🚓 TEAM PROMPT PATROL 🚨

THE IMPACT OF SYSTEM PROMPT LANGUAGE ON TOURISM-FOCUSED LARGE LANGUAGE MODEL RESPONSE QUALITY

O2PLUS

Gabriel I. Geraldo
Mentor: Nazrul Amin
Global Hyperscale AI Camp 2025
Philippines, Ateneo de Manila University

WHAT ARE

PROMPTS FOR LARGE LANGUAGE MODELS

Instructions for LLMs

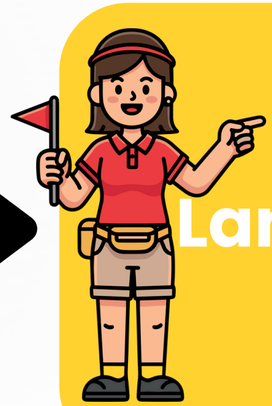
System Prompt

You are a tea shop owner and enthusiast in Jeju.

+

User Prompt

What **single** place should I go to in Jeju?



Large Language Model

Tour Guide LLM

Output

Go to O'Sulloc Tea Museum 🌿

If you only have time for one place in Jeju, this is it. You can walk through beautiful green tea fields, learn how Jeju's climate makes its tea special, and relax at the café with real Jeju-grown tea and desserts like matcha cake.

It's not just sightseeing—it's experiencing Jeju's calm, nature, and culture in one stop. 🍵

Key Takeaway: Prompts can be structured as system and user wherein user can be used for role-setting

ADJUSTING THE

SYSTEM INSTRUCTION LANGUAGE

Instructions for LLMs

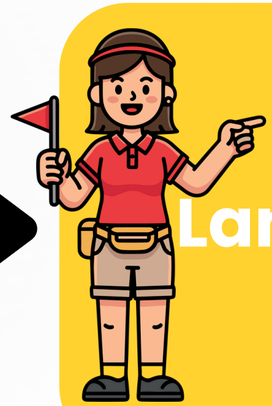
System Prompt

당신은 제주도 관광 전문 AI 어시스턴트입니다. 사용자의 제주 여행 계획을 도와주고, 관광지, 맛집, 숙소 등에 대한 정보를 제공합니다.

+

User Prompt

What **single** place should I go to in Jeju?



Large Language Model

Tour Guide LLM

Output

If you can visit just one place in Jeju, go to Seongsan Ilchulbong (Sunrise Peak).
Why this one spot stands out:

- 🌋 A dramatic volcanic crater right by the sea
- 🌅 Incredible views—especially at sunrise, but beautiful any time
- 🚶 Easy hike (about 20–30 minutes)
- 📸 One of Jeju's most iconic photo spots
- 🌍 UNESCO World Natural Heritage site

It captures Jeju's nature, history, and scenery all in one place—perfect if you only have time for a single stop.

Testing Hypothesis: Does changing system prompt language affect quality of response?

PROJECT

MOTIVATION

Why begin this project?

When the LLM sees Korean text data, it will associate the prompt with Korean information, possibly more authentic data.

System Prompt

당신은 제주도 관광 전문 AI 어시스턴트입니다. 사용자의 제주 여행 계획을 도와주고, 관광지, 맛집, 숙소 등에 대한 정보를 제공합니다.

+

User Prompt

What single place should I go to in Jeju?

이중 언어를 구사하는 한국인

Like asking a **bilingual** Korean person. They may know Korea more than English only tour guides.

Output

Output can be better informed by associated Korean data.

ADJUSTING THE

SYSTEM INSTRUCTION LANGUAGE

Instructions for LLMs

System Prompt

당신은 제주도 관광 전문 AI 어시스턴트입니다. 사용자의 제주 여행 계획을 도와주고, 관광지, 맛집, 숙소 등에 대한 정보를 제공합니다.

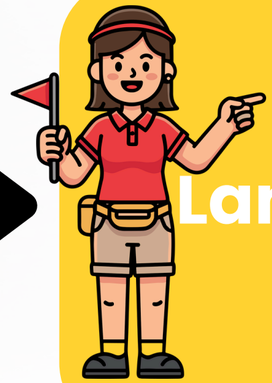
+

User Prompt

What **single** place should I go to in Jeju?

Qwen

LLaMA



Large Language Model

Tour Guide LLM

Output

If you can visit just one place in Jeju, go to Seongsan Ilchulbong (Sunrise Peak).
Why this one spot stands out:

- 🌋 A dramatic volcanic crater right by the sea
- 🌅 Incredible views—especially at sunrise, but beautiful any time
- 🚶 Easy hike (about 20–30 minutes)
- 📸 One of Jeju's most iconic photo spots
- 🌍 UNESCO World Natural Heritage site

It captures Jeju's nature, history, and scenery all in one place—perfect if you only have time for a single stop.

Testing Hypothesis: Which models display a more significant effect when using a Korean system prompt?

PROJECT

EXISTING WORK

What work have others done?

Existing Work

Haoyang Huang (2023) introduces cross-lingual-thought prompting which prompts a model to understand a non-English request. It improves benchmarks on MGSM (~+10).

Libo Qin et al. (2023) introduces cross-lingual CoT prompting by using English to understand/parse non-English queries; then using that understanding in a further prompt to resolve the query in English (multi-turn).

Ziqi Yin et al. (2024) studied the effect of politeness in prompts amongst English, Japanese, and Chinese tasks. Results show that impolite = bad results, being overpolite does not necessarily get better results, and there are optimal levels per language.

Key Takeaway: Where other work focuses on phrasing, this project focuses on studying direct usage of language.

PROJECT

PROBLEM STATEMENT

What research gap am I addressing?

Research Question

How does the language of system prompt instructions (Korean vs English) influence tour guide–style response quality and recommendation performance in a Jeju tourism-focused LLM?

Further Implications

1. Are we able to retrieve further knowledge on Jeju via the Korean prompt?
2. What models work with this method?
3. Explore the behavior of existing LLMs, study into maximizing how models are used and trained in terms of using English and region-specific non-English data.

PROJECT

DATA AND PREPARATION

What data is being used for this project?

Data Overview

Data as **Conversation** between User and LLM Agent regarding tourism queries in Jeju

Sourced from Jeju Tour Tool Calling Dataset by Chang Woo Choi from Pusan National University

Schema

```
// Represents one-turn conversation
{
    “system”: string,
    “user”: string,
    “assistant”: string
}
```

Training Dataset Size

n=900

Eval Dataset Size

n=100

Training and Eval have KOR and ENG versions

Key Takeaway: One dataset contains a system prompt in Korean, the other dataset’s system prompt is in English.

PROJECT

DATA PREPARATION

Jeju Tour Dataset

Korean Prompt Dataset

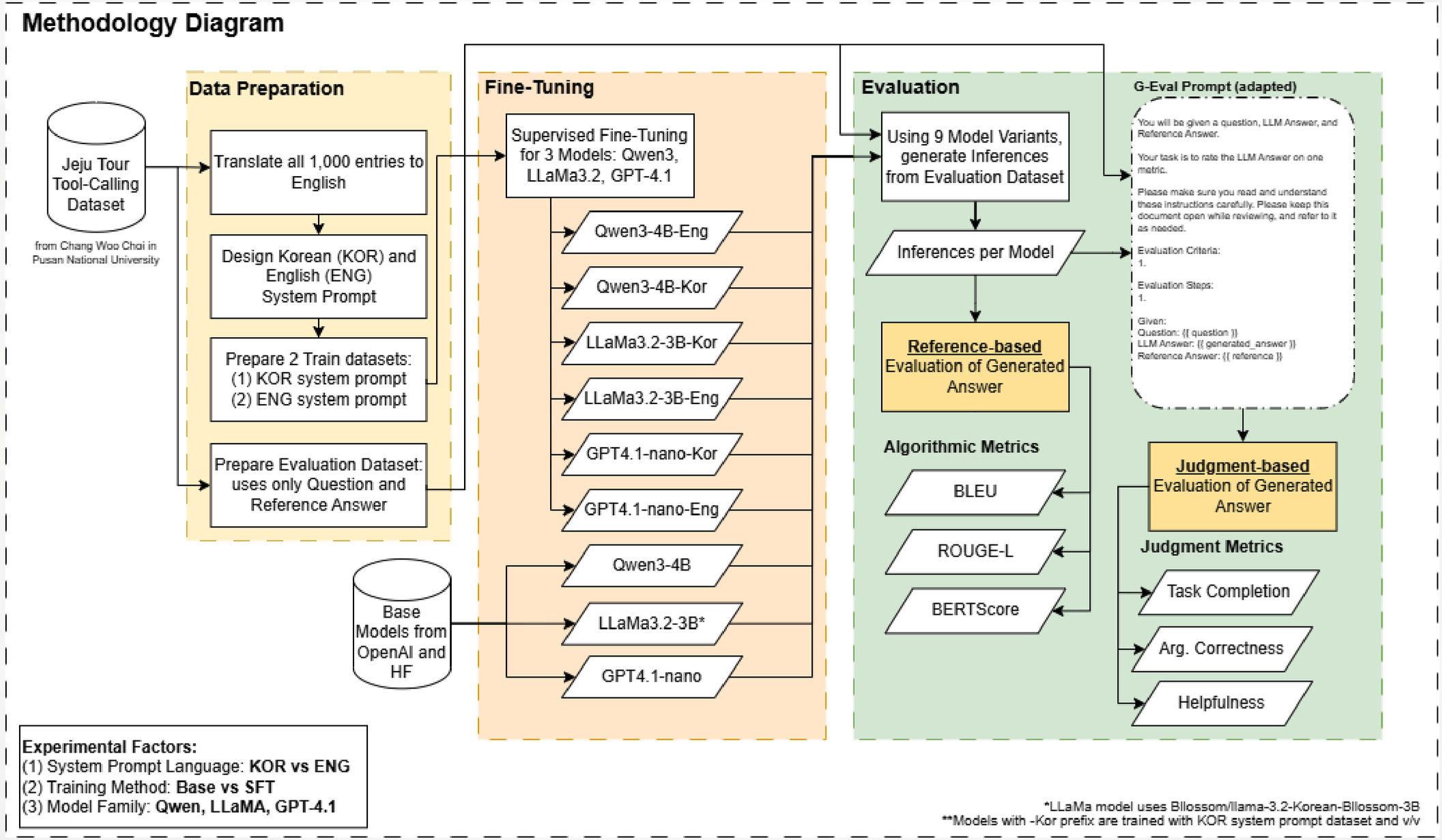
```
// Represents one-turn conversation
{
    “system”: 한국어 텍스트,
    “user”: string,
    “assistant”: string
}
```

English Prompt Dataset

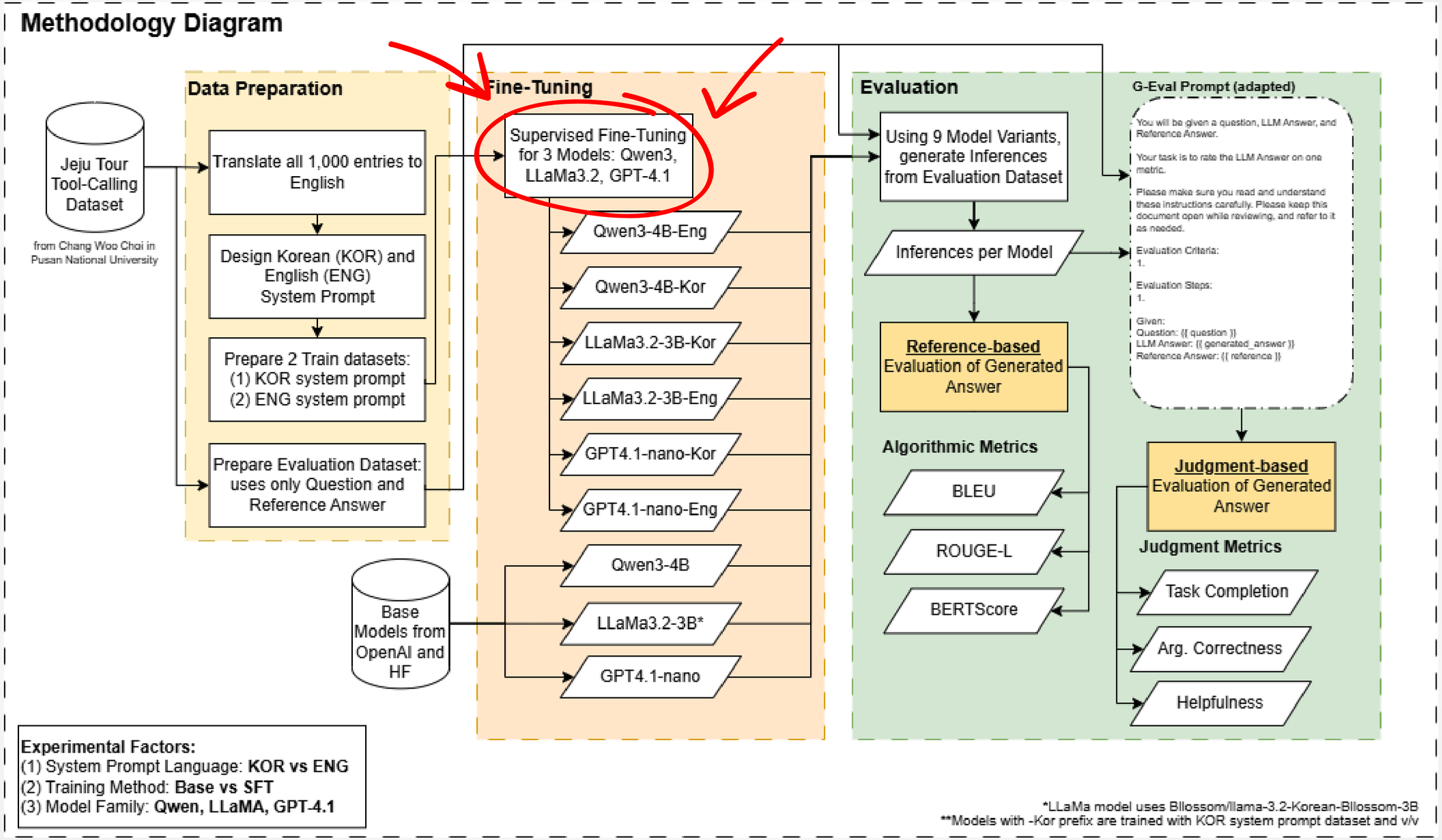
```
// Represents one-turn conversation
{
    “system”: string,
    “user”: string,
    “assistant”: string
}
```

Main Difference: One dataset contains a system prompt in Korean, the other dataset’s system prompt is in English.

METHODOLOGY DIAGRAM



METHODOLOGY DIAGRAM



TRAINING DIAGRAM

Using either,

Korean Prompt Dataset

// Represents one-turn conversation

```
{
    “system”: 한국어 텍스트,
    “user”: string,
    “assistant”: string
}
```

English Prompt Dataset

// Represents one-turn conversation

```
{
    “system”: string,
    “user”: string,
    “assistant”: string
}
```

SFT with LoRA

Qwen

LLaMA

Base

SFT with KOR

SFT with ENG

learning rate: 1e-4
epochs: 3
batch size: 1

Evaluation

Base with ENG Eval Dataset

Base with KOR Eval Dataset

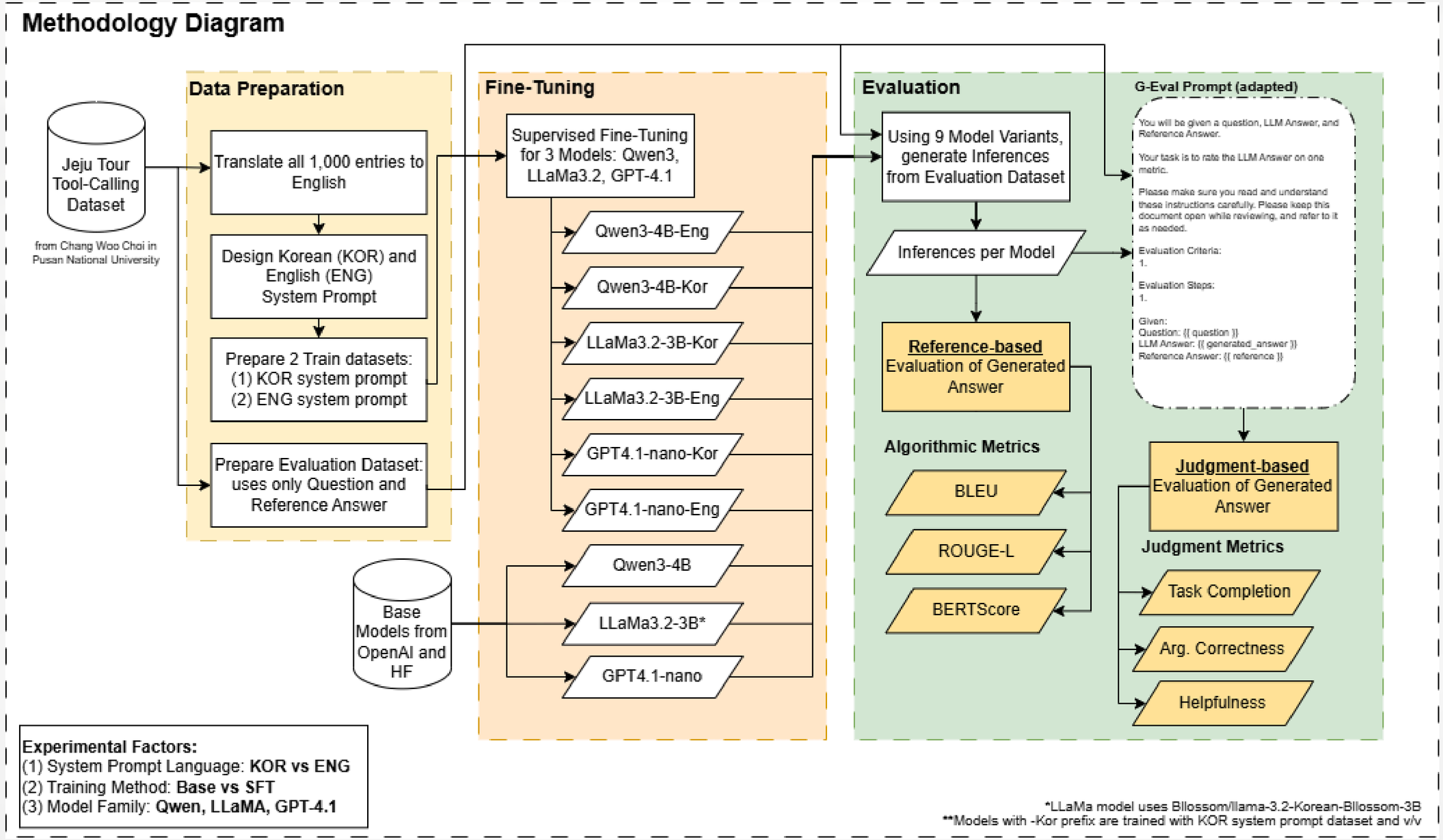
SFT-ENG with ENG Eval Dataset

SFT-KOR with KOR Eval Dataset

Control
Base Model with ENG Eval

Experimental
Base model with KOR Eval
SFT-ENG model with ENG Eval
SFT-KOR model with KOR Eval

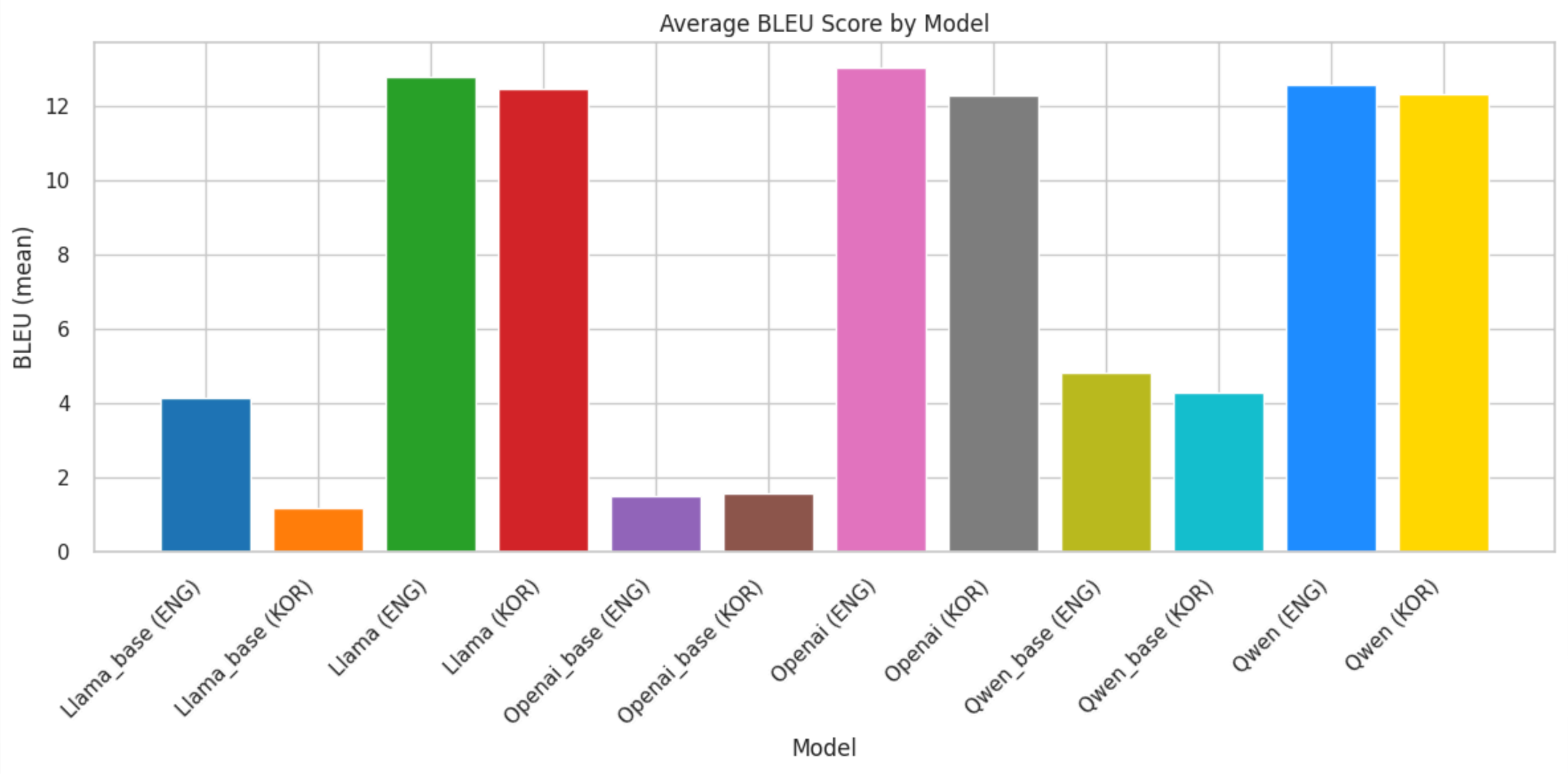
METHODOLOGY DIAGRAM



BLEU against reference

RESULTS

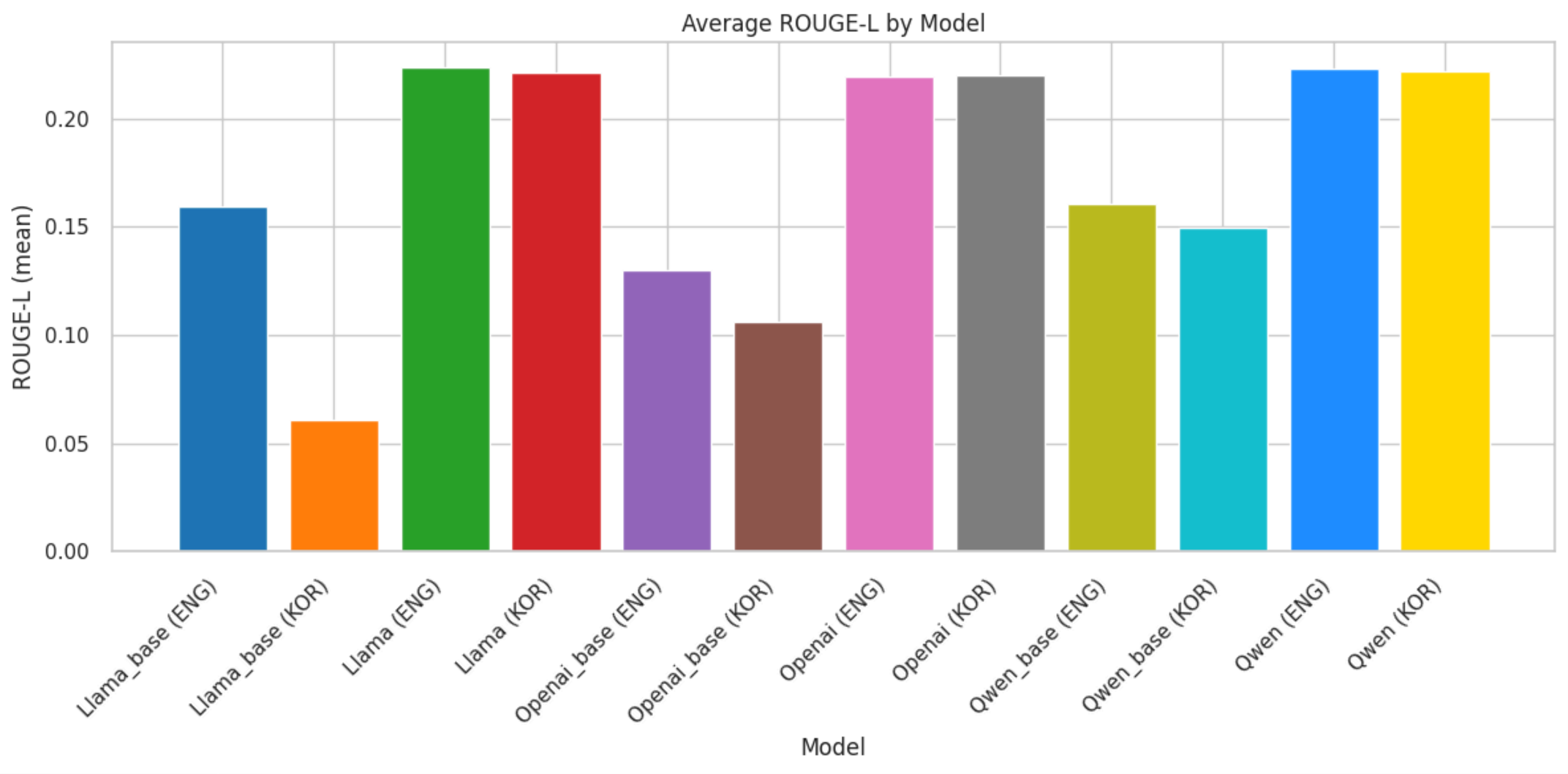
BLEU METRIC



Key Takeaway: In terms of BLEU mean, fine-tuned models perform better than base models. Base Qwen reaches higher scores than other larger base models.

RESULTS

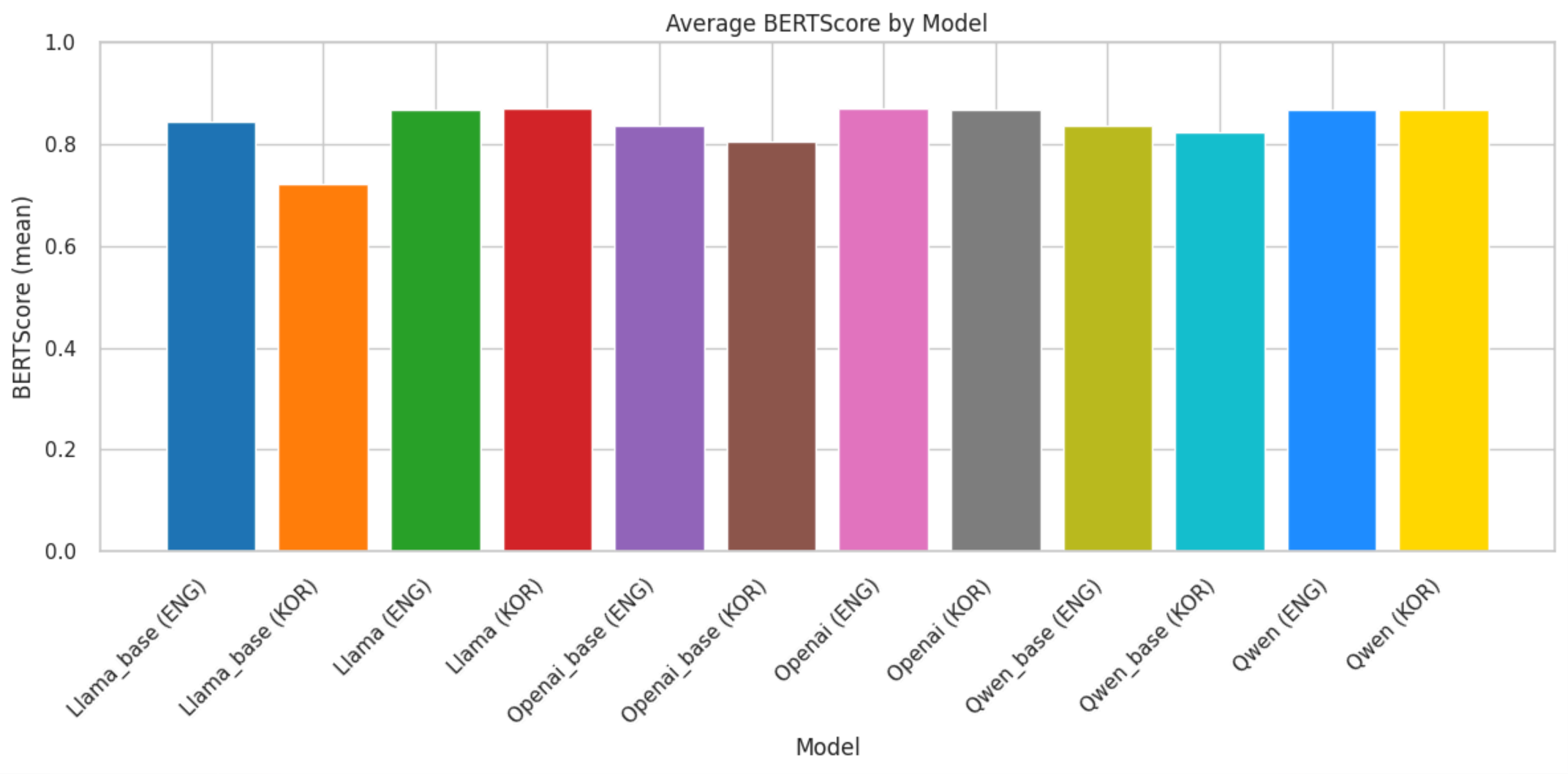
ROUGE-L



Key Takeaway: Usage of the KOR system prompt, proves to be harmful for the ROUGE-L automatic metric.

RESULTS

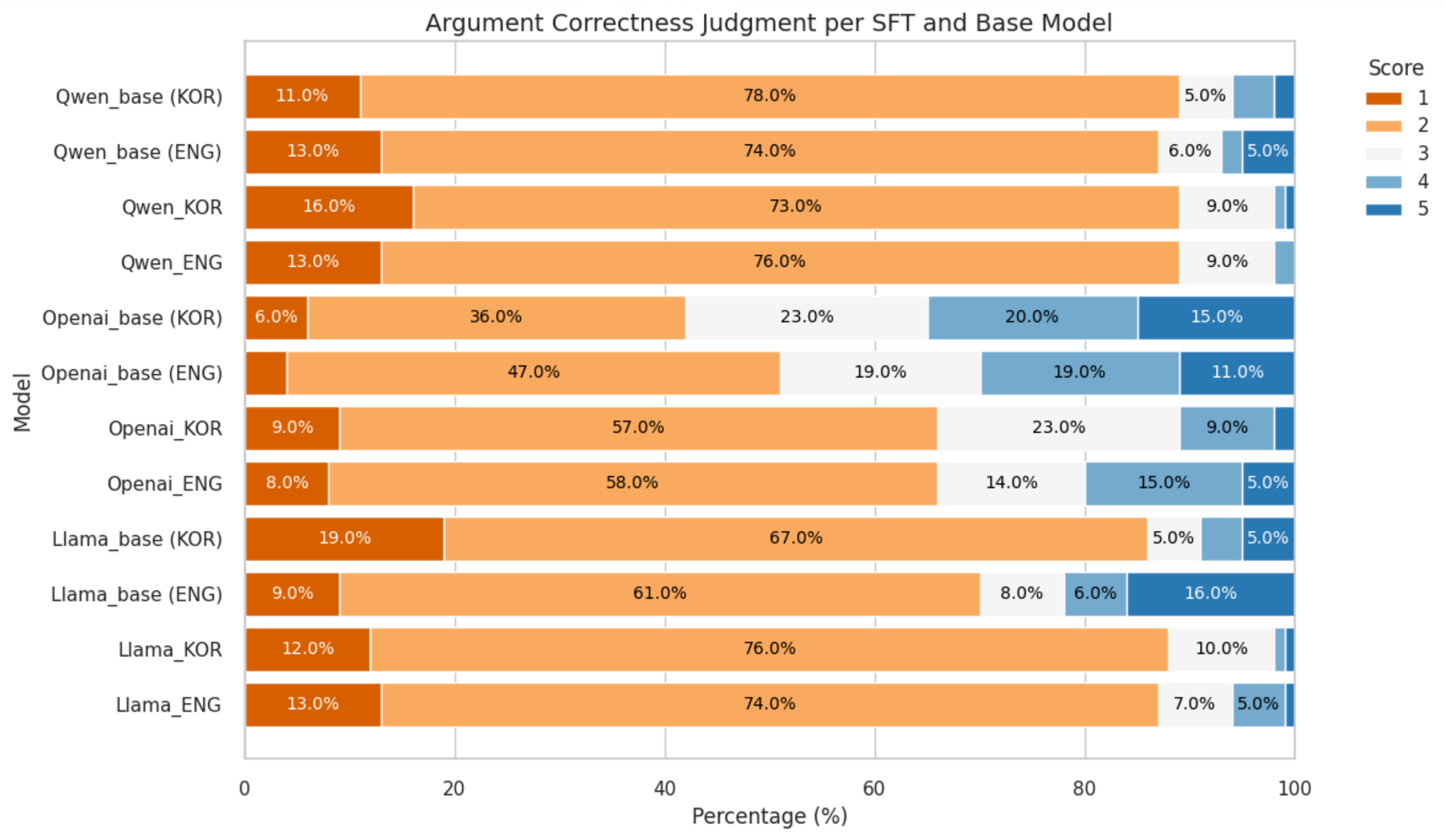
BERTSCORE



Key Takeaway: In terms of maximizing semantic similarity, fine-tuning can bridge the gap for base models.

RESULTS

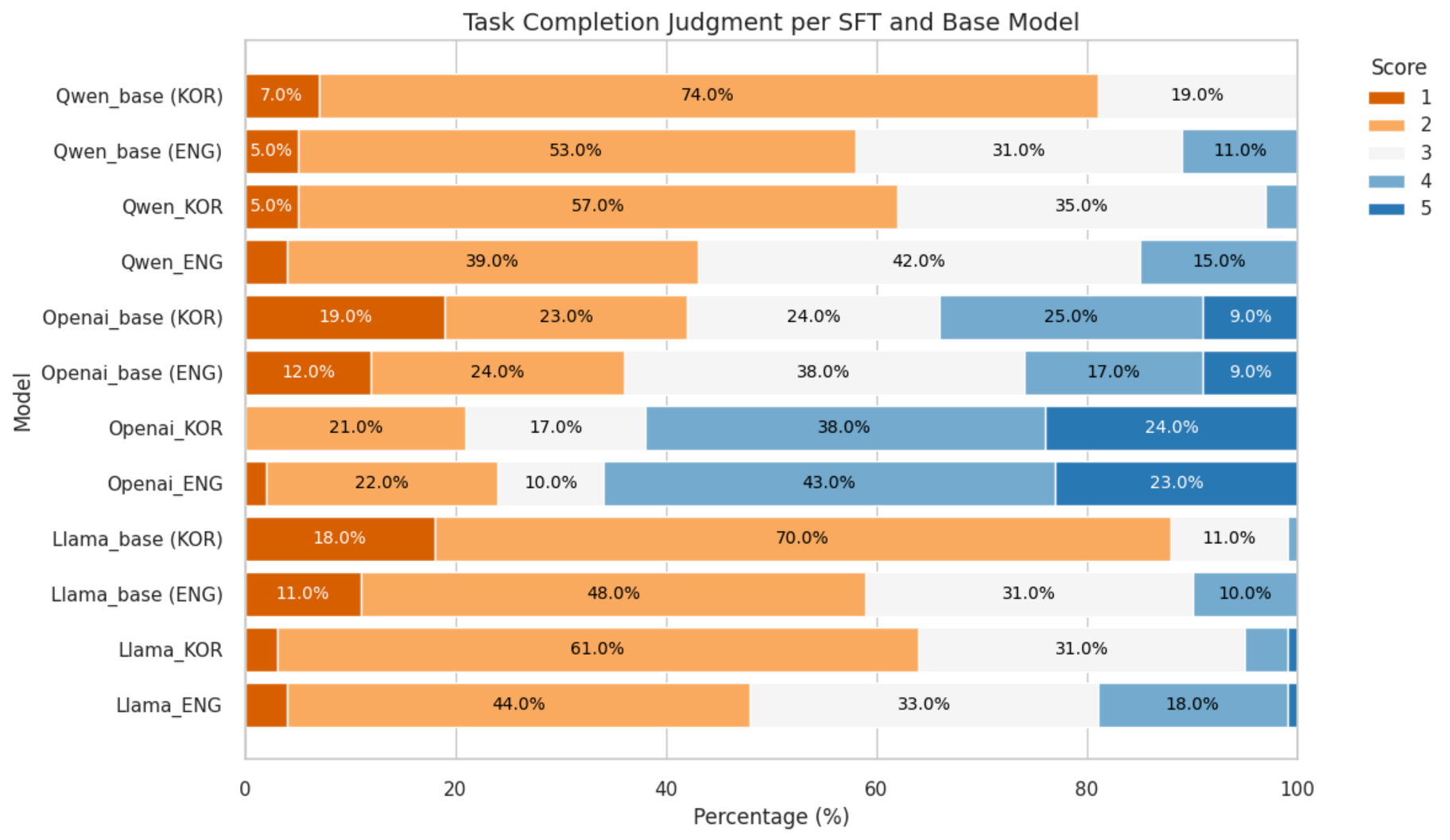
ARGUMENT CORRECTNESS



TAKEAWAY: Contradiction between automatic metrics. Slight improvement in performance when used in the OpenAI base model. Zero-shot prompting / RAG is the further direction in this study

RESULTS

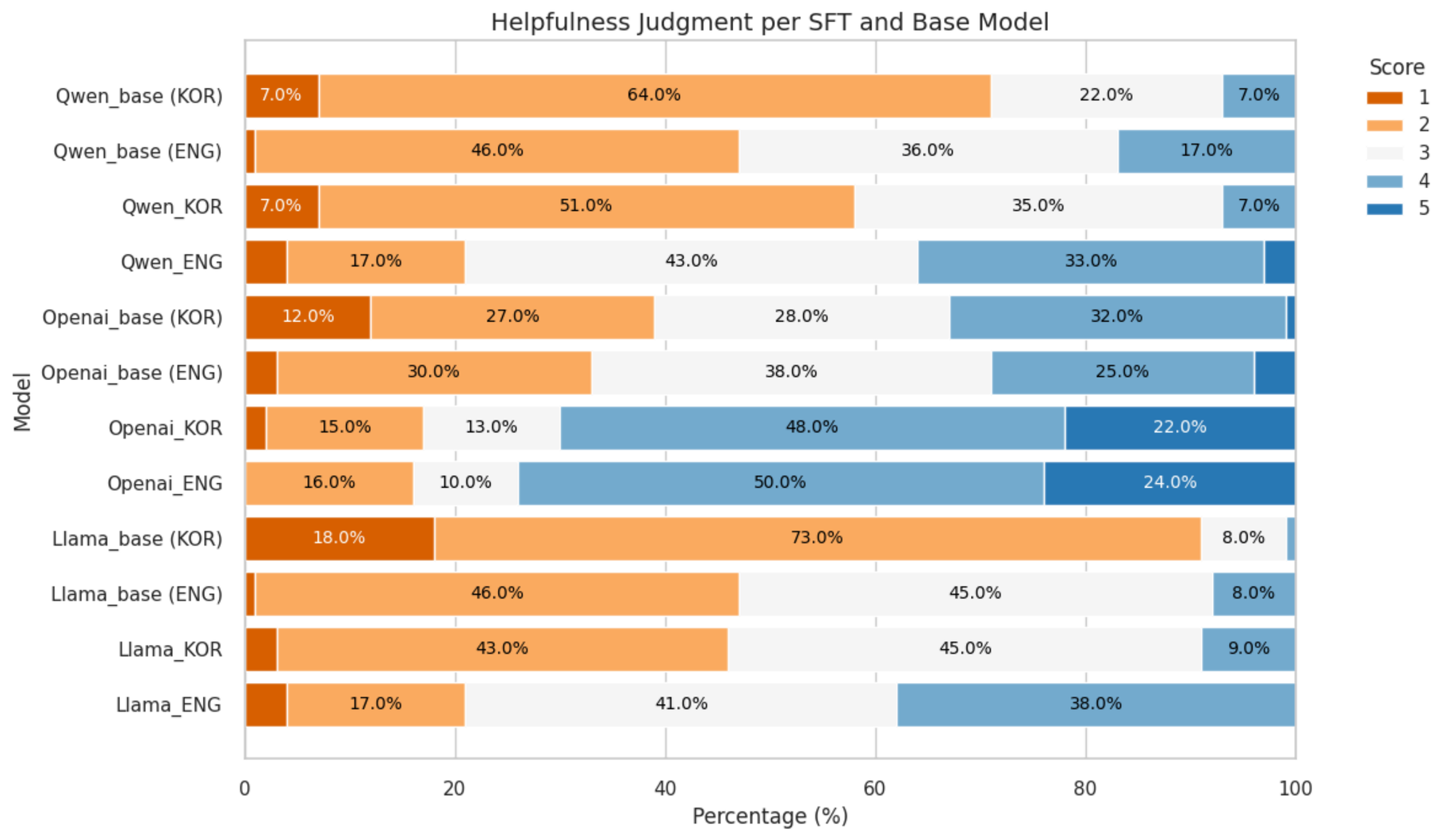
TASK COMPLETION



TAKEAWAY: Supervised fine-tuning dramatically improves task completion reliability across all model families. SFT

RESULTS

HELPFULNESS



TAKEAWAY: In smaller models, ENG system prompt significantly does better. While in larger SOTA models, KOR prompt usage does show negligible difference in helpfulness.

PROJECT

CONCLUSIONS

and Future Directions

1

The language of system instructions primarily shapes stylistic realization rather than the underlying recommendation quality. **Model family is important in the usage of a cross-lingual prompt.**

2

Automatic metrics based on reference answer now look to be unreliable as compared to chatbot metrics such as Judge correctness. **There is a need for better judgment metrics for general chatbots.**

3

Future direction for this study is to experiment the non-English system prompt setup with CoT, RAG, with hyperparameter tuning. **Zero-shot prompting with specialized base models is worth looking into.**

THANK YOU!

gabriel.geraldo27@gmail.com

PROJECT

REFERENCES

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. Findings of the Association for Computational Linguistics: EMNLP 2023 (January 2023). DOI:https://doi.org/10.18653/v1/2023.findings-emnlp.826

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (January 2023), 2695–2709. DOI:https://doi.org/10.18653/v1/2023.emnlp-main.163

Qikai Wei, Mingzhi Yang, Jinqiang Wang, Wenwei Mao, Jiabo Xu, and Huansheng Ning. 2024. TourLLM: Enhancing LLMs with Tourism Knowledge. arXiv.org. Retrieved November 14, 2025 from https://arxiv.org/abs/2407.12791v1

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (January 2023). DOI:https://doi.org/10.18653/v1/2023.emnlp-main.153

Ziqi Yin, Hao Wang, Kaito Horio, Daisuike Kawahara, and Satoshi Sekine. 2024. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024) (January 2024), 9–35. DOI:https://doi.org/10.18653/v1/2024.sicon-1.2