

# consultantAI

## Available models for korean language

1. Qwen3-4b (Highly efficient and SOTA model from alibaba and specially trained for conversational AI.) {tested}
2. IBM-Granite (Multilingual model from IBM, not tested yet!)
3. LFM2 (model with korean and english support, liquidAI, not tested yet!)
4. LFM2-1.2B-RAG (specialized model for RAG based usecases.)

Note: Large language models specialised only for korean and english language can be found with good research, otherwise korean language specific finetuning is required for the model to be efficient for day to day conversation and the described use case by the client.

## Steps

1. Choose a model
2. Test the model without finetuning with user interactions and real usecase (helps in requirement gathering, how the model should respond) (At least 1 week of interaction with the model and the history of the interaction will be used to analyse the requirements.)
3. Databases and File integrated and retrieved by the model as a reference and improvising its default response (KEY FEATURE).
4. Deployment (Xeon CPU based server with at least 32GB RAM) or GPU-providers like DigitalOcean, AWS, Google Cloud Platform to get the single H1000 GPU instances.
5. Model Size: Below 7B parameter (Above this size, GPU-based hosting will only work, CPU based servers won't work in this case and only work in case of models smaller than 7b.)

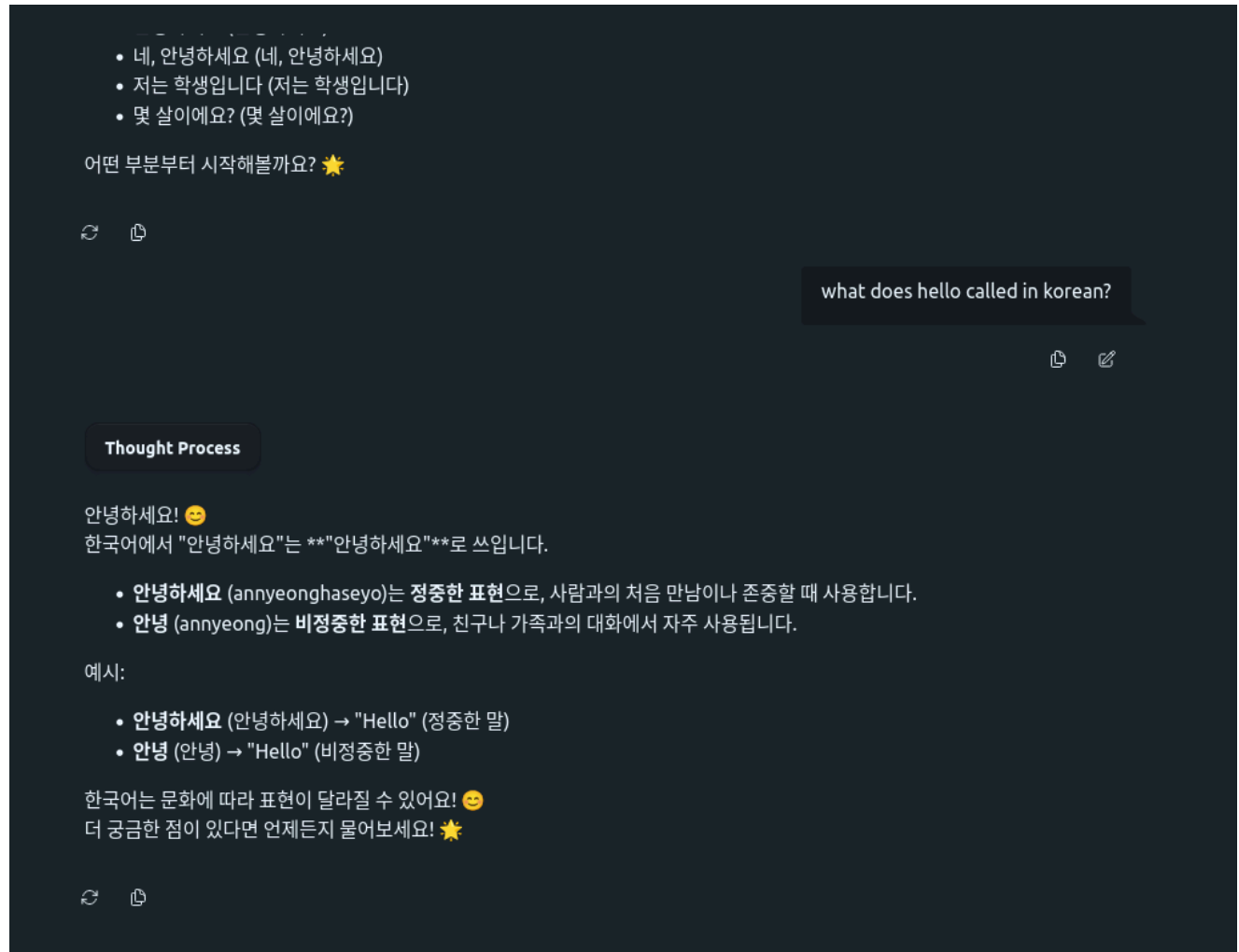
Note: Finetuning a model for this usecase is indeed very possible but the dataset development requires much more attention and manual effort as compared to the training process and training cost. So, the dataset development part must be considered seriously.

## Project Pipeline

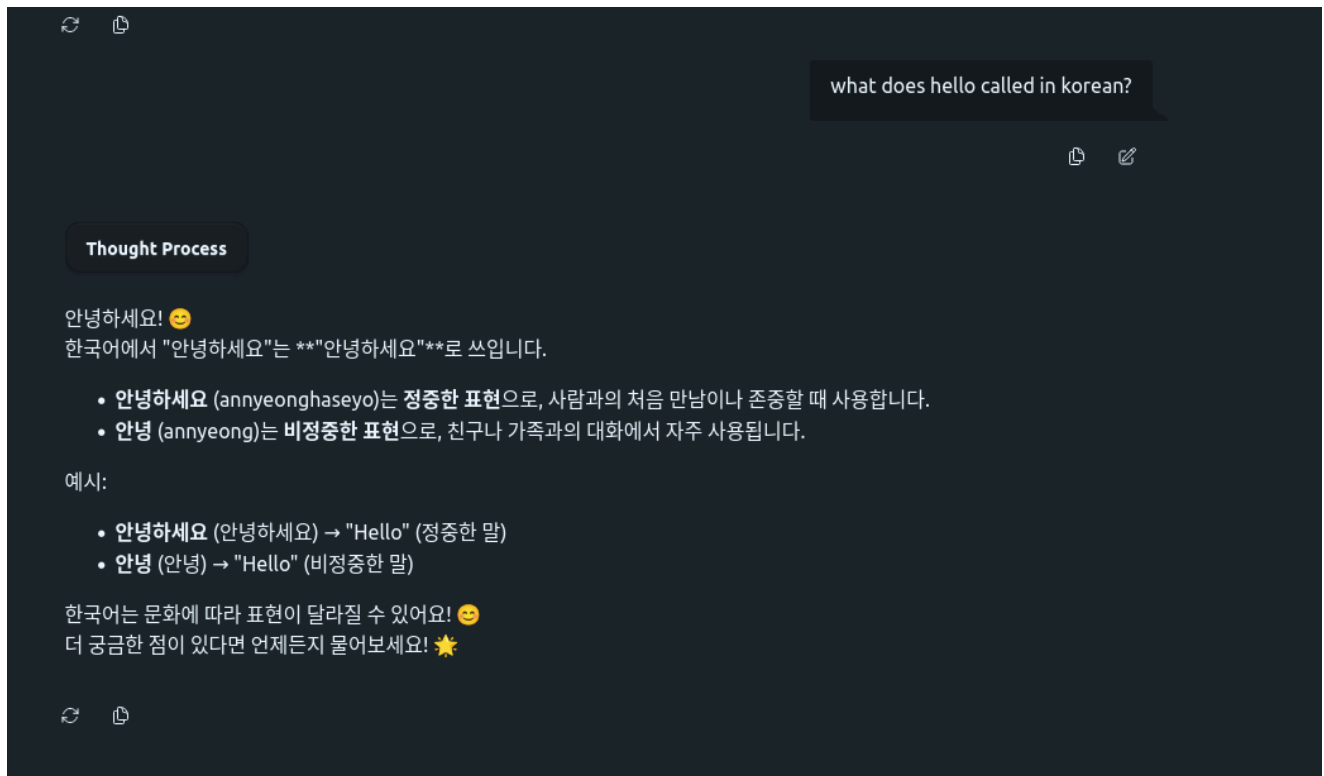
FrontEnd Chat UI (ReactJS) sends request (Text conversation) - API Endpoint takes the request to backend (Django, Node/ExpressJS) - Request converted to Python Data structure - Request (TEXT) sent to LLM hosted in GPU or CPU instance - LLM (think about the question and what the response should be [increase the response time, but coherent

responses are provided]) - LLM generates the response - Backend serves its via API - FrontEnd renders the response to the user.

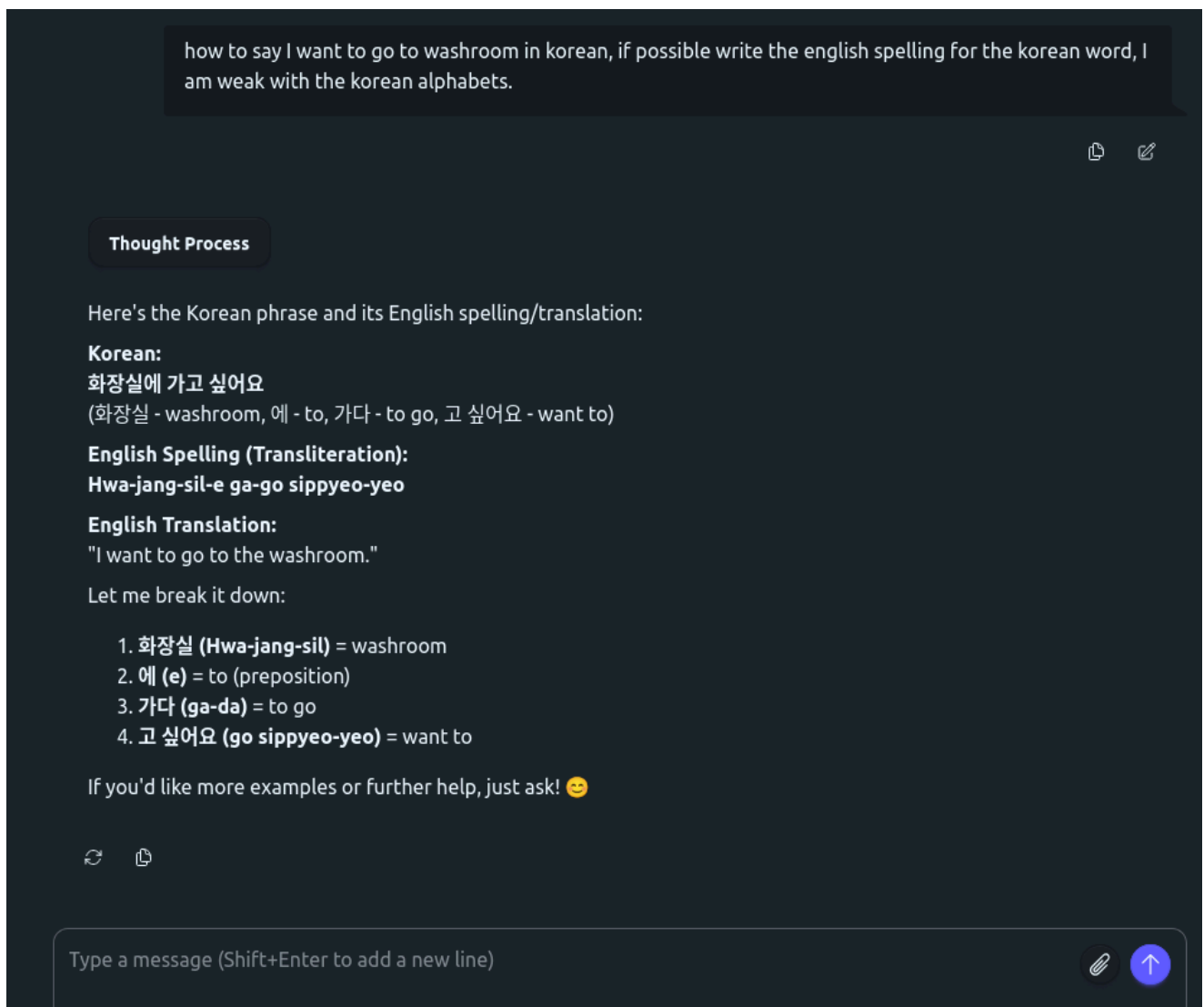
## Samples of conversations with Qwen3-4b



image\_1

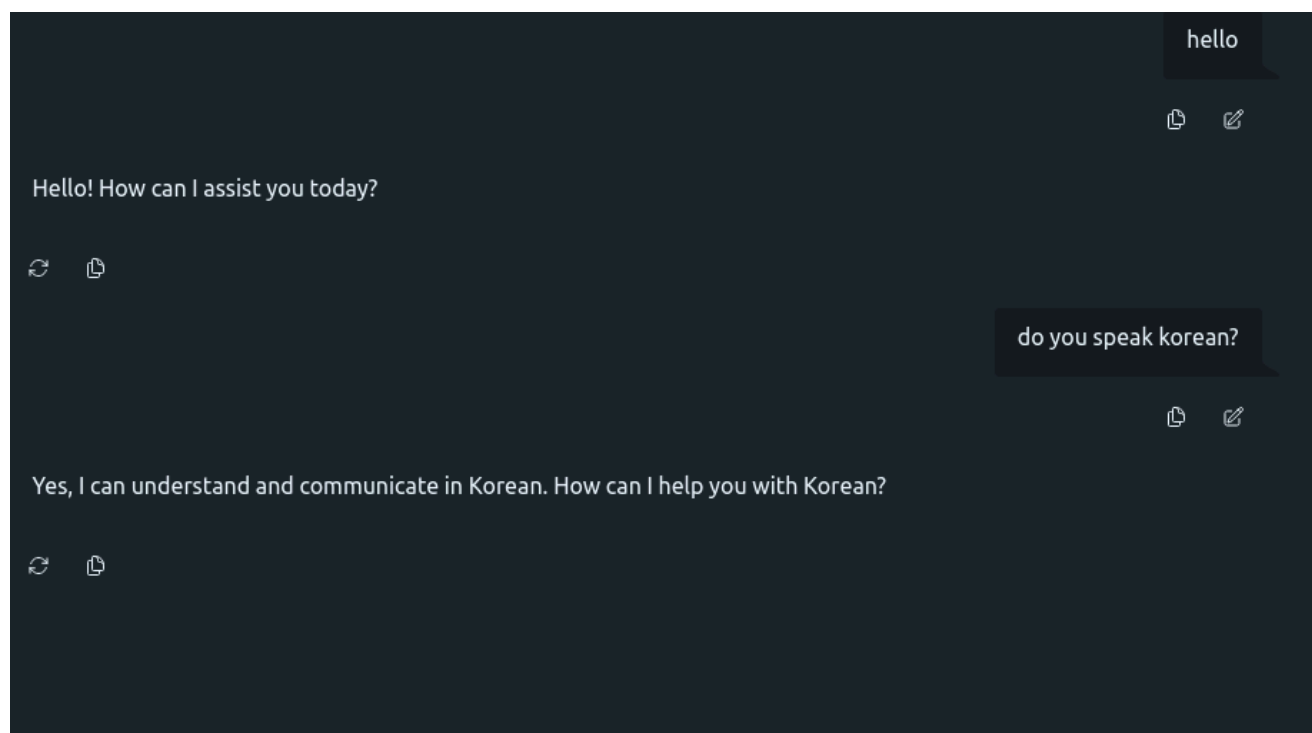


image\_2

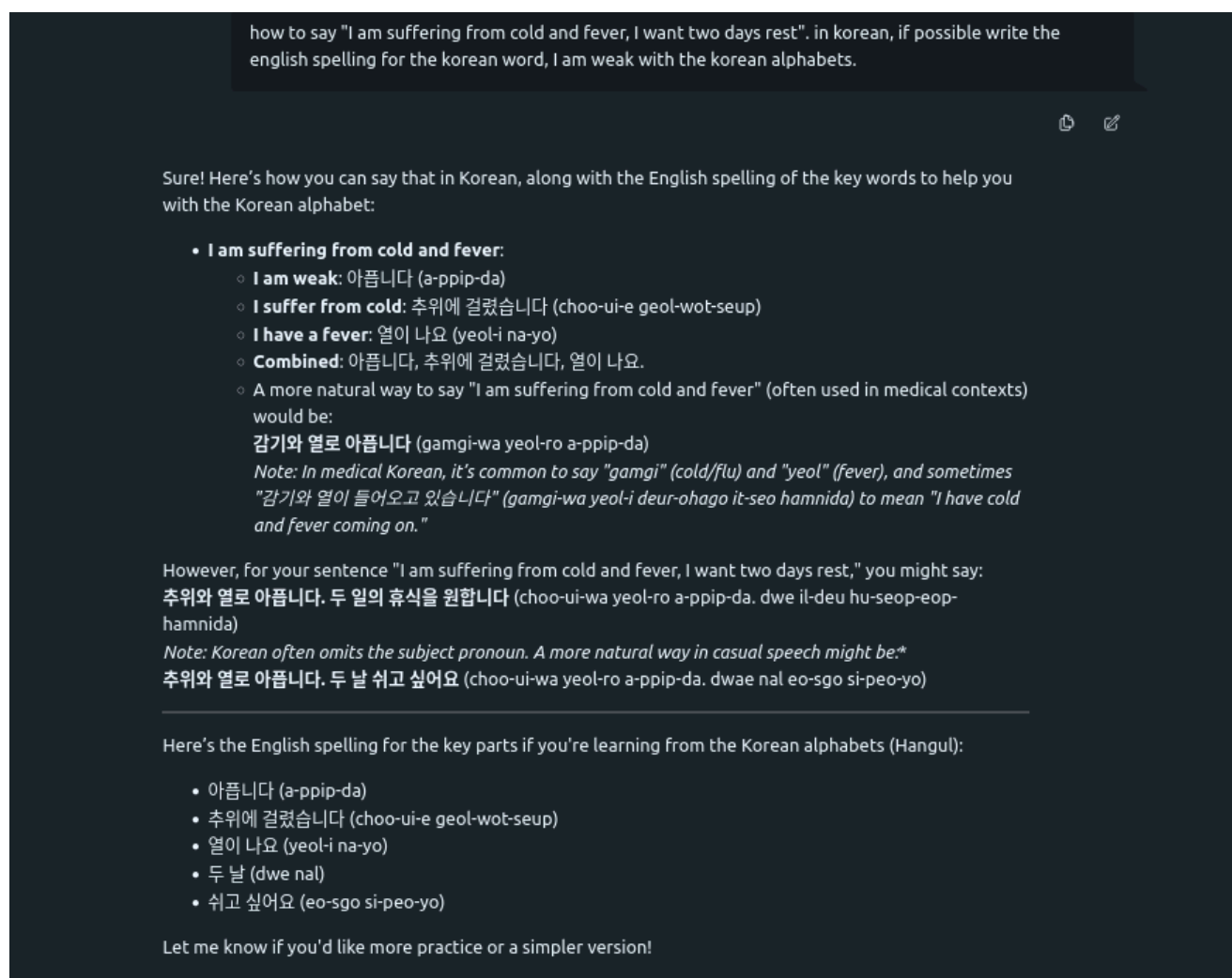


image\_3

## Sample conversation with LFM2-3.6B model



image\_1



image\_2



okay tell me greeting words in korean and their english transliteration.



Sure! Here are some common Korean greeting words along with their English transliterations:

1. 안녕하세요 (**An-nyeong-haseyo**) – Hello (formal, used throughout the day)
2. 안녕 (**An-yeong**) – Hi / Bye (casual, same as "hello" in informality)
3. 웁깁니다 (**Won-geungsida**) – Hello (very polite, often used in formal situations)
4. 어서 오세요 (**Eoseo o-seyo**) – Please come in
5. 조금 기다려 주세요 (**Jogeum gidaryeo juseyo**) – Wait a moment (to get someone's attention)
6. 갑사합니다 (**Gamsahamnida**) – Thank you
7. 안녕히 가세요 (**An-nyeong-hi gaseyo**) – Goodbye (when leaving)
8. 잘 가세요 (**Jal gaseyo**) – Goodbye (informal, to a friend)
9. 안녕히 계세요 (**An-nyeong-hi gyeo-seyo**) – Good evening (casual)
10. 잘 지내세요 (**Jal jina-seyo**) – Take care / Have a good time

---

Here's a quick breakdown:

image\_3