# New Models for Enhanced Performance on Specialized Tasks

## Advanced Math Reasoning Models

Open-source language models fine-tuned for mathematical reasoning have seen significant improvements recently. For example, **WizardMath** (based on Llama-2) and **MetaMath** are two new models explicitly optimized for math. The 13B-parameter version of MetaMath achieves about **72.3% accuracy on GSM8K** (a benchmark of grade-school math word problems) and **22.4% on the MATH competition dataset**, substantially outperforming WizardMath-13B (63.9% on GSM8K, 14.0% on MATH)[[1]](https://huggingface.co/meta-math/MetaMath-13B-V1.0#:~:text=WizardMath,7B%2069.2%2030.0)[[2]](https://dataloop.ai/library/model/skywork_skywork-13b-math/#:~:text=Model%20GSM8K%20MATH%20CMATH%20%3D%3DLLaMA,Math%2072.33%2016.98%2077.27). These models use techniques like chain-of-thought prompting and reinforcement learning from synthetic math problems to greatly boost arithmetic and algebraic problem-solving ability[[3]](https://openreview.net/forum?id=mMPMHWOdOy#:~:text=Abstract%3A%20Large%20language%20models%20,Remarkably%2C%20WizardMath)[[4]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=Model%20Checkpoint%20Paper%20GSM8k%20MATH,7%20Demo%20Llama%202). Another model, **WizardMath 70B**, even manages to slightly surpass ChatGPT-3.5 on GSM8K (81.6% vs 80.8%) when run with sufficient hardware[[5]](https://openreview.net/forum?id=mMPMHWOdOy#:~:text=this%20paper%2C%20we%20present%20WizardMath%2C,Additionally%2C%20our%20preliminary%20exploration%20highlights)[[6]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=The%20following%20figure%20shows%20that,7). For a local setup, a **13B model** like MetaMath-13B or WizardMath-13B offers a good balance between size and performance, given your 16 GB GPU – these can be quantized to 4-bit and run within ~7–8 GB VRAM with minimal quality loss[[7]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=Branch%20Bits%20GS%20Act%20Order,VRAM%20than%2032g%2C%20but%20with)[[4]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=Model%20Checkpoint%20Paper%20GSM8k%20MATH,7%20Demo%20Llama%202). In summary, adopting a math-specialized LLM (≥10B parameters) should markedly improve the system’s quantitative reasoning capabilities.

## Improved Embedding and Re-ranking Models

For embedding-based retrieval, modern **bi-encoder models** such as Microsoft’s **E5** family have become state-of-the-art. In particular, the **multilingual E5-large** model (≈600M parameters) provides universal sentence embeddings for search and semantic similarity across languages[[8]](https://huggingface.co/intfloat/multilingual-e5-large#:~:text=Benchmark%20Results%20on%20Mr). This model achieved an average **MRR@10 of 70.5** on the Mr.TyDi multilingual retrieval benchmark, a dramatic improvement over traditional BM25 retrieval (33.3 MRR)[[8]](https://huggingface.co/intfloat/multilingual-e5-large#:~:text=Benchmark%20Results%20on%20Mr). Using E5, you can encode Bible verses, questions, or theology texts (English, Greek, Hebrew, etc.) into a vector space where relevant pairs have high cosine similarity. Indeed, an earlier project for Bible study used **BAAI’s BGE embeddings** with a vector database to semantically match verses to queries[[9]](https://news.ycombinator.com/item?id=38040591#:~:text=Introducing%20Biblos%2C%20a%20simple%20tool,app), showing that modern embedding models capture scriptural semantics beyond keyword overlap.

After retrieving candidate passages via embeddings, a **re-ranking model** can further boost precision. Lightweight **cross-encoders** fine-tuned on MS MARCO or similar datasets are readily available for this task. For example, the **ms-marco-MiniLM-L12-v2** cross-encoder (110M parameters) can take a query and a passage and output a relevance score – it achieves about **39.0 MRR@10 on MS MARCO**, outperforming smaller cross-encoders and even some larger BERT-based rankers[[10]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=Model,L2%2067.43%2030.15%209000). These models are efficient (processing ~950 passages/second on a GPU)[[11]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=cross,L6%2069.57%2036.13%20680), making them suitable to re-rank top-$k$ results. If multilingual re-ranking is needed, there are also cross-encoders based on multilingual BERT (e.g. a multilingual MiniLM) achieving around **35–36 MRR** on English MS MARCO[[12]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=nboost%2Fpt,msmarco%2068.40%2035.54%20330) – likely a bit lower in English accuracy, but capable of handling non-English text[[13]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=nboost%2Fpt,msmarco%2072.82%2037.88%20720). In practice, you might embed both original-language and translated passages, then use a cross-encoder on the translated text for final ranking. Overall, combining a powerful bi-encoder like E5 for recall with a fast cross-encoder for precision will ensure high-quality retrieval and relevance ranking in the system.

## Models for Theology and Ancient Languages

Developing a “systematic theology” model that handles ancient Biblical languages is challenging, as no major LLM to date has been specifically trained on extensive Koine Greek, Biblical Hebrew, or Aramaic corpora. However, recent research trends show it is possible to adapt open models to low-resource languages. For example, **DictaLM 2.0** is a new 7B-parameter model based on Mistral, specialized for Hebrew: it was continuously pre-trained on ~100 billion tokens of Hebrew text (along with English) and then instruction-tuned for tasks like question-answering and translation[[14]](https://arxiv.org/html/2407.07080v1#:~:text=techniques%20that%20differ%20significantly%20from,work%20not%20only%20addresses%20the)[[15]](https://arxiv.org/html/2407.07080v1#:~:text=In%20response%20to%20these%20challenges%2C,ensure%20effective%20learning%20and%20adaptation). This kind of targeted training dramatically improves a model’s fluency and understanding in Hebrew. A similar approach could be considered for Biblical Hebrew and Koine Greek – i.e. extending a base model’s tokenizer to include ancient vocabulary, and fine-tuning on parallel texts (original scriptures and their translations). So far, though, such an ancient-language LLM has not been publicly released.

In the absence of a dedicated ancient-language model, a practical workaround is to integrate high-quality **machine translation** tools into your pipeline. Meta’s **No Language Left Behind (NLLB)** project and the Helsinki-NLP **OPUS-MT** models provide free translation for many low-resource languages. For instance, OPUS-MT has an English-Greek model that even includes an option for ancient Greek (grc locale) – though its out-of-the-box performance on Koine Greek is virtually nil (only 0.1 BLEU on an English→Ancient Greek test)[[16]](https://huggingface.co/Helsinki-NLP/opus-mt-en-grk#:~:text=testset%20BLEU%20chr,test.eng.multi%2045.6%200.677) without further fine-tuning. Similarly, OPUS-MT or NLLB models for Hebrew might handle modern Hebrew well and could be adapted to Biblical Hebrew. By coupling such a translator with your LLM, the system could translate a Greek or Hebrew passage on-the-fly to English, allow the LLM to reason in English, and even translate an answer or quotation back if needed. This two-step approach (translation + LLM) ensures you leverage the LLM’s strengths in reasoning and doctrinal knowledge, while not being stymied by its weaknesses in untrained ancient scripts.

Finally, to ensure the model’s responses align with Christian doctrine and use correct theological terminology, you might look at community fine-tuning efforts. Projects like **ChristianGPT** have fine-tuned Llama-2 on tens of thousands of Bible-based Q&A pairs to "align" the model with orthodox Christian perspectives[[17]](https://huggingface.co/AiForTheChurch#:~:text=The%20family%20of%20models%20presented,tuned%20on%20denomination%20specific%20datasets). For example, a base Llama-2-7B was trained on 30k filtered Q&As from a site (with GPT-3.5 generated answers) to produce a faithful Christian assistant, with further fine-tunes for specific denominations[[17]](https://huggingface.co/AiForTheChurch#:~:text=The%20family%20of%20models%20presented,tuned%20on%20denomination%20specific%20datasets). In your case, starting with a larger base (13B+ for better quality) and incorporating original-language data and translations during fine-tuning could yield a powerful local theology model. In summary, while no single off-the-shelf model meets all requirements (ancient Greek, Aramaic, Hebrew plus theological training), a combination of **an open LLM (>=13B)** fine-tuned on Biblical content and **auxiliary translation models or extended multilingual training** appears to be the best path. This hybrid strategy would allow the **local model** to interpret ancient texts on the fly and provide doctrinally sound answers – effectively acting as a specialized “Christian GPT” to support your current Christian AI system.

[[1]](https://huggingface.co/meta-math/MetaMath-13B-V1.0#:~:text=WizardMath,7B%2069.2%2030.0) meta-math/MetaMath-13B-V1.0 · Hugging Face

<https://huggingface.co/meta-math/MetaMath-13B-V1.0>

[[2]](https://dataloop.ai/library/model/skywork_skywork-13b-math/#:~:text=Model%20GSM8K%20MATH%20CMATH%20%3D%3DLLaMA,Math%2072.33%2016.98%2077.27) Skywork 13B Math · Models · Dataloop

<https://dataloop.ai/library/model/skywork_skywork-13b-math/>

[[3]](https://openreview.net/forum?id=mMPMHWOdOy#:~:text=Abstract%3A%20Large%20language%20models%20,Remarkably%2C%20WizardMath) [[5]](https://openreview.net/forum?id=mMPMHWOdOy#:~:text=this%20paper%2C%20we%20present%20WizardMath%2C,Additionally%2C%20our%20preliminary%20exploration%20highlights) WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct | OpenReview

<https://openreview.net/forum?id=mMPMHWOdOy>

[[4]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=Model%20Checkpoint%20Paper%20GSM8k%20MATH,7%20Demo%20Llama%202) [[6]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=The%20following%20figure%20shows%20that,7) [[7]](https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ#:~:text=Branch%20Bits%20GS%20Act%20Order,VRAM%20than%2032g%2C%20but%20with) TheBloke/WizardMath-13B-V1.0-GPTQ · Hugging Face

<https://huggingface.co/TheBloke/WizardMath-13B-V1.0-GPTQ>

[[8]](https://huggingface.co/intfloat/multilingual-e5-large#:~:text=Benchmark%20Results%20on%20Mr) intfloat/multilingual-e5-large · Hugging Face

<https://huggingface.co/intfloat/multilingual-e5-large>

[[9]](https://news.ycombinator.com/item?id=38040591#:~:text=Introducing%20Biblos%2C%20a%20simple%20tool,app) Show HN: Biblos – Semantic Bible Embedded Vector Search and Claude LLM | Hacker News

<https://news.ycombinator.com/item?id=38040591>

[[10]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=Model,L2%2067.43%2030.15%209000) [[11]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=cross,L6%2069.57%2036.13%20680) [[12]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=nboost%2Fpt,msmarco%2068.40%2035.54%20330) [[13]](https://www.sbert.net/docs/pretrained-models/ce-msmarco.html#:~:text=nboost%2Fpt,msmarco%2072.82%2037.88%20720) MS MARCO Cross-Encoders — Sentence Transformers documentation

<https://www.sbert.net/docs/pretrained-models/ce-msmarco.html>

[[14]](https://arxiv.org/html/2407.07080v1#:~:text=techniques%20that%20differ%20significantly%20from,work%20not%20only%20addresses%20the) [[15]](https://arxiv.org/html/2407.07080v1#:~:text=In%20response%20to%20these%20challenges%2C,ensure%20effective%20learning%20and%20adaptation) Adapting LLMs to Hebrew: Unveiling DictaLM 2.0 with Enhanced Vocabulary and Instruction Capabilities

<https://arxiv.org/html/2407.07080v1>

[[16]](https://huggingface.co/Helsinki-NLP/opus-mt-en-grk#:~:text=testset%20BLEU%20chr,test.eng.multi%2045.6%200.677) Helsinki-NLP/opus-mt-en-grk · Hugging Face

<https://huggingface.co/Helsinki-NLP/opus-mt-en-grk>

[[17]](https://huggingface.co/AiForTheChurch#:~:text=The%20family%20of%20models%20presented,tuned%20on%20denomination%20specific%20datasets) AiForTheChurch (AI for the Church)

<https://huggingface.co/AiForTheChurch>