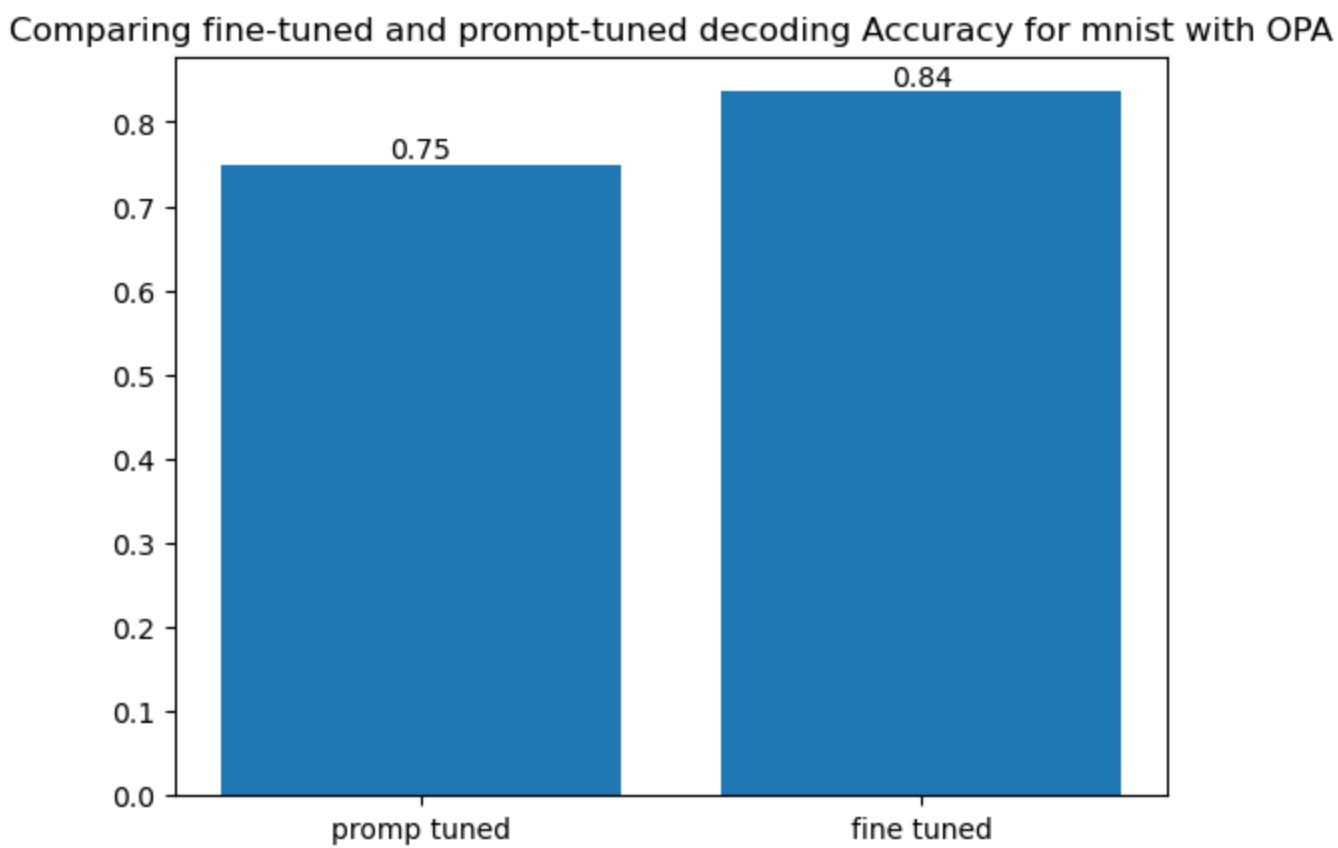
# Prompt-tuning

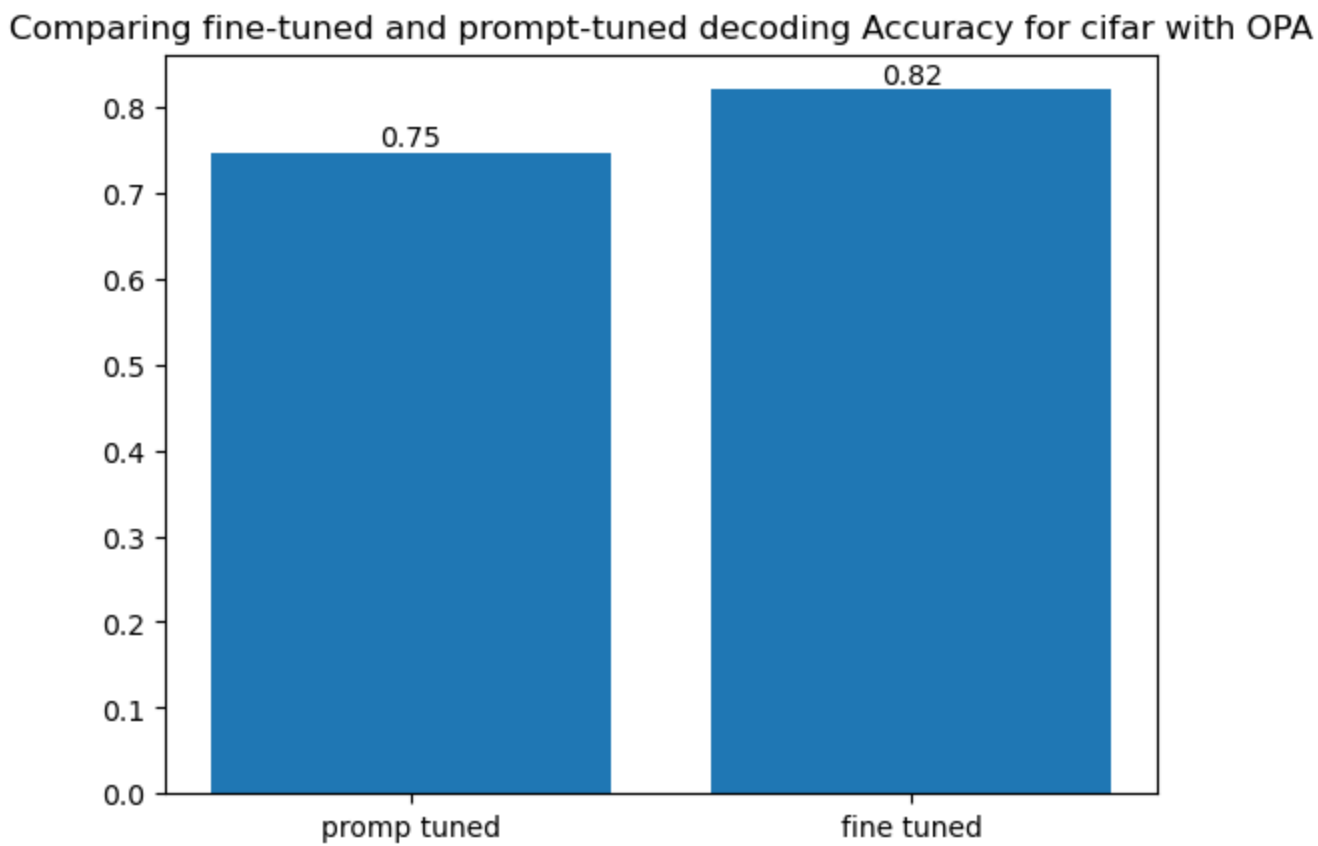


# Prompt-tuning vs Fine-tuning (Brain Encoding)

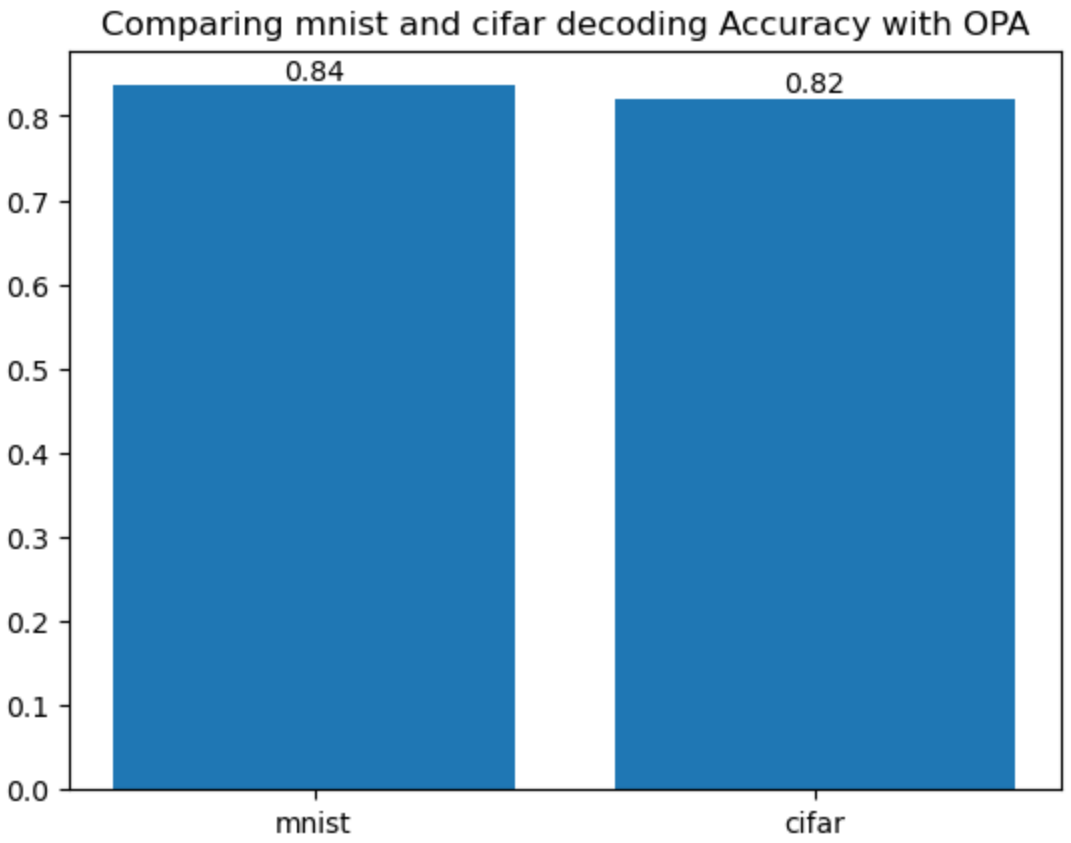
- Though the MNIST dataset doesn't signify a difference, the CIFAR dataset pronounces that ViT Base Fine-tuned on CIFAR better accounts for brain representations than ViT prompt tuned on CIFAR.
- Again the trend of fine-tuned models performing better continues.

# OPA ROI

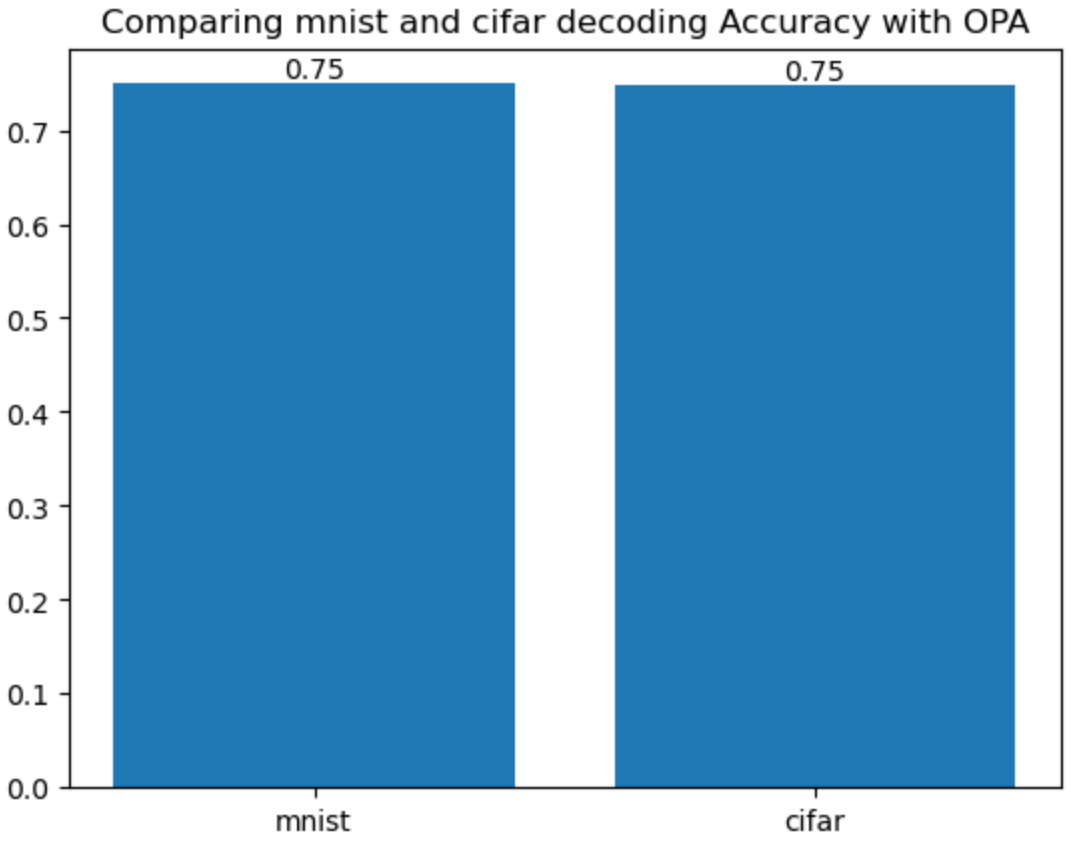

Comparing fine-tuned and prompt-tuned decoding Accuracy for mnist with OPA
0.84
0.75
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for cifar with OPA
0.82
0.75
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned

# Fine-tuning

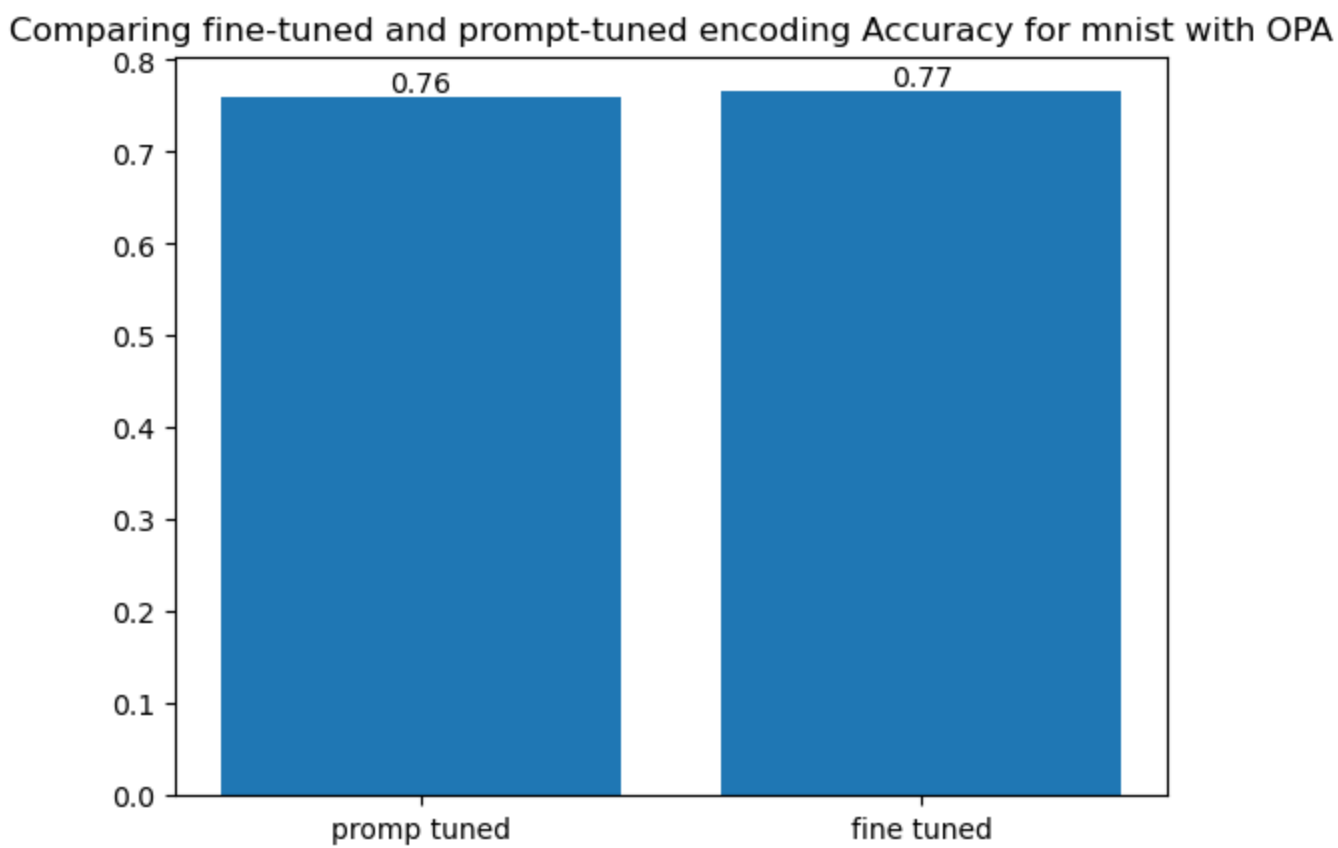

Comparing mnist and cifar decoding Accuracy with OPA
0.84
0.82
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mnist
cifar

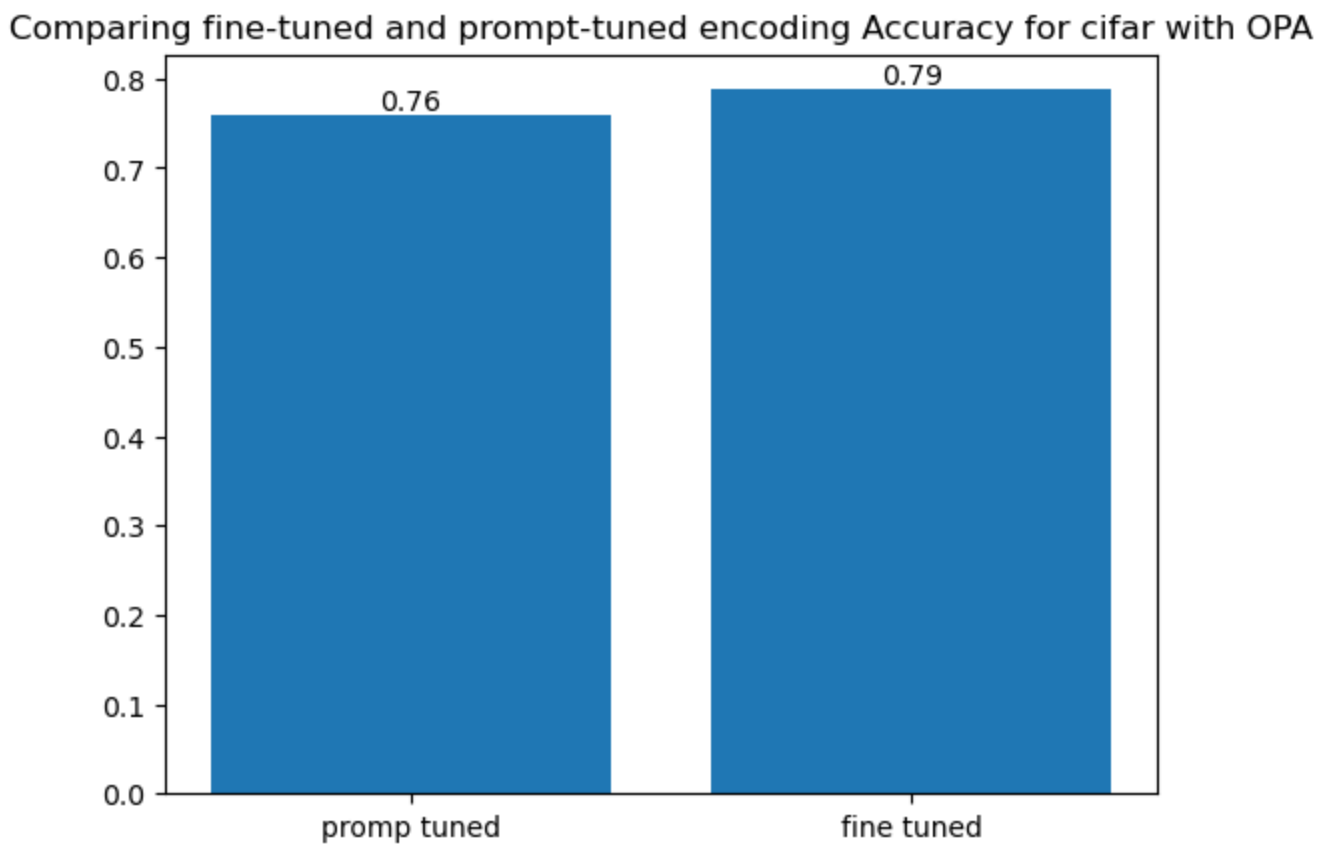
# Prompt-tuning



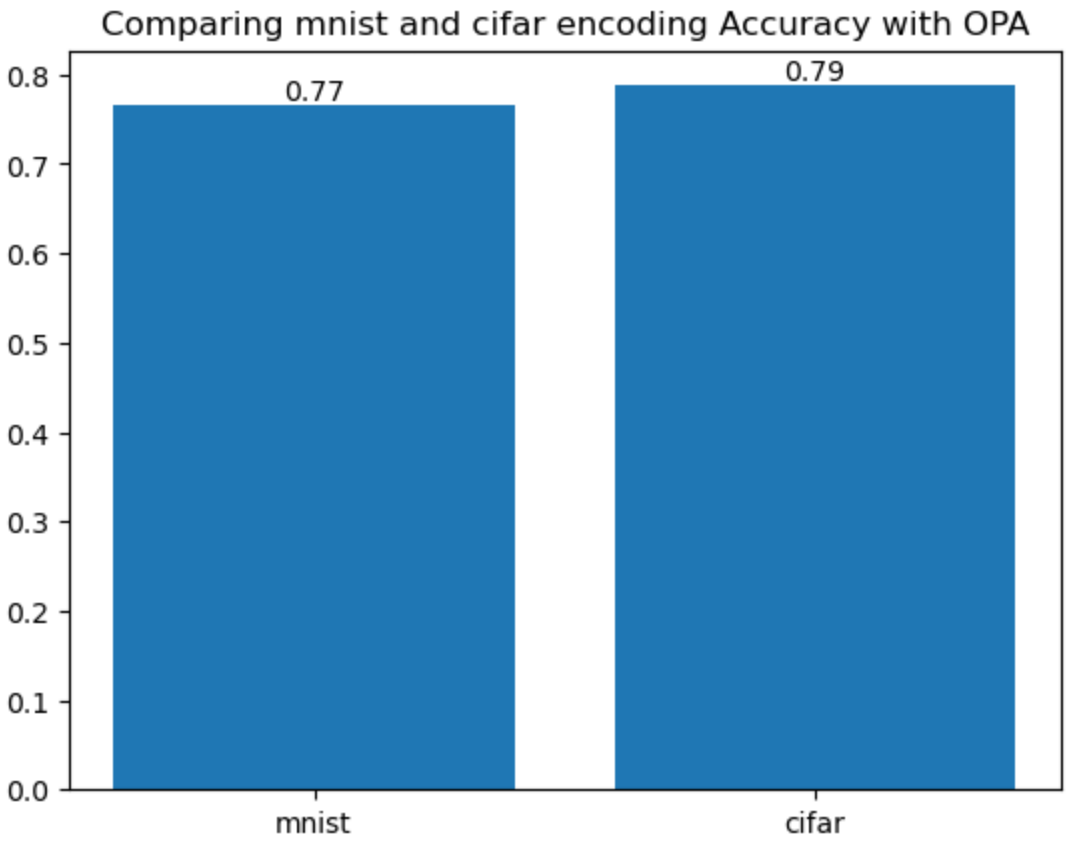
# Prompt-tuning vs Fine-tuning (Brain Decoding)

- With a clear trend, **fine-tuning models outperform prompt-tuning models.**
- The MNIST model seems to outperform the CIFAR model in brain decoding the images from BOLD5000 dataset
- Also both the prompt-tuned models, one on MNIST and CIFAR gave the same accuracy.

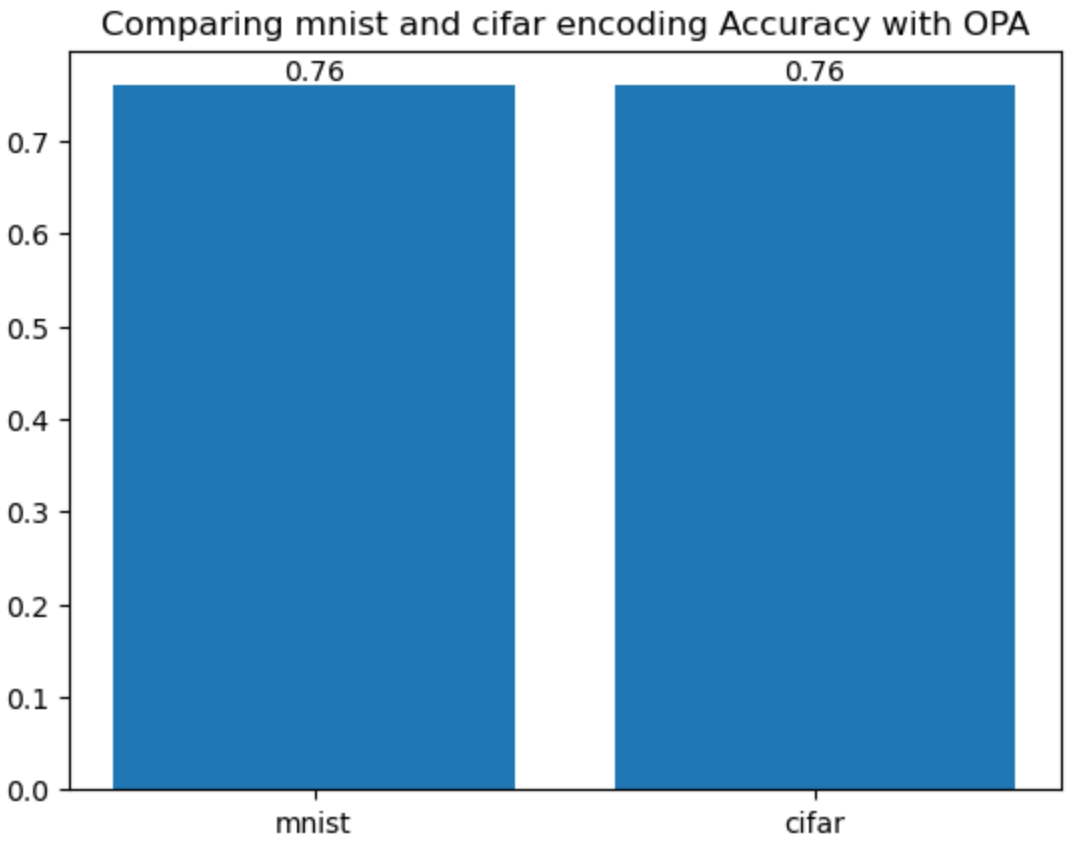

Comparing fine-tuned and prompt-tuned encoding Accuracy for mnist with OPA
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.76
0.77
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for cifar with OPA
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.76
0.79
promp tuned
fine tuned

# Fine-tuned


Comparing mnist and cifar encoding Accuracy with OPA
0.77
0.79
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mnist
cifar

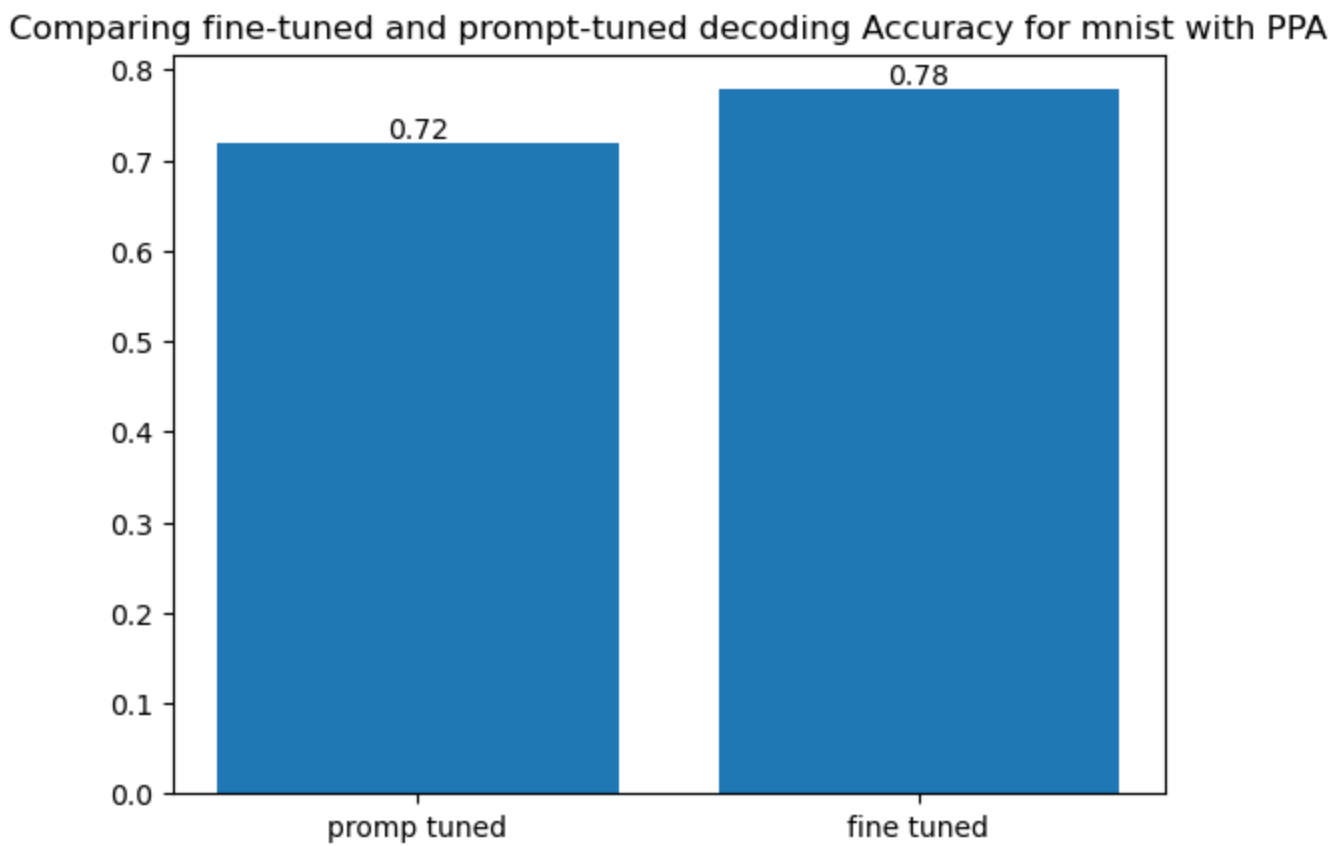
# Prompt-tuned

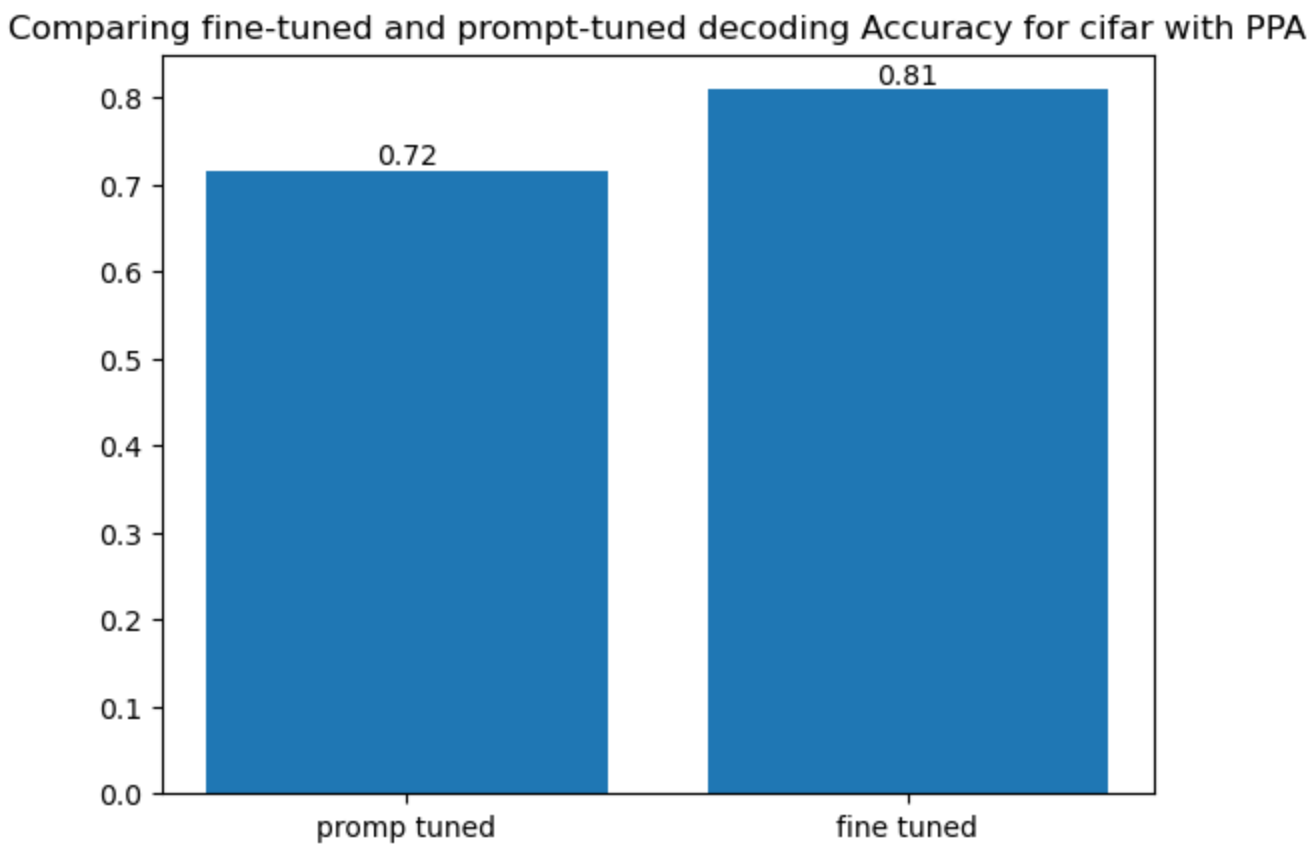


# Prompt-tuning vs Fine-tuning (Brain Encoding)

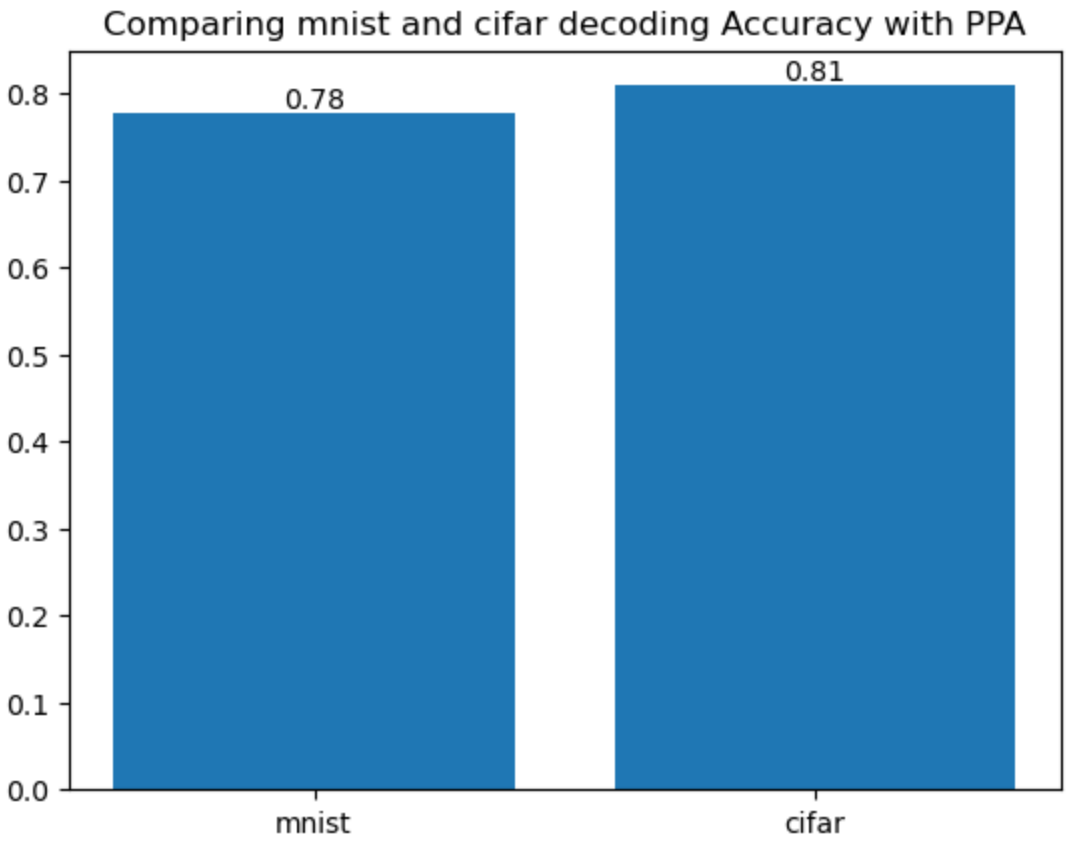
- The margin is less, but still in both the models, Fine-tuning beats prompt-tuning.
- The both models, one prompt tuned on MNIST and one on CIFAR-10 perform at an equal level indicating similar effect of prompt-tuning.

# PPA ROI

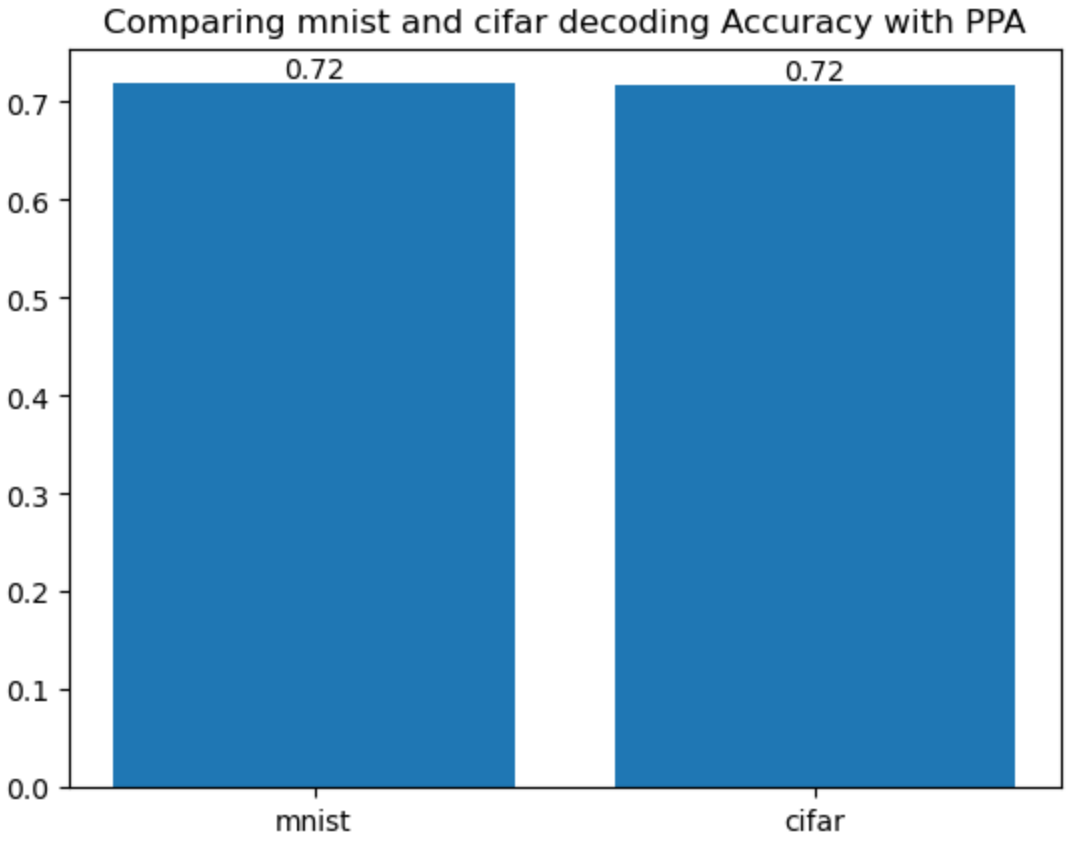

Comparing fine-tuned and prompt-tuned decoding Accuracy for mnist with PPA
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.72
0.78
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for cifar with PPA
0.81
0.72
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned

# Fine-tuning

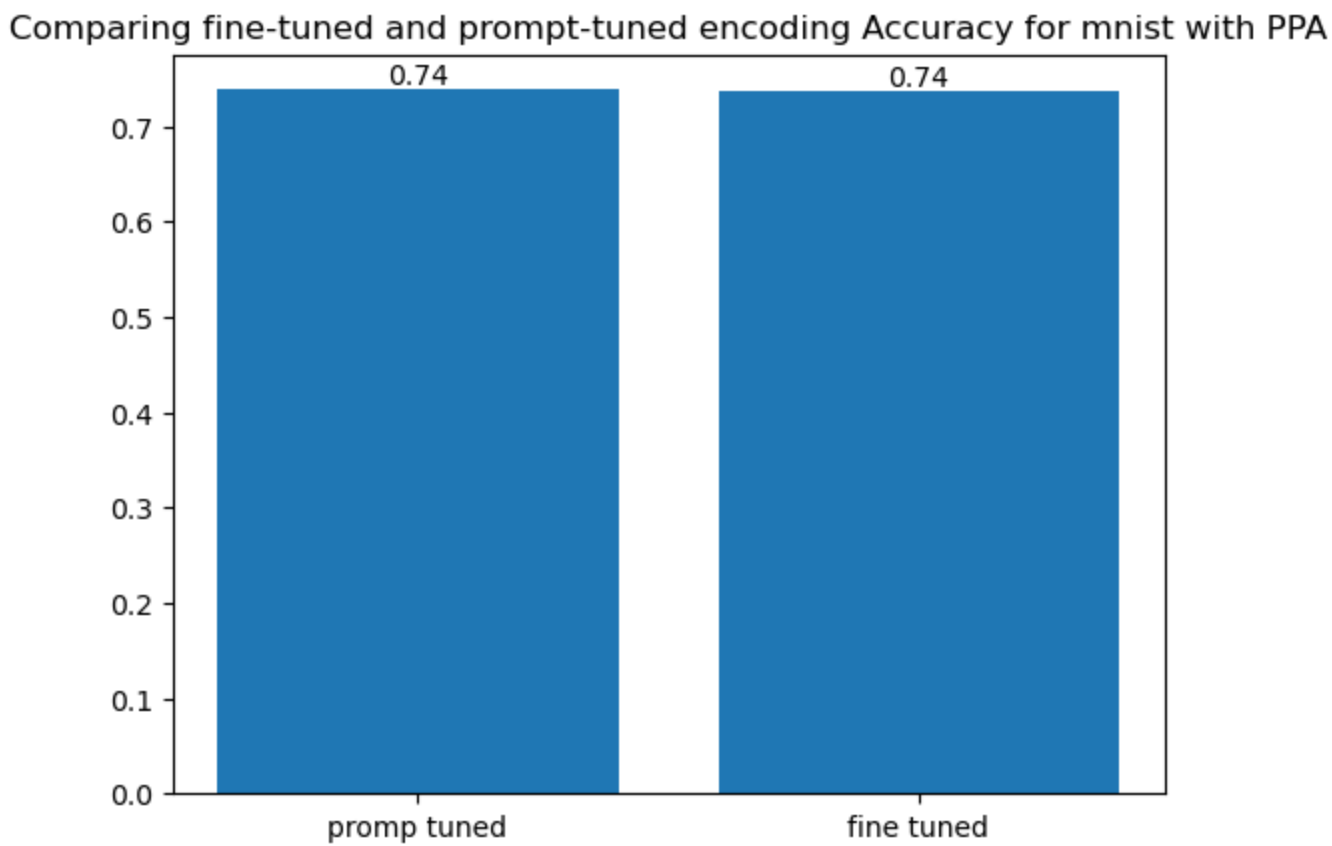

Comparing mnist and cifar decoding Accuracy with PPA
0.78
0.81
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mnist
cifar

# Prompt-tuning

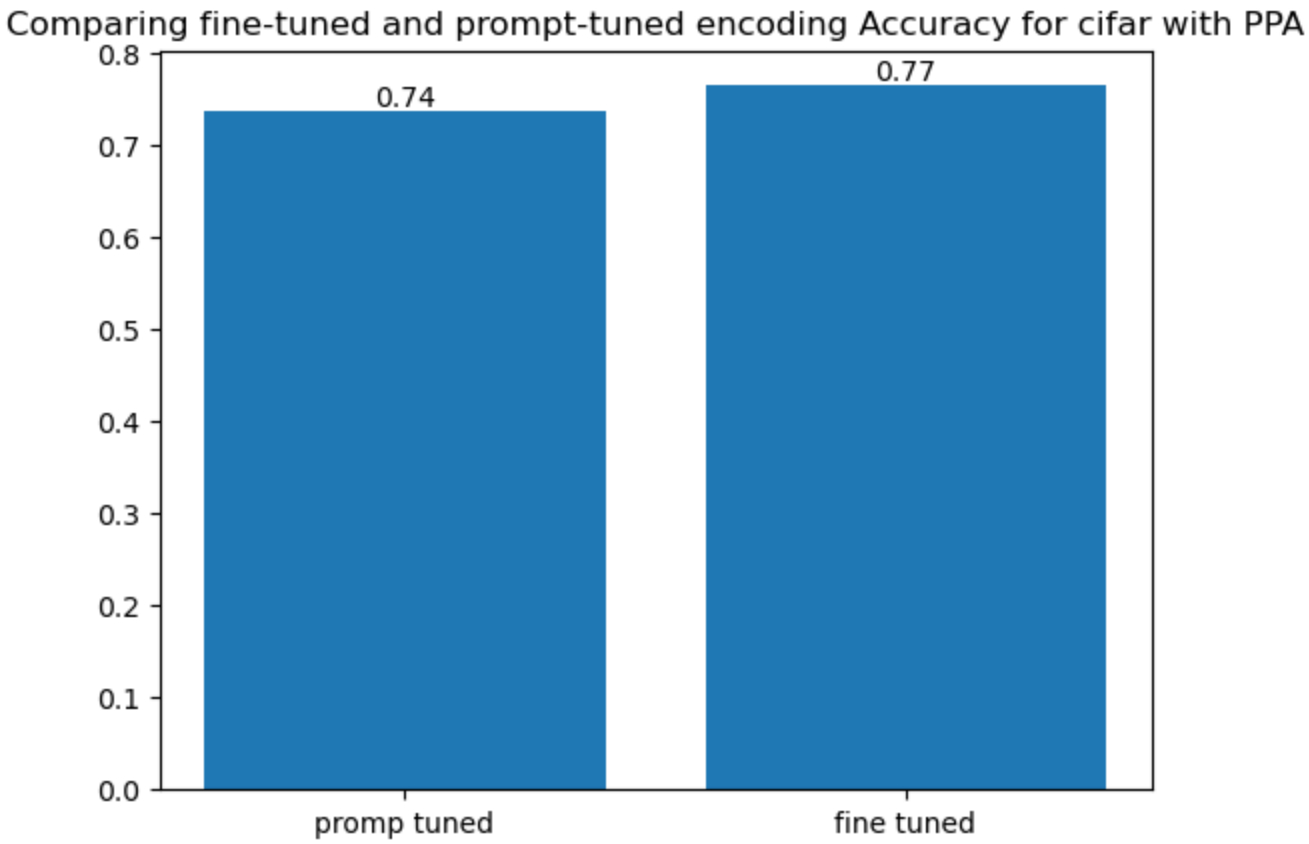


# Prompt-tuning vs Fine-tuning (Brain Decoding)

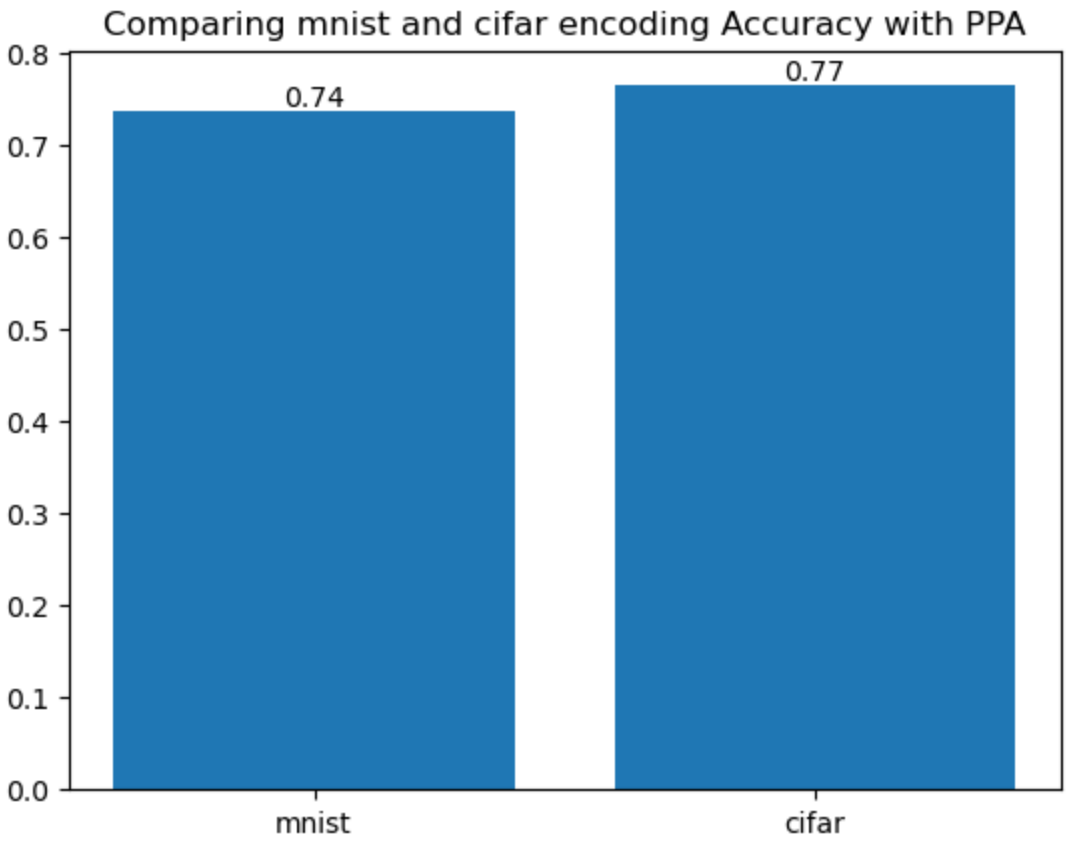
- Again, with a clear trend, **fine-tuning models outperform prompt-tuning models.**
- Also both the prompt-tuned models, one on MNIST and CIFAR give the same accuracy.

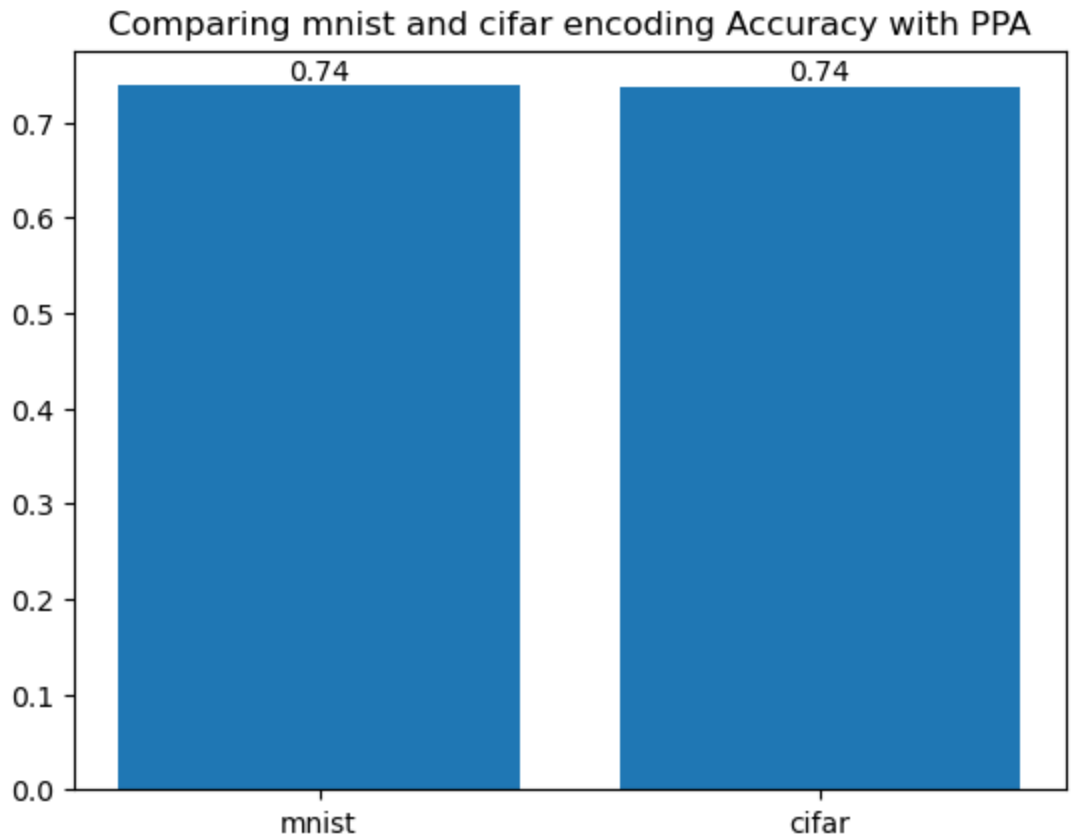

Comparing fine-tuned and prompt-tuned encoding Accuracy for mnist with PPA
0.74
0.74
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned

# Fine-tuning


Comparing fine-tuned and prompt-tuned encoding Accuracy for cifar with PPA
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.74
0.77
promp tuned
fine tuned

# Prompt-tuning

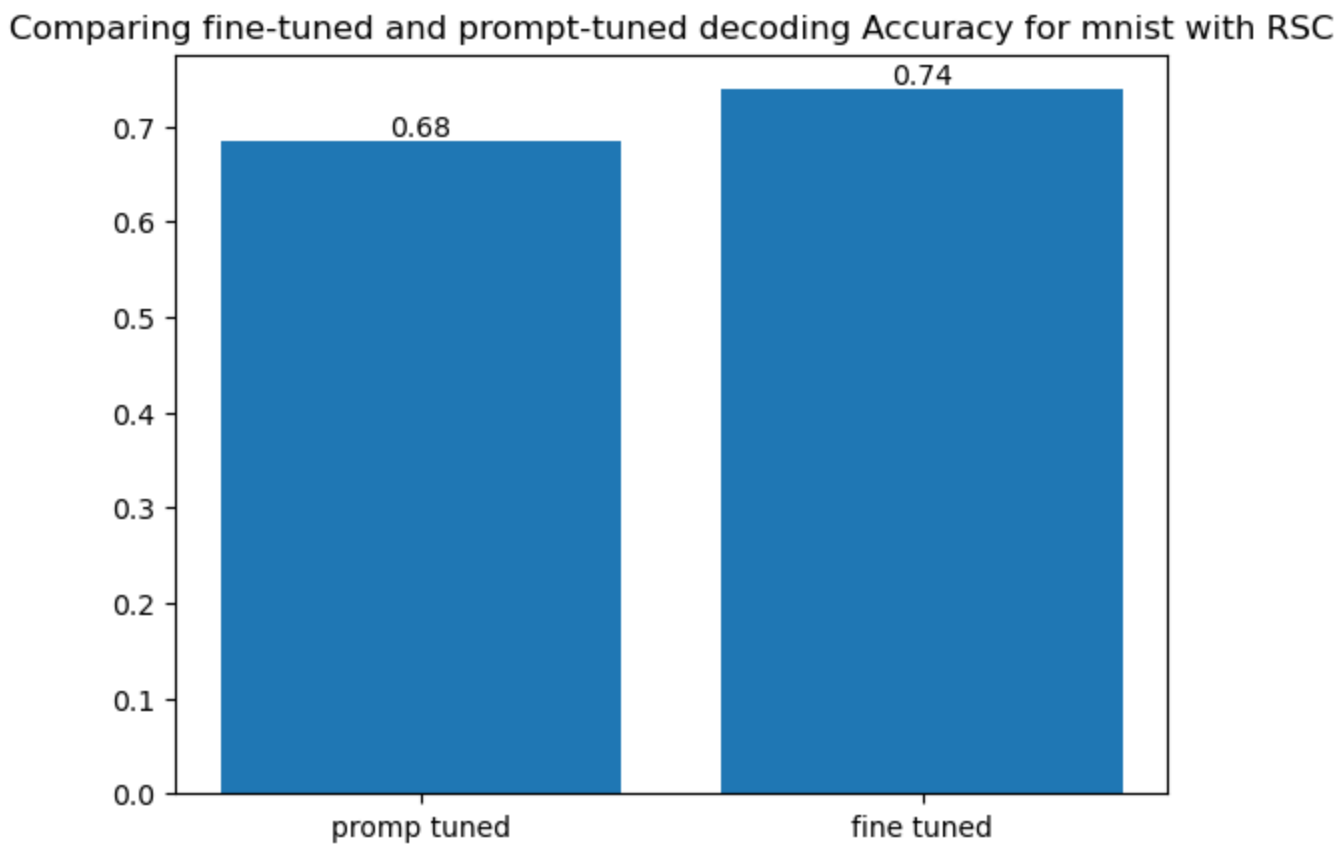


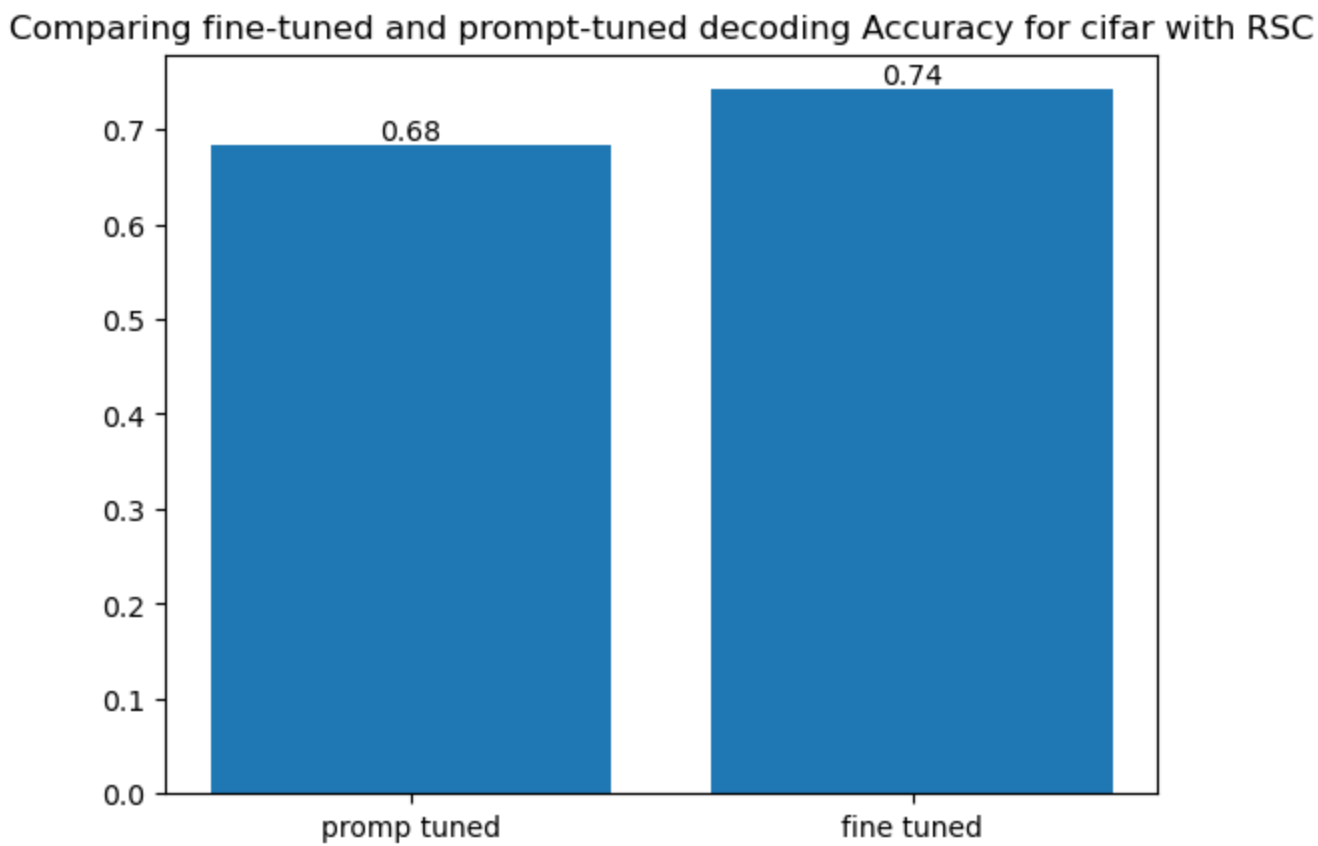

Comparing mnist and cifar encoding Accuracy with PPA
0.74
0.74
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mnist
cifar

# Prompt-tuning vs Fine-tuning (Brain Encoding)

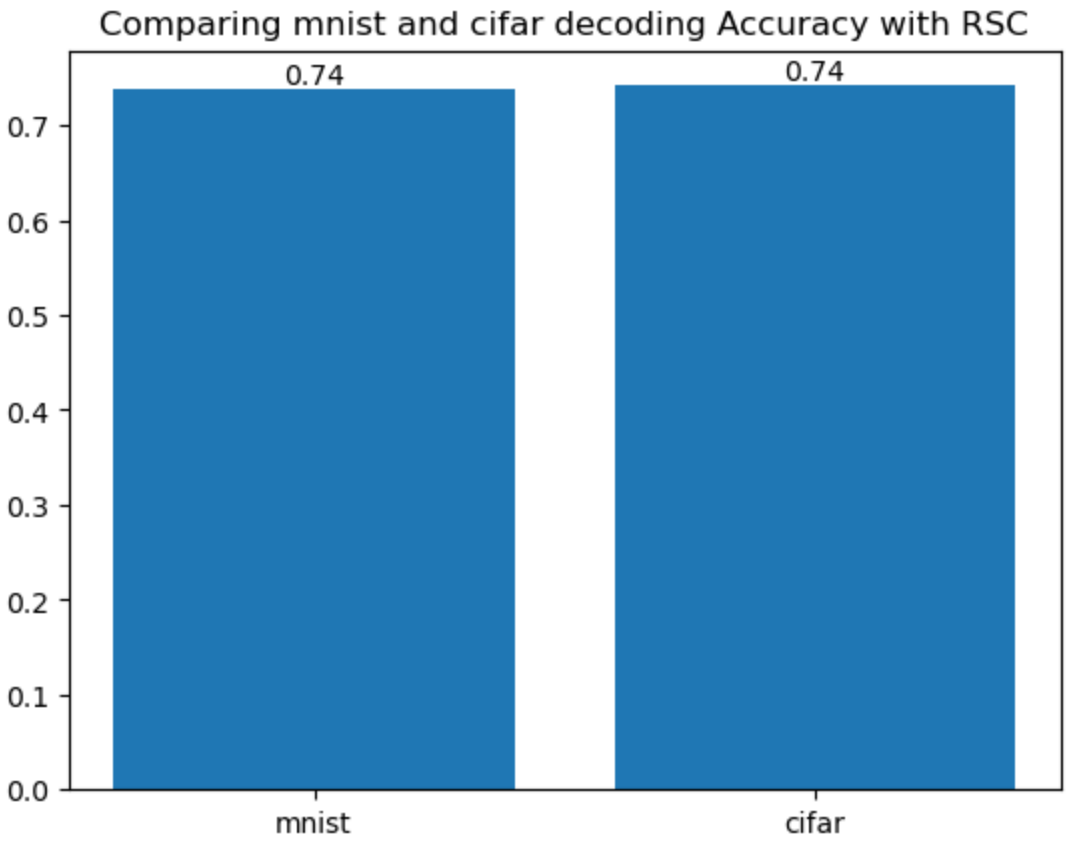
- Though the MNIST dataset doesn't signify a difference, the CIFAR dataset pronounces that ViT Base Fine-tuned on CIFAR better accounts for brain representations than ViT prompt tuned on CIFAR.
- Somehow, or sometimes, the capacity of the fine-tuned ViT on MNIST is equaling the prompt-tuned ViT on MNIST.
- Again the trend of fine-tuned models performing better continues.

# RSC ROI

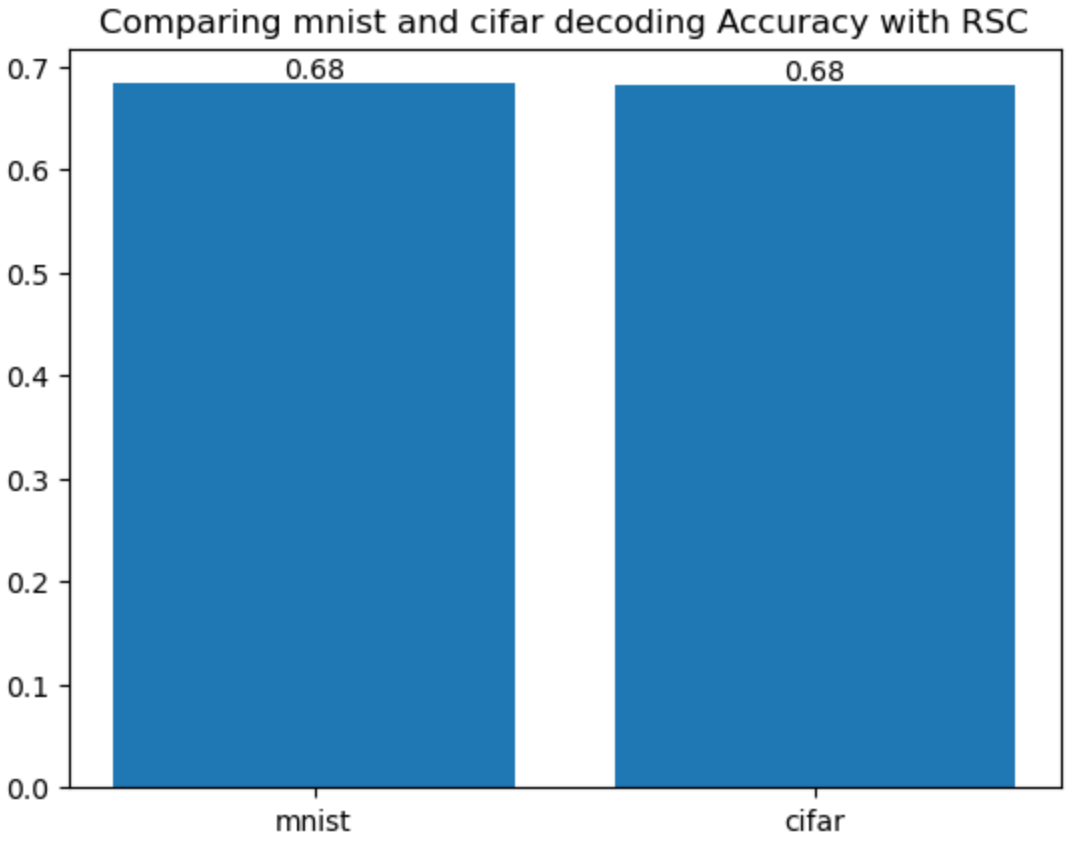

Comparing fine-tuned and prompt-tuned decoding Accuracy for mnist with RSC
0.74
0.68
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for cifar with RSC
0.74
0.68
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned

# Fine-tuning

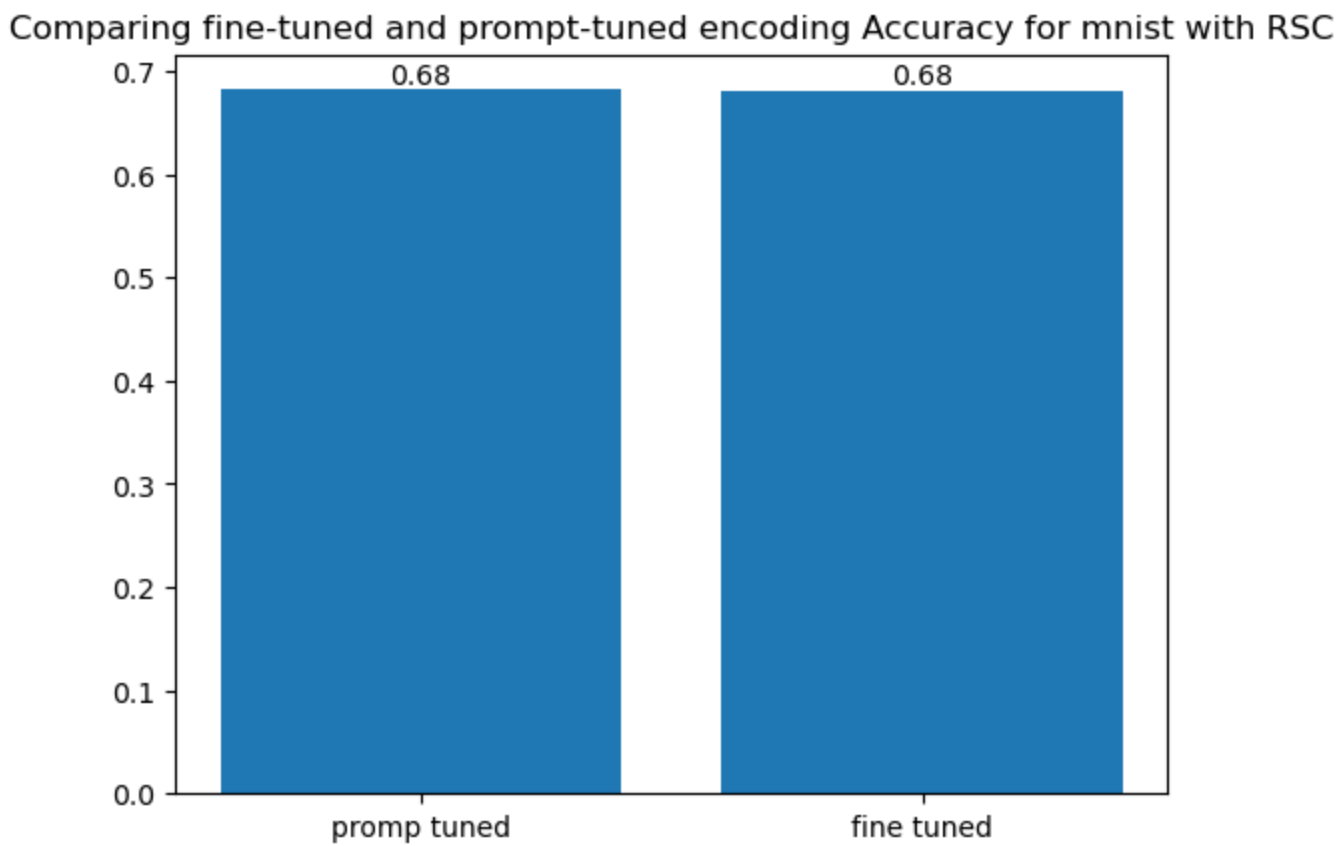

Comparing mnist and cifar decoding Accuracy with RSC
0.74
0.74
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
mnist
cifar

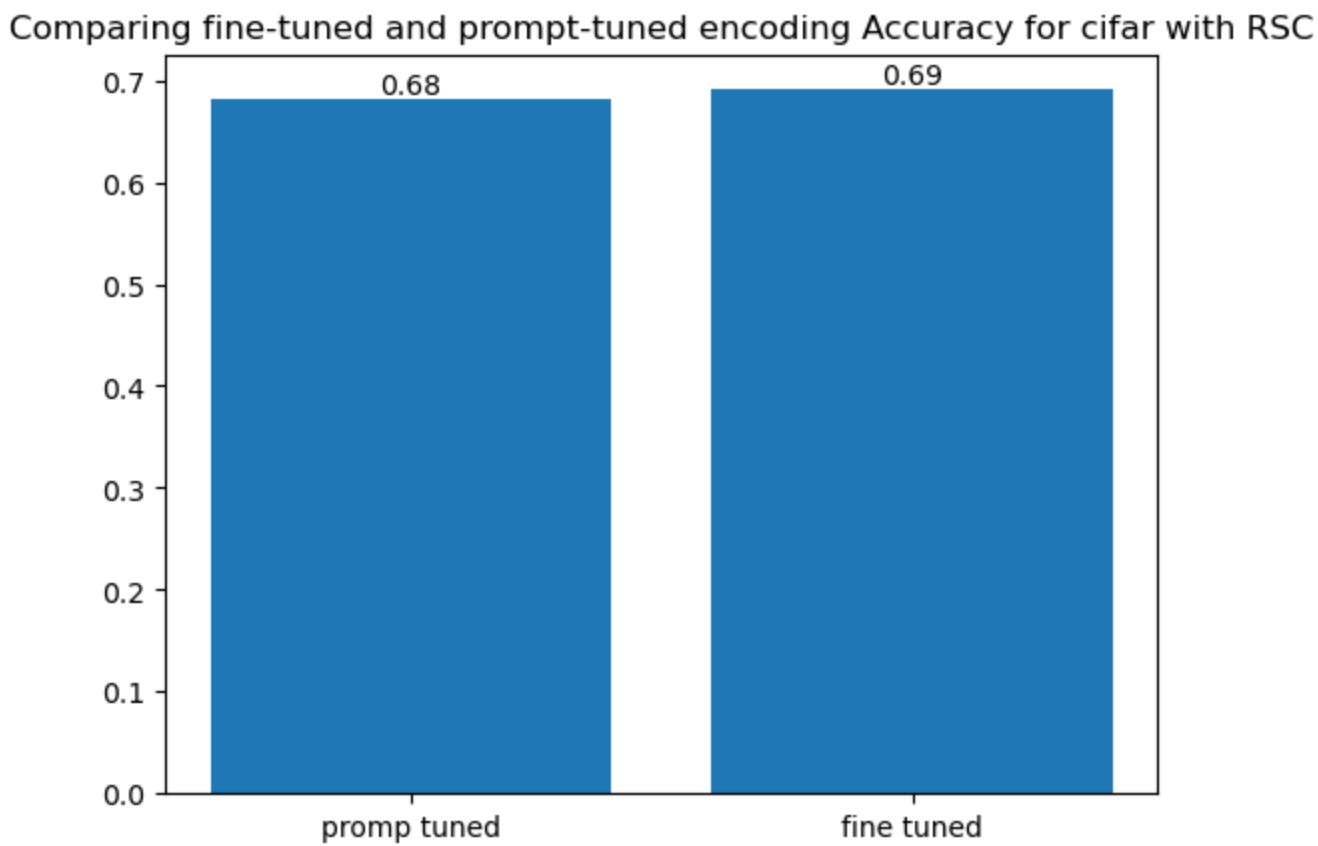
# Prompt-tuning



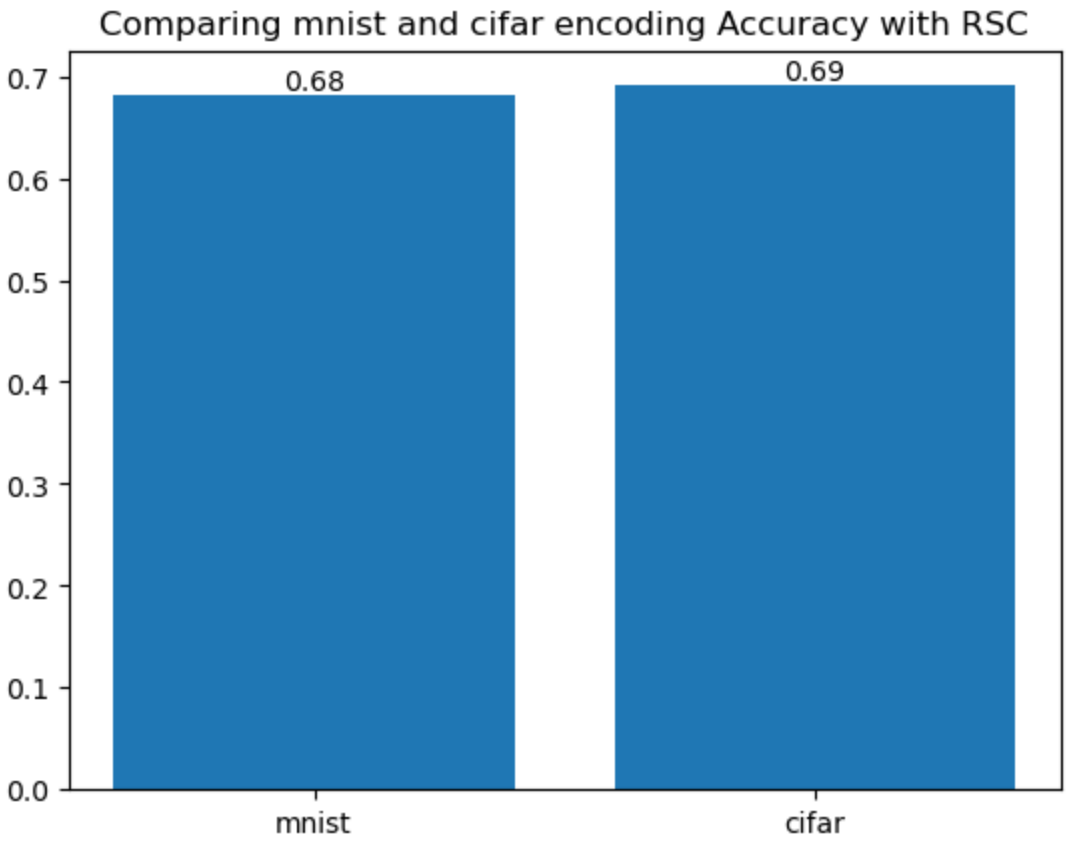
# Prompt-tuning vs Fine-tuning (Brain Decoding)

- With a clear trend, **fine-tuning models outperform prompt-tuning models.**
- Also both the prompt-tuned models, one on MNIST and CIFAR give the same accuracy.

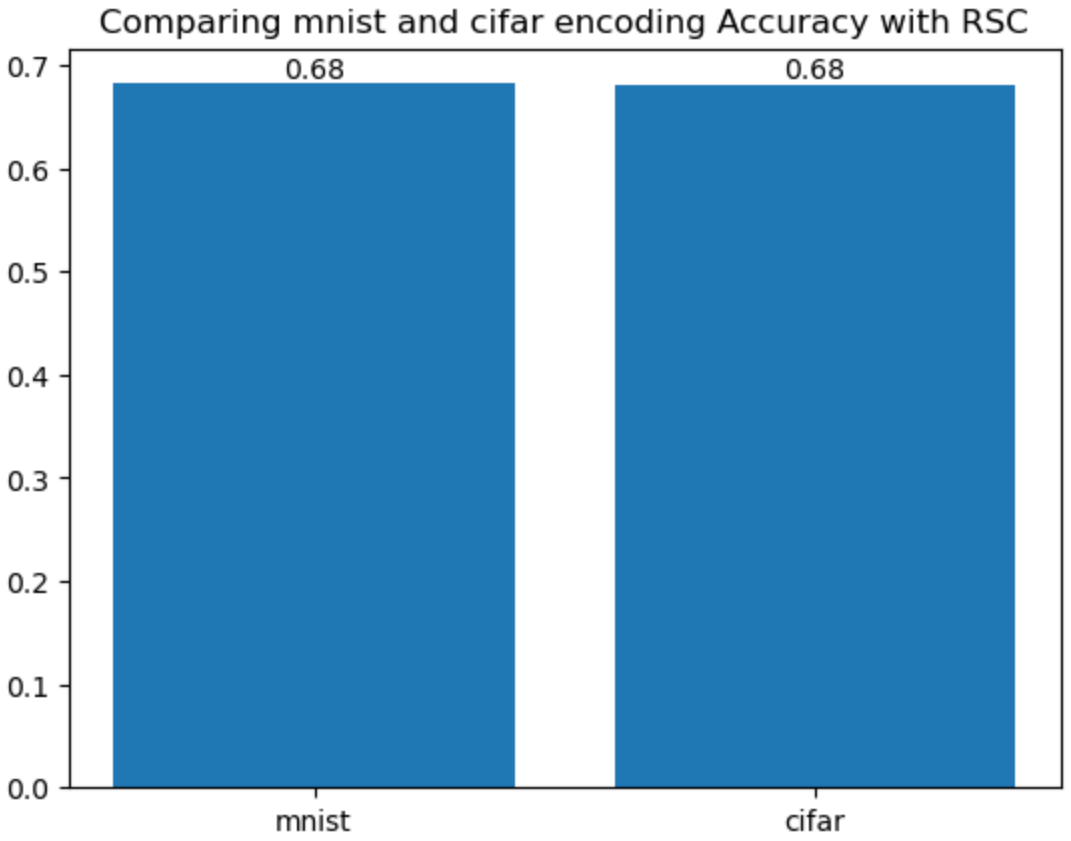

Comparing fine-tuned and prompt-tuned encoding Accuracy for mnist with RSC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.68
0.68
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for cifar with RSC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.68
0.69
promp tuned
fine tuned

# Fine-tuning


Comparing mnist and cifar encoding Accuracy with RSC
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.68
0.69
mnist
cifar

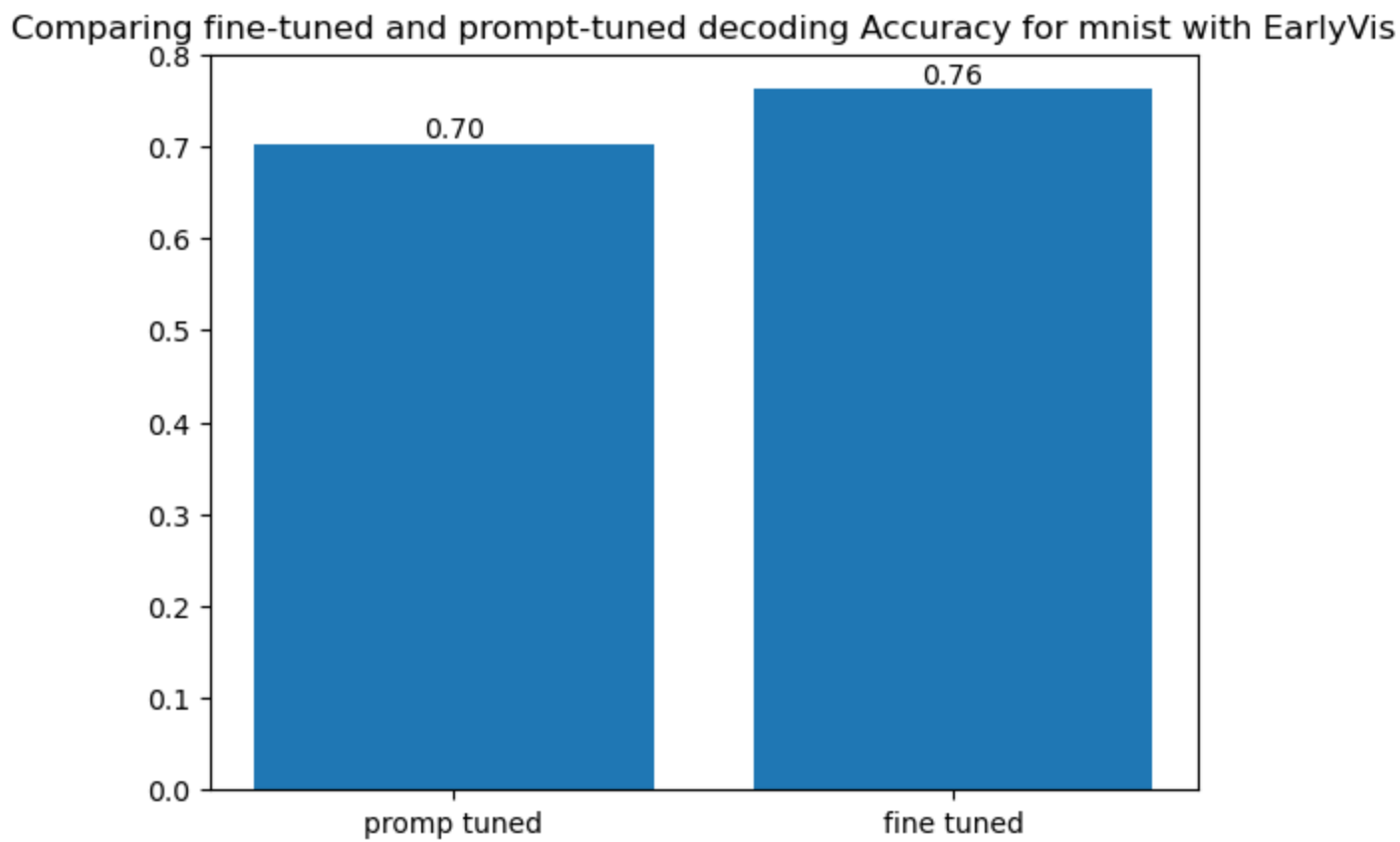
# Prompt-tuning

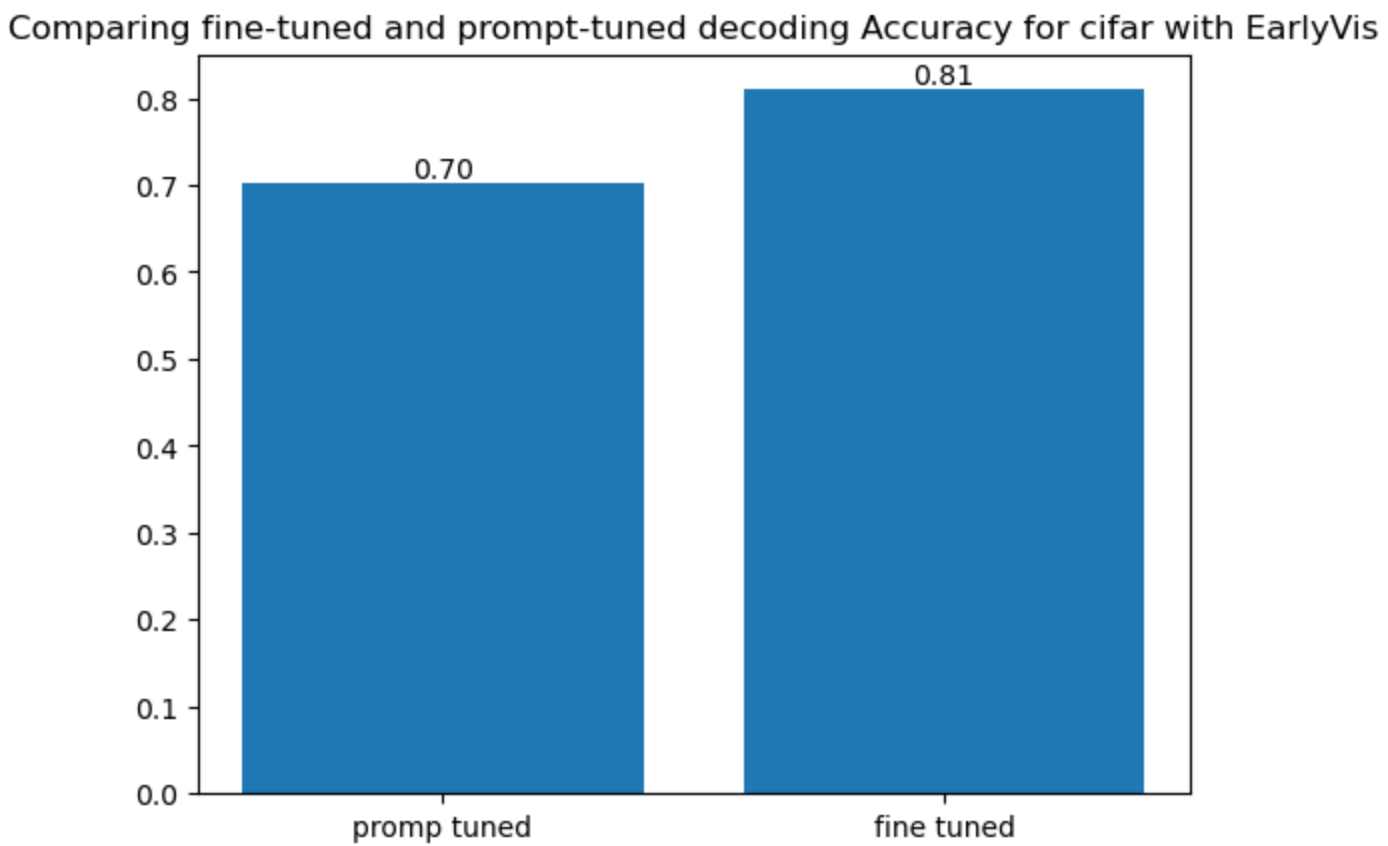


# Prompt-tuning vs Fine-tuning (Brain Encoding)

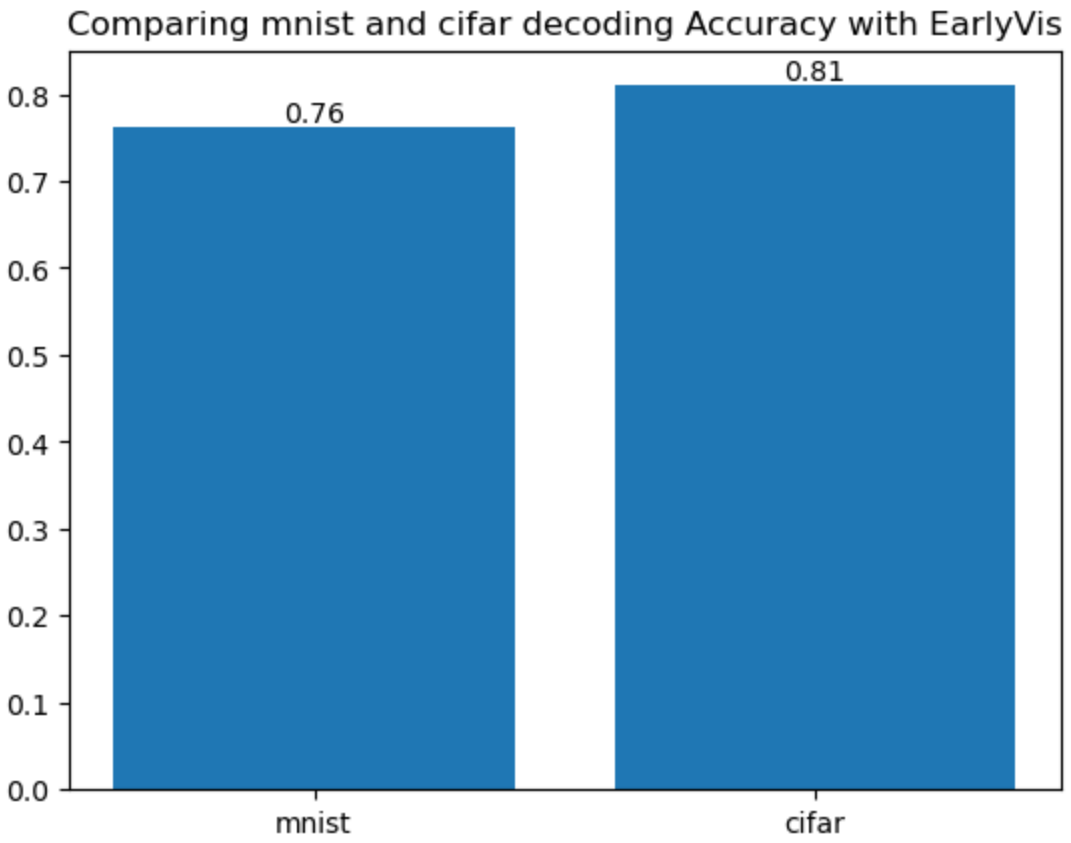
- Though the MNIST dataset doesn't signify a difference, the CIFAR dataset pronounces that ViT Base Fine-tuned on CIFAR better accounts for brain representations than ViT prompt tuned on CIFAR.
- But the CIFAR dataset also gives only a small advantage to Fine-tuning over Prompt-tuning.
- Again the trend of fine-tuned models performing better continues.

# EarlyVis ROI

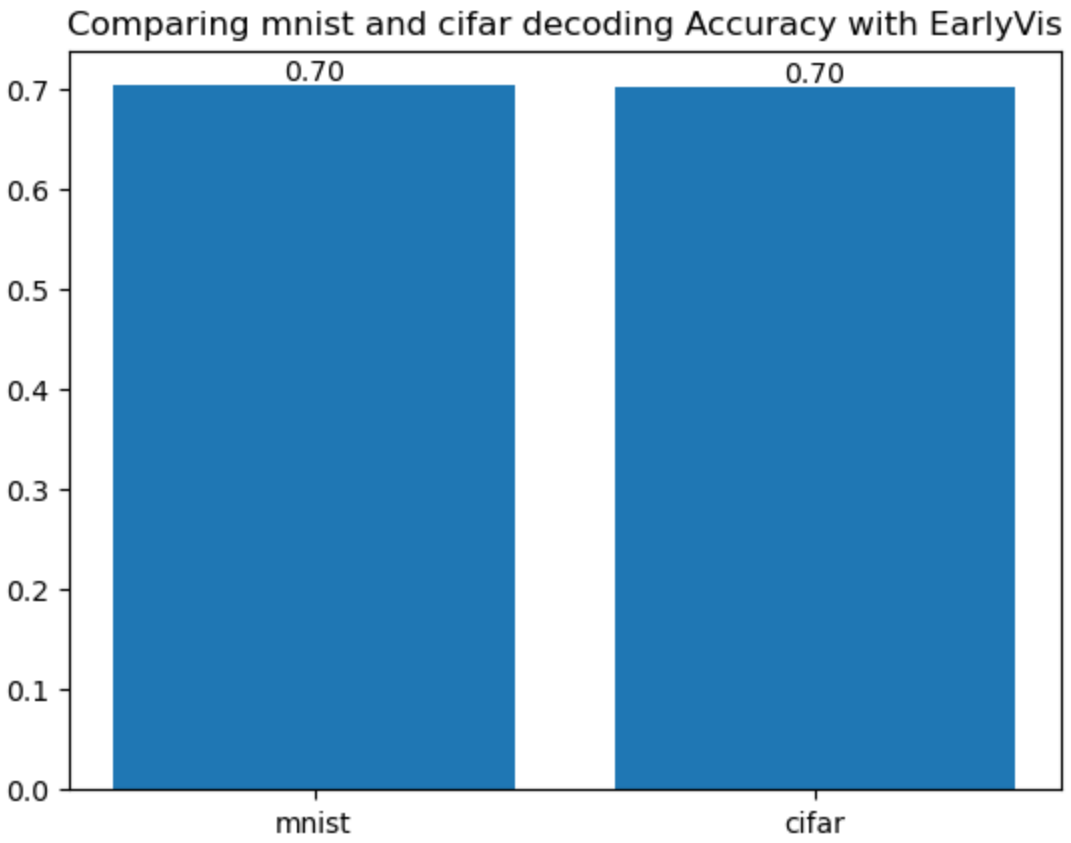

Comparing fine-tuned and prompt-tuned decoding Accuracy for mnist with EarlyVis
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.70
0.76
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned decoding Accuracy for cifar with EarlyVis
0.81
0.70
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned

# Fine-tuning

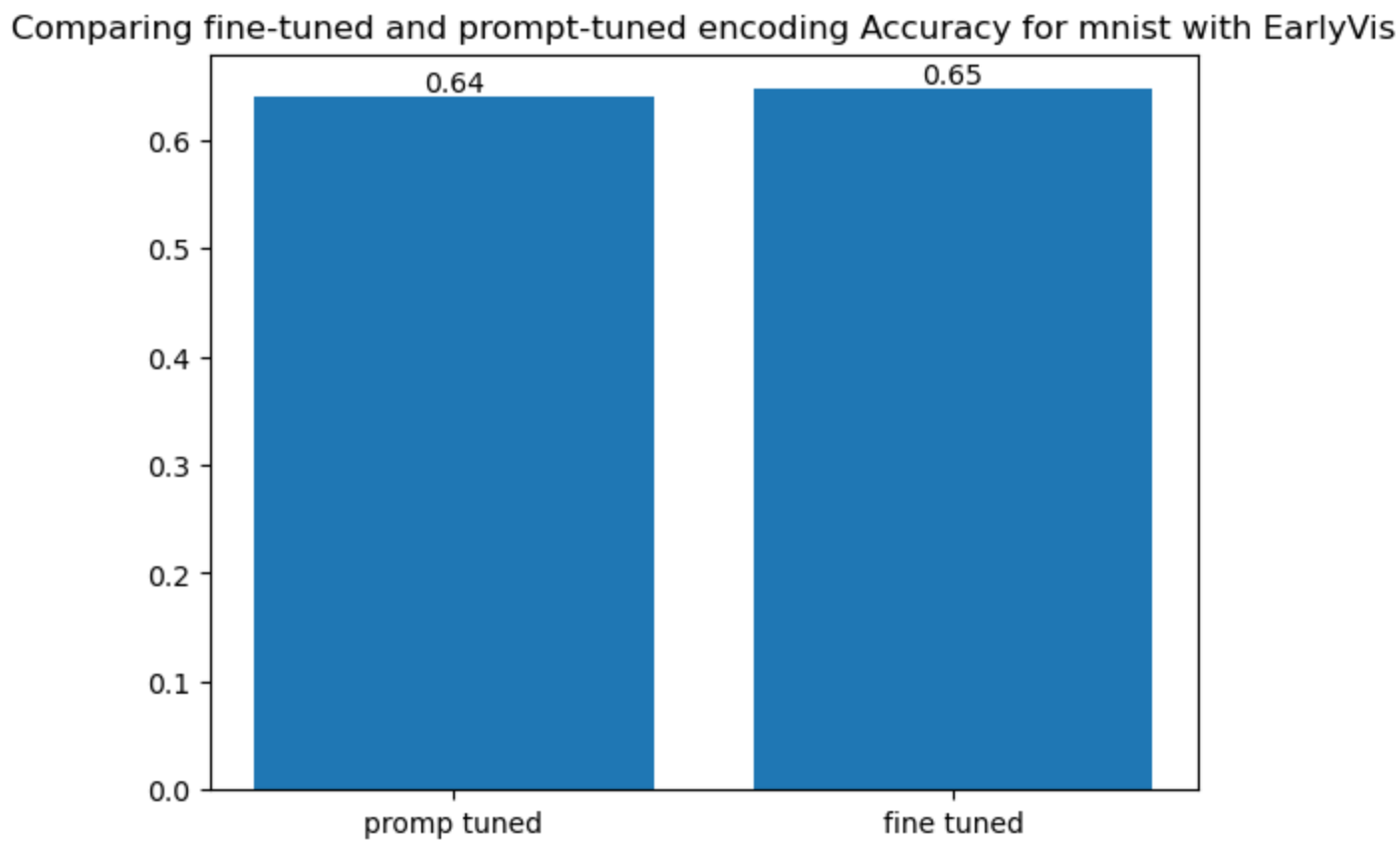

Comparing mnist and cifar decoding Accuracy with EarlyVis
0.76
0.81
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
mnist
cifar

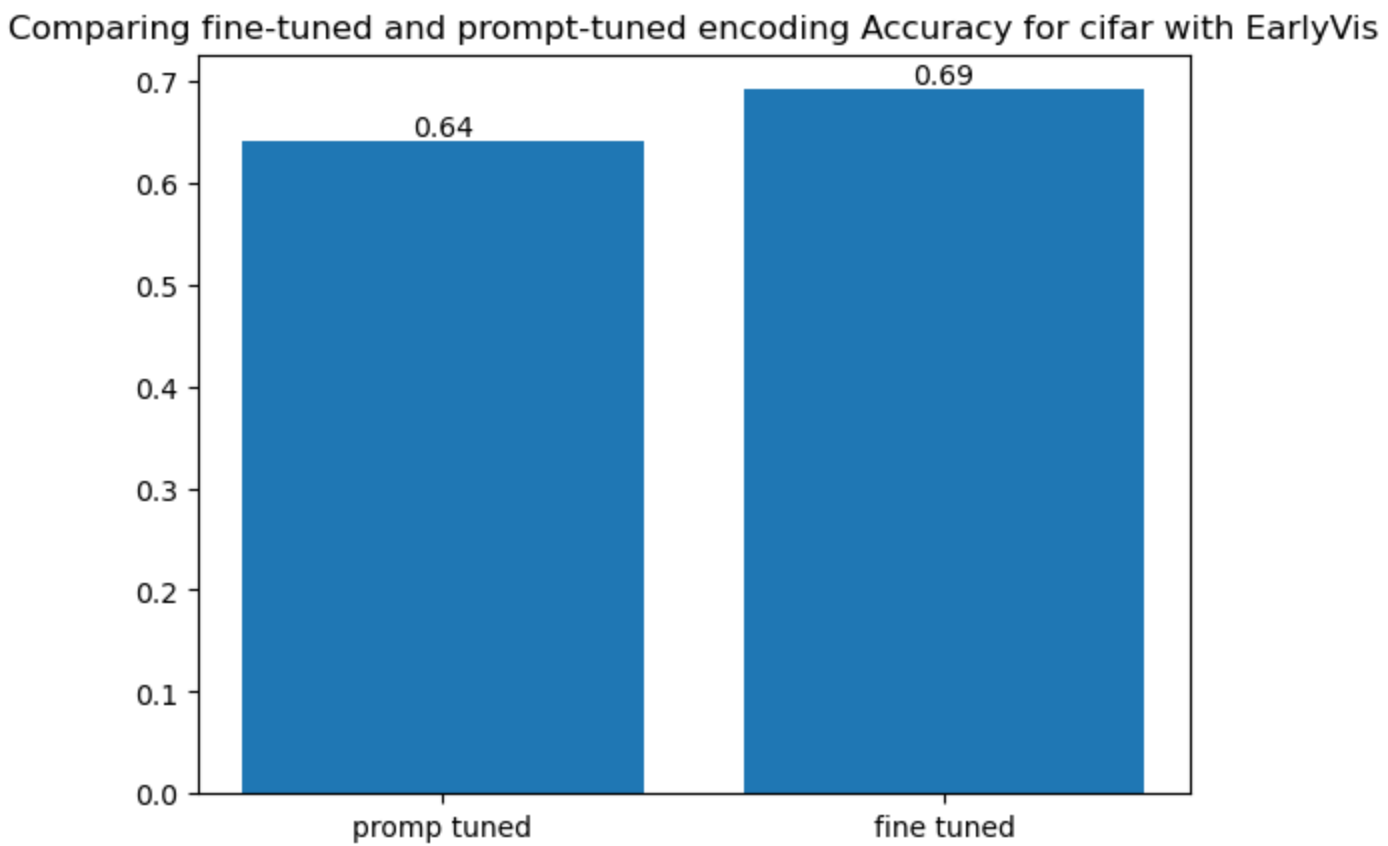
# Prompt-tuning



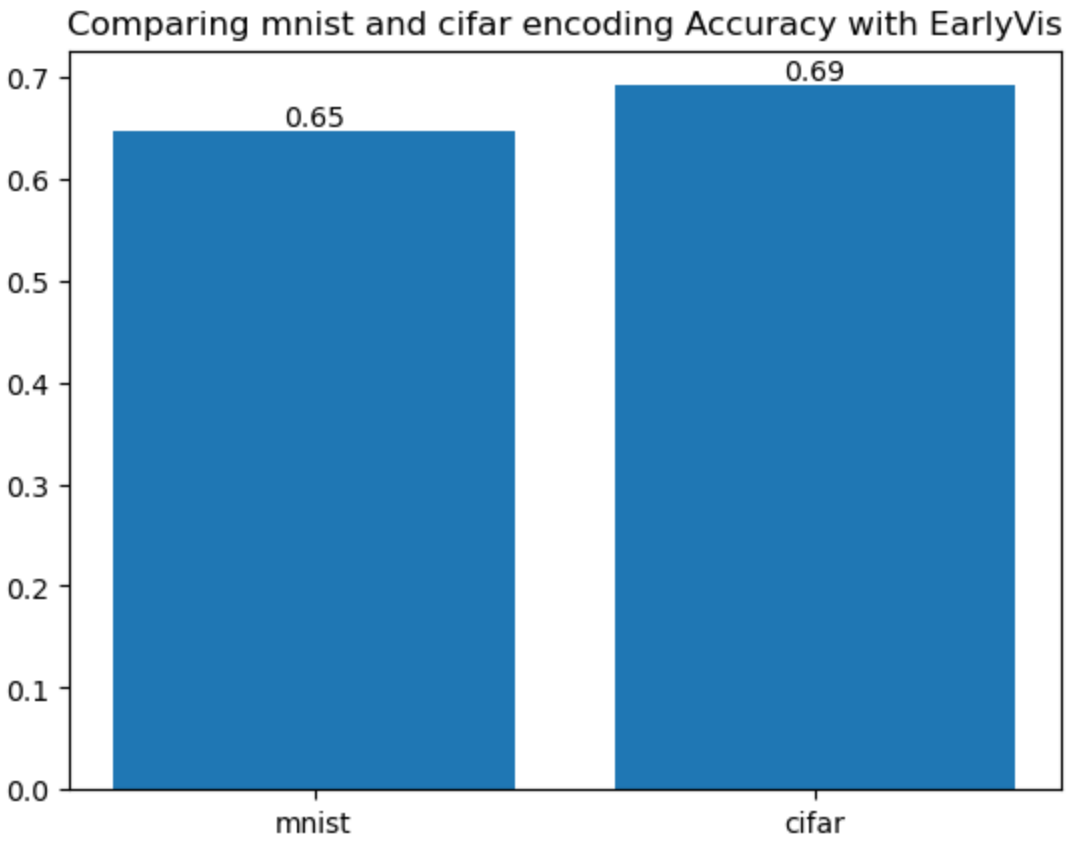
# Prompt-tuning vs Fine-tuning (Brain Decoding)

- With a clear trend, **fine-tuning models outperform prompt-tuning models.**
- Also both the prompt-tuned models, one on MNIST and CIFAR give the same accuracy.
- One **overall Decoding trend:** Fine-tuned ViT **always** better accounts for brain decoding in visual stimuli.

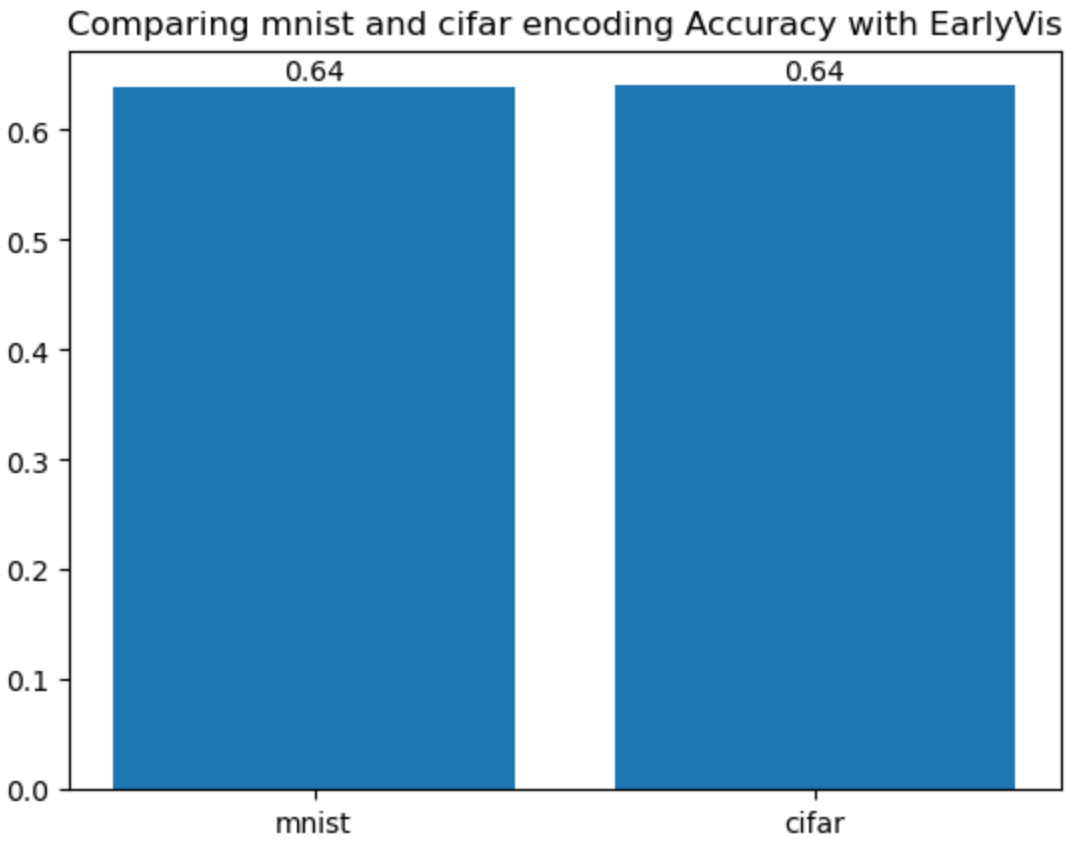

Comparing fine-tuned and prompt-tuned encoding Accuracy for mnist with EarlyVis
0.64
0.65
0.6
0.5
0.4
0.3
0.2
0.1
0.0
promp tuned
fine tuned


Comparing fine-tuned and prompt-tuned encoding Accuracy for cifar with EarlyVis
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.64
0.69
promp tuned
fine tuned

# Fine-tuning


Comparing mnist and cifar encoding Accuracy with EarlyVis
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.65
0.69
mnist
cifar

# Prompt-tuning



# Prompt-tuning vs Fine-tuning (Brain Encoding)

- Both the fine-tuned models, beat the corresponding prompt-tuned models
- Again the trend of fine-tuned models performing better continues.
- One recurring Encoding trend: ViT fine-tuned on CIFAR-10 always is better for brain encoding than its counterpart, wheras for MNIST, they perform equally in some cases but **never** does Prompt-tuned model perform better.

# Discussion

- In Text:
  - Prompt tuning vs fine tuning?
  - Summarization vs Question Answering vs Translation?

- In Vision
  - Prompt tuning vs fine tuning?
  - MNIST vs CIFAR

# Contributions

- In Text:
  - Fine Tuned vs Prompt Tuned Representation of NLG tasks
  - Which are better aligned to the brain?

- In Vision:
  - Implemented Visual Prompt Tuning (code not available on the internet)
  - Finetuned vs Prompt Tuned Representations of Images
  - Which are better aligned to the brain?

Thank you!