Generative AI suite powered by state-of-the-art models and providing advanced AI/AGI functions. It features AI personas, AGI functions, multi-model chats, text-to-image, voice, response streaming, code highlighting and execution, PDF import, presets for developers, much more. Deploy on-prem or in the cloud

About:

Generative AI suite powered by state-of-the-art models and providing advanced AI/AGI functions. It features AI personas, AGI functions, multi-model chats, text-to-image, voice, response streaming, code highlighting and execution, PDF import, presets for developers, much more. Deploy on-prem or in the cloud.  
  
Support for the newly released Gemini Pro 1.5 models  
Support for the new OpenAI 2024-04-09 Turbo models  
Beam: the multi-model AI chat. find better answers, faster - a game-changer for brainstorming, decision-making, and creativity.  
Managed Deployments Auto-Configuration: simplify the UI models setup with backend-set models.  
Message Starring : star important messages within chats, to attach them later  
Enhanced the default Persona  
Anthropic Claude-3 model family support.  
New Perplexity and Groq integration (thanks @Penagwin).  
LocalAI deep integration, including support for model galleries  
Mistral Large and Google Gemini 1.5 support  
Performance optimizations: runs much faster, saves lots of power, reduces memory usage  
Enhanced UX with auto-sizing charts, refined search and folder functionalities, perfected scaling  
And with more UI improvements, documentation, bug fixes (20 tickets), and developer enhancements

AI models supported vendors  
Opensource Servers LocalAI (multimodal) · Ollama · Oobabooga  
Local Servers LM Studio  
Multimodal services Azure · Google Gemini · OpenAI  
Language services Anthropic · Groq · Mistral · OpenRouter · Perplexity · Together AI  
Image services Prodia (SDXL)  
Speech services ElevenLabs (Voice synthesis / cloning)

More integrations  
Web Browse Browserless · Puppeteer-based  
Web Search Google CSE  
Code Editors CodePen · StackBlitz · JSFiddle  
Sharing [Paste.gg](https://paste.gg/) (Paste chats)  
Tracking Helicone (LLM Observability)

backend:  
  
A Gradio web UI for Large Language Models. Supports transformers, GPTQ, AWQ, EXL2, llama.cpp (GGUF), Llama models.  
  
Features  
3 interface modes: default (two columns), notebook, and chat.  
Multiple model backends: Transformers, llama.cpp (through llama-cpp-python), ExLlamaV2, AutoGPTQ, AutoAWQ, GPTQ-for-LLaMa, QuIP#.  
Dropdown menu for quickly switching between different models.  
Large number of extensions (built-in and user-contributed), including Coqui TTS for realistic voice outputs, Whisper STT for voice inputs, translation, multimodal pipelines, vector databases, Stable Diffusion integration, and a lot more. See the wiki and the extensions directory for details.  
Chat with custom characters.  
Precise chat templates for instruction-following models, including Llama-2-chat, Alpaca, Vicuna, Mistral.  
LoRA: train new LoRAs with your own data, load/unload LoRAs on the fly for generation.  
Transformers library integration: load models in 4-bit or 8-bit precision through bitsandbytes, use llama.cpp with transformers samplers (llamacpp\_HF loader), CPU inference in 32-bit precision using PyTorch.  
OpenAI-compatible API server with Chat and Completions endpoints

update:  
  
Image Analysis: Got an image? We've got answers! Upload or paste any image and ask away.  
Built-In Code Interpreter: Analyze graphs and create files using our integrated Python-GPT4 interpreter.  
Web Searches: The BingChat API support makes web searches a breeze.  
Flowcharts and MindMap Generation: Generate perfect flowcharts and mindmaps on any topic, as easily as wishing upon a star.  
Large Text Summary: Large chunks of text? No problem! We break them down into bite-sized summaries.  
Free: No need to pay for gpt-4-usage. It's like a limitless universe of possibilities!