

Chatting With Docs

Privacy-friendly alternatives to OpenAI

Hidéo SNES

wave@hideosnes.online

<https://hideosnes.online>

FV: homahuki.eu@hideosnes



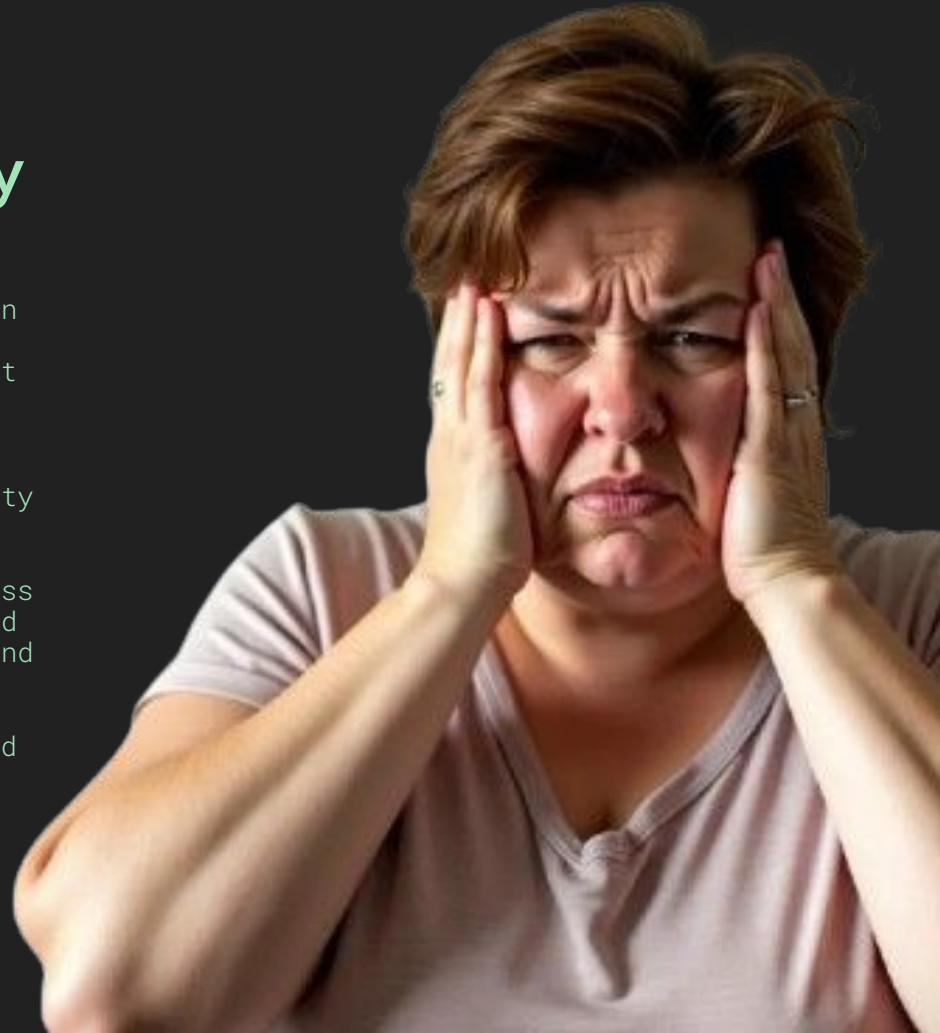
Privacy & Data Integrity

Data Exposure Risk: Personal information stored in databases and documents can be accidentally exposed when RAG-LLMs retrieve and combine information from various sources. **If an error occurs** a person will read your chat history. It may even have legal consequences (eg. NDA).

Identity Tracking: Interaction with a RAG-LLM service, creates a digital trail that can be linked to an identity through patterns in queries and responses. (Journalism)

Sensitive Content Access: RAG-LLMs can potentially access and process sensitive private information from unsecured databases, including medical records, financial data, and personal communications.

Long-term Storage: Queries and the information retrieved to answer them can be permanently stored in training data, creating a permanent record of interests, behaviors, and personal details. (Thumbs up or down!)



LLM Alphabet Soup #1

Model Size Indicator

B (Billion): Indicates number of parameters (70B = 70 billion parameters)

M (Million): Smaller models (125M = 125 million parameters)

x (Multiplier): Indicates mixture of experts architecture (8x7B = 8 expert groups with 7B parameters each)

Version Numbers

vX (Version): Model version number (v0.1 = version 0, revision 1)

X.X (Decimal): Major version number (3.2 = third major, second minor version)

Architecture & Training

R1, R2, etc.: Release number of the model

o1, o2, etc.: Optimization version number

MoE (Mixture of Experts): Architecture type using multiple specialized models

Task-Specific Indicators

Instruct: Fine-tuned to follow instruct.

Chat: Optimized for conversational tasks

Medical: Specialized for healthcare

Code: Optimized for programming tasks

LLM Alphabet Soup #2

Size Categories

mini: Compact version of the model
flash: Lightweight, fast version
S (Small): Smallest quantized version
M (Medium): Balanced size and performance
L (Large): Full-size version

Quantization Formats (*)

GGML: GPT-Generated Mixed List format
GGUF: GPT-Generated Unified Format
QX (Q4, Q8,...): Quantization of X bits
_0, _1: Uniform quantization method
_K: K-quant method (advanced rounding)

Optimization Indicators

INT8: 8-bit integer quantization
(Memory optimized)
FP16: 16-bit floating-point
(Balanced performance)
FP32: 32-bit floating-point
(Maximum precision)

File Formats

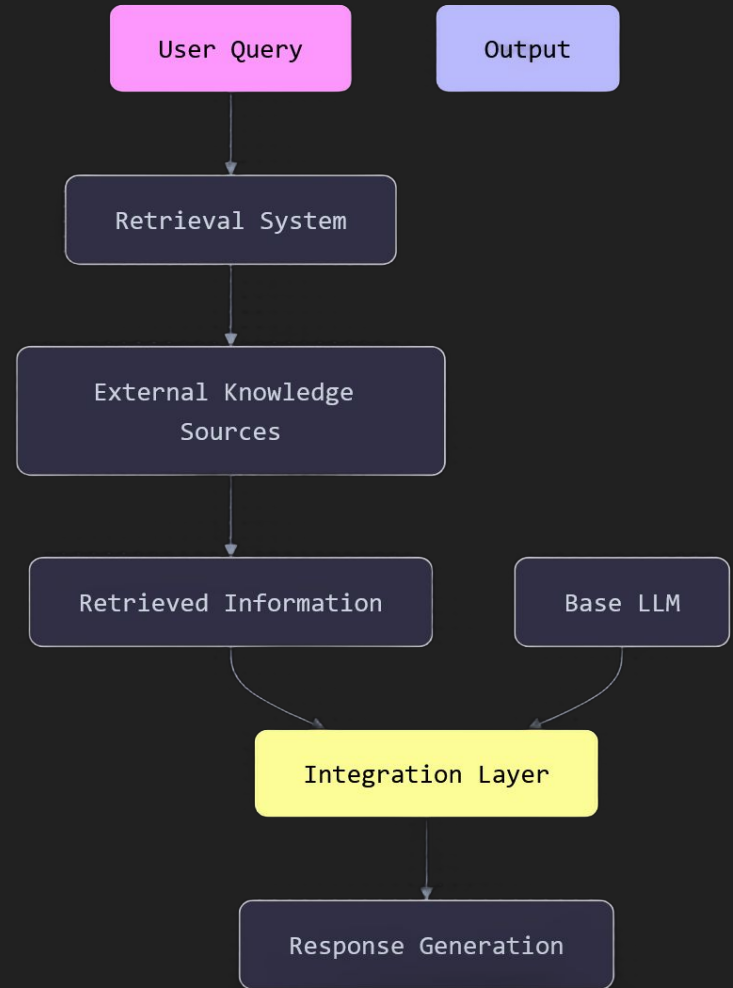
.bin: Binary model file
.pt: PyTorch model file
.onnx: Open Neural Network Exchange
.tflite: TensorFlow Lite format

** It's like converting a high-resolution image to a lower resolution version (lower size & detail)*

RAG-LLM

(Retrieval-Augmented Generative Large Language Model)

1. The Base LLM works alongside the Integration Layer to **process both retrieved information and generate responses**.
2. External Knowledge Sources represent **databases, documentation, or web content** that the model can draw upon.
3. The Integration Layer combines retrieved information with the base model's knowledge to **generate coherent responses**.

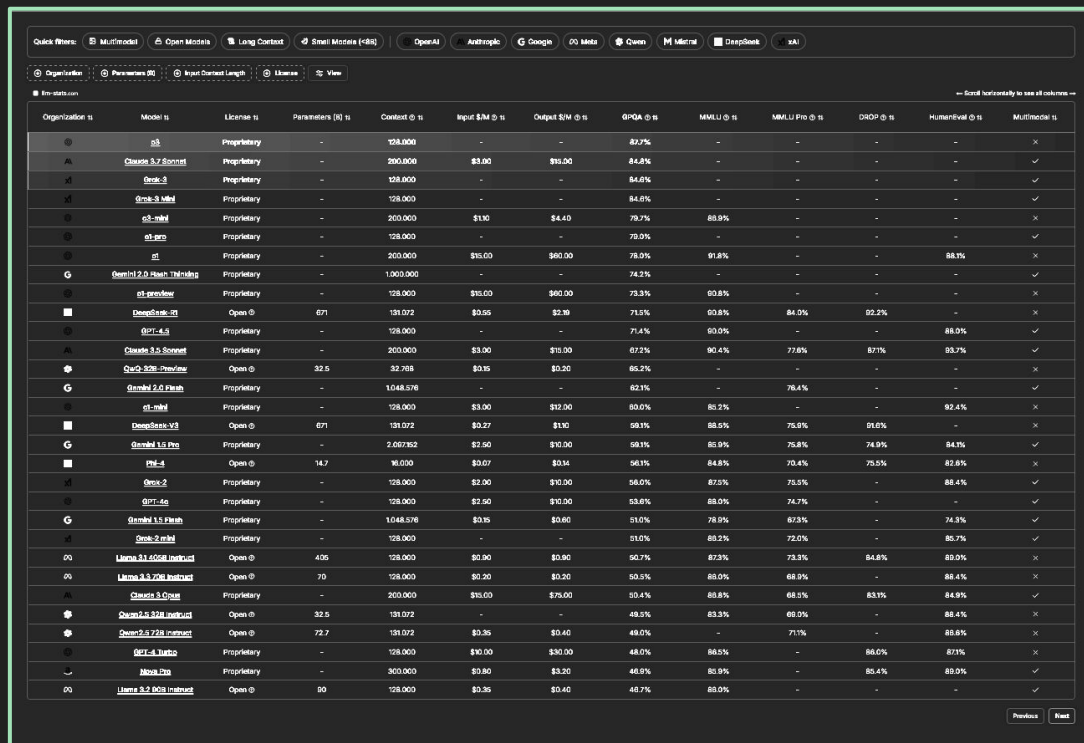


Every Model has Pro's & Con's

DROP is a specialized evaluation framework that tests LLMs' ability to extract and process discrete information from text, focusing on mathematical reasoning and information extraction. It's particularly relevant, where precision is crucial.

>> **Benchmarks** >>

(<https://llm-stats.com/>)
Info: Huggingface-Cards



Organization	Model	License	Parameters	Context	Input \$/M	Output \$/M	QoQ	MMLU	MMLU Pro	DROP	HumanEval	Multimodal		
ai	gpt-4o	Proprietary	-	128,000	-	-	82.7%	-	-	-	-	×		
AI	Claude 3.7 Sonnet	Proprietary	-	200,000	\$3.00	\$10.00	84.8%	-	-	-	-	✓		
ai	o1	Proprietary	-	128,000	-	-	84.0%	-	-	-	-	✓		
ai	o1-mini	Proprietary	-	128,000	-	-	84.0%	-	-	-	-	✓		
ai	o1-mini	Proprietary	-	200,000	\$1.00	\$4.40	79.7%	88.9%	-	-	-	×		
ai	o1-pro	Proprietary	-	128,000	-	-	79.0%	-	-	-	-	✓		
ai	o1	Proprietary	-	200,000	\$15.00	\$60.00	78.0%	91.8%	-	-	-	×		
G	Gemini 2.0 Flash Thinking	Proprietary	-	1,000,000	-	-	74.2%	-	-	-	-	✓		
ai	o1-preview	Proprietary	-	128,000	\$10.00	\$60.00	73.3%	93.8%	-	-	-	×		
■	DeepSeek-V3	Open ©	671	131,072	\$0.55	\$2.30	71.5%	92.8%	84.0%	92.2%	-	×		
ai	gpt-4.5	Proprietary	-	128,000	-	-	71.4%	93.0%	-	-	-	88.0%	✓	
AI	Claude 3.5 Sonnet	Proprietary	-	200,000	\$3.00	\$10.00	67.2%	90.4%	77.8%	87.1%	-	83.7%	✓	
+	Qwen 3.1 72B Instruct	Open ©	32.5	32,768	\$0.15	\$0.20	65.2%	-	-	-	-	-	×	
G	Gemini 2.0 Flash	Proprietary	-	1,048,576	-	-	62.1%	-	-	-	-	-	✓	
ai	o1-mini	Proprietary	-	128,000	\$3.00	\$12.00	60.0%	85.2%	-	-	-	92.4%	×	
■	DeepSeek-V3	Open ©	671	131,072	\$0.27	\$1.00	59.1%	88.5%	75.9%	91.6%	-	-	×	
G	Gemini 1.5 Pro	Proprietary	-	2,067,952	\$2.50	\$10.00	59.1%	85.9%	75.8%	74.9%	-	84.1%	✓	
■	Phi-4	Open ©	14.7	96,000	\$0.07	\$0.14	56.1%	84.8%	70.4%	75.5%	-	87.8%	×	
ai	o1	Proprietary	-	128,000	\$2.00	\$10.00	56.0%	87.0%	75.5%	-	-	88.4%	✓	
ai	gpt-4o	Proprietary	-	128,000	\$2.50	\$10.00	53.6%	88.0%	74.7%	-	-	-	✓	
G	Gemini 1.5 Flash	Proprietary	-	1,048,576	\$0.15	\$0.60	51.0%	78.9%	87.3%	-	-	74.3%	✓	
ai	o1	Proprietary	-	128,000	-	-	51.0%	80.2%	72.0%	-	-	85.7%	✓	
00	Llama 3.1 608B Instruct	Open ©	405	128,000	\$0.80	\$0.80	50.7%	87.3%	73.3%	84.8%	-	89.0%	×	
00	Llama 3.1 70B Instruct	Open ©	70	128,000	\$0.20	\$0.20	50.5%	88.0%	68.9%	-	-	88.4%	×	
AI	Claude 3.5 Opus	Proprietary	-	200,000	\$10.00	\$78.00	50.4%	88.5%	68.5%	83.1%	-	84.9%	✓	
+	Qwen 3.1 72B Instruct	Open ©	32.5	131,072	-	-	48.5%	83.3%	68.0%	-	-	88.4%	×	
+	Qwen 3.1 72B Instruct	Open ©	72.7	131,072	\$0.35	\$0.60	48.0%	-	71.1%	-	-	88.8%	×	
ai	o1-mini	Proprietary	-	128,000	\$10.00	\$30.00	48.0%	86.5%	-	-	-	86.0%	×	
ai	o1	Proprietary	-	300,000	\$0.80	\$3.20	46.0%	80.9%	-	-	-	85.4%	89.0%	✓
00	Llama 3.1 80B Instruct	Open ©	80	128,000	\$0.35	\$0.40	45.7%	88.0%	-	-	-	-	✓	

Benchmarks

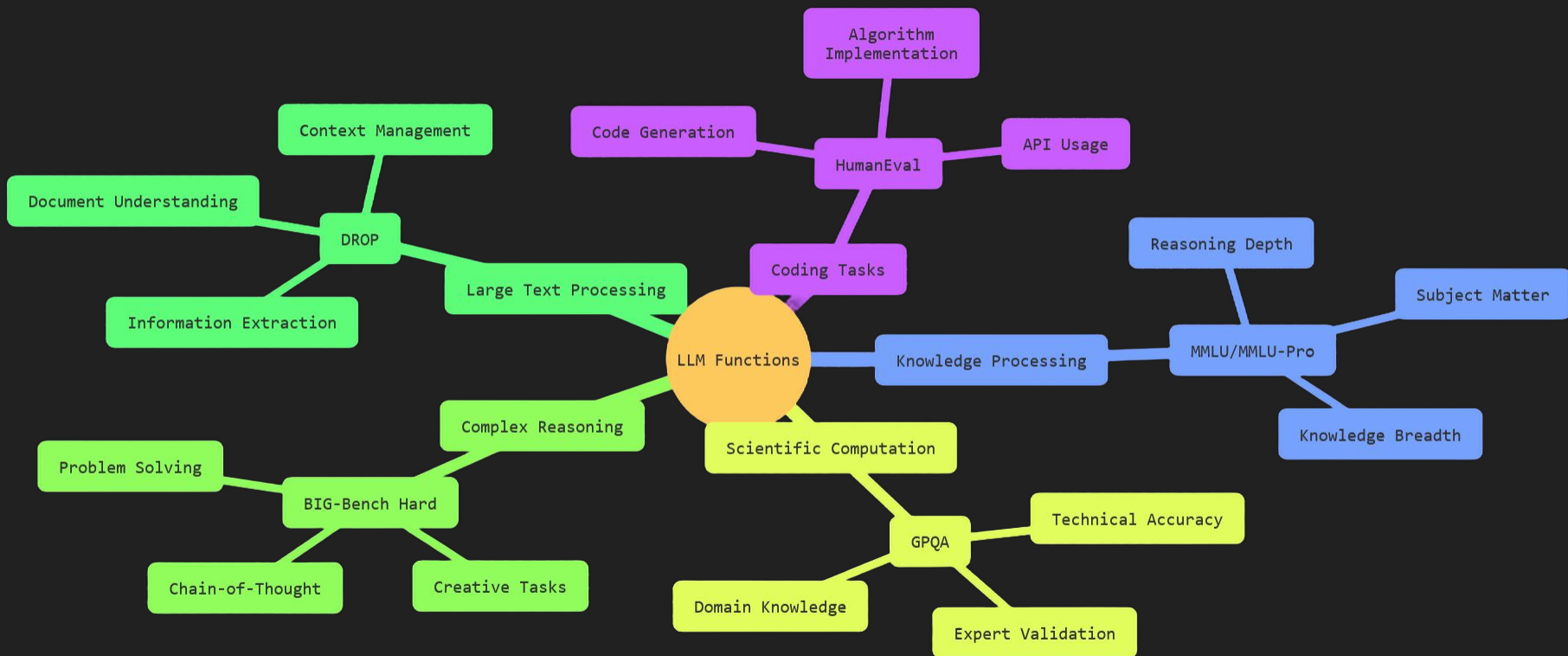
MMLU/MMLU-Pro is a comprehensive evaluation framework that assesses language models' knowledge across 57 subjects, from elementary to advanced levels, using multiple-choice questions that require zero-shot or few-shot reasoning abilities, with answers scored based on exact matches, and its enhanced version, MMLU-Pro, which features more complex reasoning-based questions, increased options, and an expanded dataset, resulting in significantly lower performance scores and a higher level of difficulty that better challenges top language models.

HumanEval is a benchmark evaluating programming capabilities through 164 practical coding tasks, focusing on functional correctness rather than code similarity.

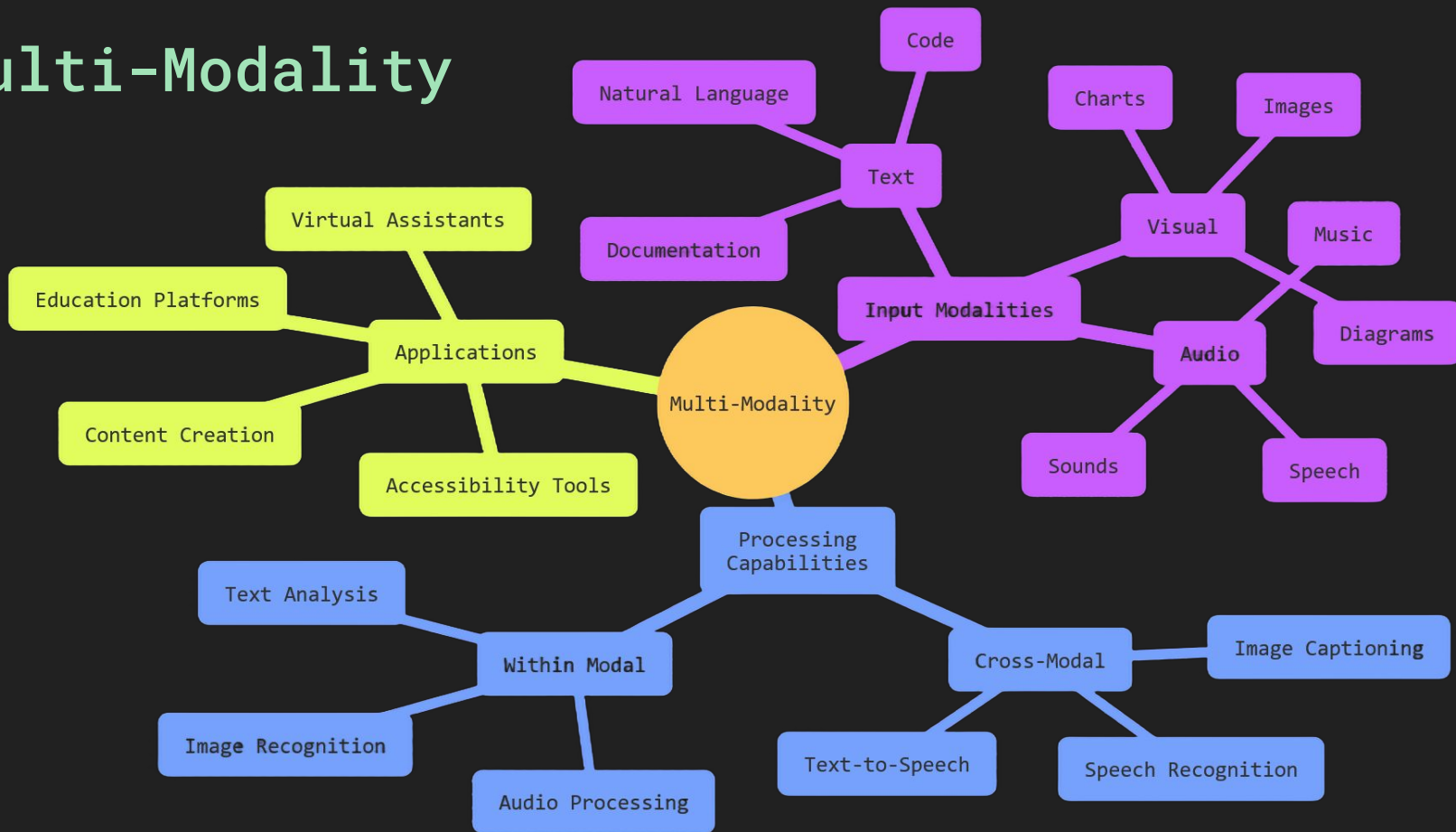
GPQA is a challenging evaluation framework that consists of graduate-level questions in scientific domains such as physics, chemistry, and biology, crafted by PhD holders to be difficult even with internet access, and designed to test understanding of complex concepts, technical jargon, and nuanced topics.

BIG-Bench Hard (BBH) is a subset of 23 challenging tasks from the BIG-Bench suite, covering diverse domains such as linguistics, math, and social bias detection, which are particularly effective in evaluating language models when combined with Chain-of-Thought prompting, and have been shown to demonstrate significant performance improvements when using this reasoning approach.

Function > Benchmark > Application



Multi-Modality



Quantization

High Precision (32-bit)

- Maximum accuracy and performance
- Largest memory footprint
- Highest computational requirements
- Typically used for training and high-stakes applications

Medium Precision (16-bit)

- Good balance of accuracy and size
- Common in production environments
- Significant memory reduction
- Minimal performance impact

Low Precision (8-bit or less)

- Smallest memory footprint
- Fastest inference speed
- Noticeable accuracy reduction
- Ideal for edge devices and mobile applications

Quantization Effects (<8bq)

- Reduced context accuracy
- More pronounced quantization effects
- Optimal for memory-constrained devices
- May require context refresh strategies

Efficient Prompting & UX

Outline the context:

Describe your current situation and where you want to go. Ask yourself: "What is it about?" and "What text would spark my own interest?". Add qualitative information (tone, medium, ...). If possible use the system prompt and the initial prompt in tandem.

Topic Introduction:

Introduce the topic and all the information you have. Keywords and unfinished questions work best. Tipp: Add "ask me questions to fulfill this task to the best of your abilities!" (Superlatives signal importance)

Correct Errors:

Tell the AI exactly what it did wrong and use the Recreate-Button. (1-2 rounds max!)

Rewrite the text:

Let the AI rewrite the text for the actual audience. Go into detail describing it. Let the AI know where you want to go by adding the most important keywords you wish to convey. If there are issues with the text, use the Recreate-Button. If you like an older version better, use the Arrow-Buttons. Tipp: Append "This is not an ad!"

>> Example: huggingface.co/chat

- Sources
- + Add source
 - Select all sources
 - CV-HidéoSnes.pdf
 - Deep Histories-HidéoSnes-2.pdf
 - Document

Chat



3 sources

Hidéo SNES is presented as a multifaceted individual with a background in digital design and project development for major companies and social sector initiatives. This person possesses a unique combination of artistic and technical skills, including art direction and DevOps knowledge, which allows for effective project management. **The entity "uWu," also identified as Hidéo SNES, is described as a constantly shifting, non-singular artistic force that challenges traditional notions of artistic production. This artistic practice encompasses interactive digital media, installations, performance, and the use of artificial intelligence.** The document thus introduces a professional with diverse expertise and a conceptually driven artistic persona.

- Save to note
- Add note
- Audio Overview
- Briefing doc

Studio

Audio Overview

Generating conversation...
This may take a few minutes. No need to stick around!

- Notes
- + Add note
 - Study guide
 - Briefing doc
 - FAQ
 - Timeline
 - Hidéo SNES: Generative AI Art and Projects**
Detailed Timeline of Hidéo SNES's Main Events: Pre-2017: Studies in Applied Arts: Hidéo SNES completes studies in Applied Arts in Vienna and Oslo. Work in Digit...
 - Hidéo SNES: Deep Histories and AI Art**
FAQ: Exploring the Work of Hidéo SNES and Deep Histories 1. Who is Hidéo SNES and what is their background? Hidéo SNES is an artist with a background in...

https://notebooklm.google.com

3 sources

How did Hidéo SNES's diverse professional background shape their unique approach to creative and tech...

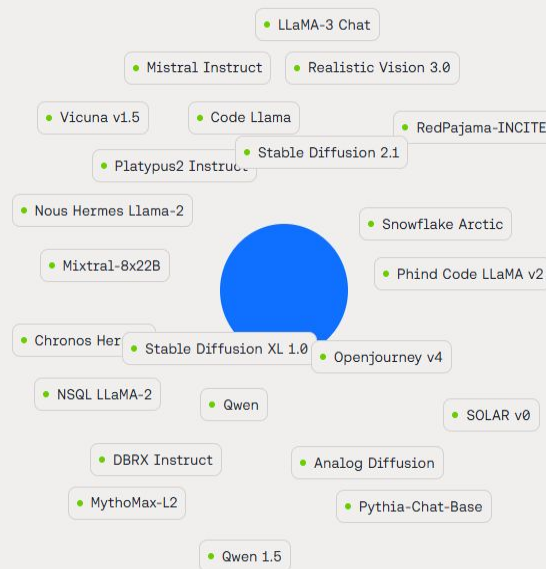
together .products

Build and run generative AI applications with accelerated performance, maximum accuracy, and lowest cost at production scale.

Start building now

Docs

Products ▾ For Business ▾ For Developers ▾ Pricing ▾ Research Company ▾ Docs Contact Get Started ↗



<https://together.ai>

Welcome to GPT4All

The privacy-first LLM chat application



Start Chatting

Chat with any LLM



LocalDocs

Chat with your local files



Find Models

Explore and download models

Latest News

GPT4All v3.10.0 was released on February 24th. Changes include:

Remote Models:

- The Add Model page now has a dedicated tab for remote model providers.
- Groq, OpenAI, and Mistral remote models are now easier to configure.

• **CUDA Compatibility:** GPUs with CUDA compute capability 5.0 such as the GTX 750 are now supported by the CUDA backend.

• **New Model:** The non-MoE Granite model is now supported.

Translation Updates:

- The Italian translation has been updated.
- The Simplified Chinese translation has been significantly improved.

• **Better Chat Templates:** The default chat templates for OLMoE 7B 0924/0125 and Granite 3.1 3B/8B have been improved.

• **Whitespace Fixes:** DeepSeek-R1-based models now have better whitespace behavior in their output.

• **Crash Fixes:** Several issues that could potentially cause GPT4All to crash have been fixed.



<https://gpt4all.io>

<https://freeconvert.com/pdf-to-txt>

Attach .xlsx & .xml

Connect Cloud Storage (Drive, One, etc.)

