Potential Large Language Models

The following generative LLMs can be integrated with an information retrieval architecture.

LLaMa 2
LLaMa 2 is a family of transformer-based autoregressive causal language models developed by Meta. Compared to major proprietary models, LLaMa2 is of modest size but is a highly efficient model optimized for lower resource consumption (1).

- Developed by Meta
- Open-source
- Cost: Free for commercial use, restrictions apply for companies with more than 700 million monthly active users
- Available in variants of 7 billion, 13 billion, and 70 billion parameters
- Pros
    - Excels with minimal hardware demands
    - Outperforms other open-source language models at comprehension, reasoning, general intelligence, math, and more
- Cons
    - Limited multilingual capabilities
    - Parameter size is significantly less than proprietary models, which can impact its generative abilities
    - Includes preventative behavior that can lead to refused queries if they are deemed even mildly inappropriate
    - Poor coding ability
    - May require fine-tuning to implement

Falcon
Falcon is a family of open-source large language models developed by the Technology Innovation Institute (TII) in the UAE. It is designed for a wide range of natural language processing tasks, including text generation, summarization, translation, and more. Its large size allows it to handle complex queries and generate high-quality, coherent text (2).

- Developed by the Technology Innovation Institute (TII)
- Open-source
- Cost: Free for commercial use under the Apache 2.0 license (smaller variants) or TII (Technology Innovation Institute) OpenRAIL-M License (180 billion parameter variant)
- Available in 7 billion, 40 billion, and 180 billion parameter variants
- Pros
    - Outperforms Meta's LLaMa 2 (3)
    - Performs exceptionally well in tasks such as reasoning, coding, proficiency, and knowledge tests
    - Available under a permissive license allowing for commercial use
- Cons
    - Requires significant computational resources to run
    - May require fine-tuning to implement

T5

The T5 language model frames all natural language processing tasks as a text-to-text problem, meaning both the input and output are in the form of text. This allows the model to be highly versatile across a wide range of tasks such as translation, summarization, and question answering (4).

- Developed by Google Research
- Open-source
- Cost: Free for commercial use under the Apache 2.0 license
- Available in 60 million, 220 million, 770 million, 3 billion, and 11 billion parameter variants
- Pros
    - Allows practitioners to guide the model's behavior and tailor its responses to specific tasks (5)
    - Captures contextual dependencies, leading to more accurate and context-aware language understanding and generation
    - Exhibits strong performance in multilingual tasks
- Cons
    - Requires significant computational resources for training and inference
    - Fine-tuning T5 on task-specific data often necessitates a sizable amount of labeled data
    - The model may suffer from overfitting when fine-tuned on small or biased datasets

Bloom

BLOOM is a multilingual large language model developed by a global team of researchers. It supports numerous human languages and programming languages, offering open access for study and use. Unlike many proprietary models, BLOOM is fully transparent and available under a Responsible AI License, making it accessible for both research and practical applications. It is integrated with Hugging Face, it's easy to use, and its capabilities will continue to evolve through ongoing community collaboration (6).

- Developed by the BigScience project
- Open-source
- Cost: Free for commercial use under the BigScience Open RAIL-M license
- Available in 560 million, 1 billion, 3 billion, 7 billion, and 176 billion parameter variants
- Pros
    - Ideal for applications involving non-English languages and global use cases
    - Supports 46 human languages and 13 programming languages
- Cons
    - May produce incorrect information as if it were factual (7)
    - May generate hateful, abusive, or violent language
    - May generate irrelevant or repetitive outputs

GPT-4

GPT-4 is a state-of-the-art multimodal large language model developed by OpenAI. It builds on the capabilities of its predecessor, GPT-3, with improved performance in understanding and generating human-like text. GPT-4 can handle a wide range of tasks, including text generation, summarization, translation, and more, making it highly versatile for various applications (8).

- Developed by OpenAI
- Proprietary
- Cost: $2.50 ~ $5.00 per 1M input tokens, $7.50 ~ $15.00 per 1M output tokens
- Available versions (smallest to largest): GPT-4o mini, GPT-4o
- Pros
    - Generates more accurate and context-aware text compared to earlier models
    - Suitable for a broad array of use cases across different industries
    - Continuous updates
    - No fine-tuning necessary
- Cons
    - Requires API integration
    - Usage fees and limitations on commercial deployment
    - Can require significant computational resources

Claude 3

Claude 3 is an advanced language model designed to enhance user interaction with improved performance and longer response capabilities. It is accessible via API and a beta website, claude.ai. It shows significant advancements in tasks involving coding, math, and reasoning, scoring well on standardized tests such as the Bar exam and GRE. It functions as a user-friendly personal assistant, enabling it to handle extensive texts and generate longer documents efficiently (9).

- Developed by Anthropic
- Proprietary
- Cost: $20.00+ per person per month
- Available versions (smallest to largest): Haiku, Sonnet, Opus
- Pros
    - Strong focus on creating AI that closely aligns with human values and safety
    - Demonstrates enhanced capabilities in reasoning, coding, and mathematical tasks compared to previous models
    - Claude 3 has been reported to surpass GPT-4 in certain benchmarks
- Cons
    - As a newer release, there may be limited community support or resources compared to more established models (10)
    - May require substantial computational resources
    - The quality of outputs can vary based on the clarity and specificity of user prompts
    - Requires API integration
    - Monthly subscription per user

PaLM 2

PaLM 2 is Google's next-generation language model, offering enhanced multilingual, reasoning, and coding capabilities. It outperforms previous models, including PaLM, in tasks such as code generation, math, classification, question answering, translation, and natural language generation. This improvement is attributed to its compute-optimal scaling, refined dataset mixture, and architectural advancements. Built with a focus on responsible AI, PaLM 2 undergoes rigorous evaluations for potential biases and harms. It powers advanced models like Sec-PaLM and is integrated into generative AI tools such as the PaLM API (11).

- Developed by Google
- Proprietary
- Cost: No upfront cost but usage fees apply (not listed)
- Available versions (smallest to largest): Gecko, Otter, Bison, Unicorn (12)
- Pros
    - Extensive language support (13)
    - Integrates seamlessly with other Google products
    - Prowess in reasoning and code generation, demonstrates the potential to assist in complex problem-solving scenarios
- Cons
    - Requires API integration
    - Needs substantial computational resources and energy consumption for training
    - Diverse training data used for pretraining model raises questions about data privacy and potential biases
    - Usage fees

Jurassic-2

Jurassic-2 is a large language model designed for natural language understanding and generation. It follows the success of the earlier Jurassic-1 models, with Jurassic-2 offering improvements in language fluency, coherence, and contextual understanding. Jurassic-2 is available in multiple variants, with different sizes optimized for various use cases, making it versatile for tasks such as content generation, question answering, and code writing (14).

- Developed by AI21 Labs
- Proprietary
- Cost: $0.0001 ~ $0.002 per 1K input tokens, $0.0005 ~ $0.010 per 1K output tokens
- Available versions (smallest to largest): Light, Mid, Ultra
- Pros
    - Handles a wide range of tasks, including text generation, question answering, and translation
    - Supports multiple human languages
- Cons
    - Requires API integration
    - Jurassic-2 is relatively new, and there may be limitations in areas such as complex reasoning compared to other models
    - Full customization or control is limited to the licensing terms set by AI21 Labs
    - Usage fees

Works Cited

(1) https://www.simform.com/blog/llama-2-comprehensive-guide/
(2) https://huggingface.co/tiiuae/falcon-180B
(3) https://falconllm.tii.ae/falcon-180b.html
(4) https://huggingface.co/docs/transformers/en/model_doc/t5
(5) https://databasecamp.de/en/ml-blog/t5-model
(6) https://bigscience.huggingface.co/blog/bloom
(7) https://huggingface.co/bigscience/bloom
(8) https://platform.openai.com/docs/models/continuous-model-upgrades
(9) https://www.anthropic.com/claude
(10) https://blog.gopenai.com/claude-llm-pros-and-cons-compared-with-other-llms-14f6a89cf50f
(11) https://ai.google/discover/palm2/
(12) https://blog.google/technology/ai/google-palm-2-ai-large-language-model/
(13) https://ai.plainenglish.io/the-language-model-showdown-googles-new-palm-2-challenges-gpt-4-for-ai-dominance-cf738773b0e2
(14) https://docs.ai21.com/docs/jurassic-2-models