Monthly Cost Analysis of Paid LLM Models for a Personal AI Assistant

Executive Summary: This report analyzes the potential monthly costs associated with utilizing prominent paid Large Language Models (LLMs) such as those offered by OpenAI, Google (Gemini), and Anthropic (Claude) to develop a personal AI assistant. The analysis estimates monthly expenses based on typical usage patterns for such an assistant, identifies the key factors contributing to these costs, and explores the landscape of open-source LLM alternatives. The findings suggest that the monthly cost can vary significantly depending on the chosen model and the intensity of usage, with lower-cost models potentially offering a more economical solution for personal use. Open-source LLMs present a viable alternative, particularly for users with technical expertise, although they come with their own set of challenges and cost considerations.

Understanding Paid Large Language Model Pricing Structures:

OpenAI provides a diverse range of LLM models with varying capabilities and price points. For building a personal AI assistant, the gpt-4o, gpt-4o-mini, and gpt-3.5-turbo models are particularly relevant [1]. The gpt-4o model, designed for high intelligence and complex tasks, costs $2.50 per 1 million input tokens and $10.00 per 1 million output tokens, with a context window of 128K tokens [1]. A more affordable option is gpt-4o-mini, intended for fast, everyday tasks, priced at $0.15 per 1 million input tokens and $0.60 per 1 million output tokens, also with a 128K context window [1]. For general applications, gpt-3.5-turbo costs $0.50 per 1 million input tokens and $1.50 per 1 million output tokens, with a 16K context window [1]. OpenAI also offers cached input pricing, which is generally half the price of regular input tokens for certain models [1].

The OpenAI API offers different tiers, including the Chat Completions API, Assistants API, and Responses API, each potentially having different cost implications, especially concerning the use of built-in tools [1]. For instance, using tools within the Assistants API, such as Code Interpreter, incurs a cost of $0.03 per session, and File Search Storage costs $0.10 per GB per day after an initial free GB [1]. It is important to distinguish these API costs from the subscription plans for ChatGPT itself, which offer direct access to models through a web interface or application and are not the primary consideration for building a custom assistant using the API [4]. The significant variation in price points across OpenAI's models, with newer models like gpt-4o aiming for a better balance of performance and cost compared to older generations, underscores that the choice of model will be a critical factor in determining the overall

expense [1]. Furthermore, the specific API chosen for implementation, particularly if it involves the use of additional tools, can influence the final cost [1].

Table 1: Comparative Pricing of Prominent OpenAI Models (per 1M tokens)

Model	Input Price	Output Price	Cached Input Price	Context Window
gpt-4o	$2.50	$10.00	$1.25	128K
gpt-4o-mini	$0.15	$0.60	$0.075	128K
gpt-3.5-turbo	$0.50	$1.50	-	16K

Google's Gemini family of models also presents a tiered pricing structure [8]. Key models for a personal AI assistant include Gemini 1.5 Pro and Gemini 1.5 Flash, with the newer Gemini 2.0 Flash and Gemini 2.0 Flash-Lite also available [9]. For Gemini 1.5 Pro, the input price is $1.25 per 1 million tokens for prompts up to 128K tokens and $2.50 for longer prompts. Output costs are $5.00 per 1 million tokens for shorter prompts and $10.00 for longer ones [9]. Gemini 1.5 Flash is priced at $0.075 per 1 million input tokens (up to 128K) and $0.15 for longer prompts, with output at $0.30 (<=128K) and $0.60 (>128K) per million tokens [9]. Similar to OpenAI, Google offers a free tier for experimentation with its models, which includes limitations on requests per minute, tokens per minute, and total daily requests, varying by model [8].

Pricing for Google Workspace integrations and Vertex AI, while relevant for enterprise applications, is less directly applicable to the cost of API usage for a personal assistant [8]. Grounding with Google Search, which allows the model to access and incorporate real-time information, incurs a cost of $35 per 1,000 requests [8]. The availability of significantly cheaper 'Flash' models compared to the more powerful 'Pro' models indicates a similar strategy to OpenAI, providing options based on the desired performance and budget [9]. The free tier offered by Gemini could be a valuable starting point for personal use, allowing for initial development and experimentation within the specified usage limits [8].

Table 2: Comparative Pricing of Prominent Google Gemini Models (per 1M

tokens)

Model	Input Price (<=128k)	Input Price (>128k)	Output Price (<=128k)	Output Price (>128k)	Context Window
Gemini 1.5 Pro	$1.25	$2.50	$5.00	$10.00	2M
Gemini 1.5 Flash	$0.075	$0.15	$0.30	$0.60	1M

Anthropic's Claude platform offers several subscription plans for direct access, but specific API pricing details were not readily available in the provided material [5]. However, pricing for the Claude 3.5 Haiku model is mentioned, starting at $0.80 per million input tokens and $4 per million output tokens [15]. Other models in the Claude 3.5 family, such as Sonnet and Opus, exist but their API pricing for direct use in building an application was not detailed. Anthropic does mention potential cost savings through prompt caching and the Message Batches API for Haiku [15]. While the subscription plans (Free, Pro, Team, Enterprise) offer different usage levels and features for interacting with Claude directly, these costs may not directly translate to the expenses incurred when using Claude models through an API for a custom-built assistant [5]. The pricing for Claude 3.5 Haiku appears competitive, particularly on the input side, suggesting it could be a cost-effective option if its capabilities align with the requirements of the personal AI assistant [15].

Table 3: Comparative Pricing of Anthropic Claude Models (per 1M tokens)

Model	Input Price	Output Price	Context Window
Claude 3.5 Haiku	$0.80	$4.00	200K

Estimating Token Usage for a Personal AI Assistant:

A personal AI assistant could perform a variety of tasks, including scheduling appointments, retrieving information, assisting with creative writing, drafting emails, setting reminders, and managing tasks. Each of these functionalities involves both input from the user and output generated by the AI. For scheduling, a user might provide details about a meeting, and the AI would output a confirmation or update to a calendar, potentially requiring a few hundred tokens [18]. Information retrieval could involve a user's query and the AI's response summarizing or providing the requested data, potentially ranging from a few dozen to several hundred tokens depending on the complexity [23]. Creative writing tasks, such as generating a short story or poem, could involve a prompt and a longer output, potentially consuming thousands of tokens [28]. Similarly, drafting emails might involve a user's instructions and the AI generating the email content, with token usage varying based on length [32]. Even simple interactions can accumulate tokens over time [37].

Estimating monthly token usage requires considering the frequency of these tasks. In 2017, smartphone AI assistants were used around 10 times per month on average [42]. However, with the increasing integration of AI, current usage is likely higher [43]. For a low-usage scenario, one might estimate around 30 interactions per month, with an average of 200 tokens per interaction (100 input, 100 output), totaling 6,000 tokens. A medium-usage scenario could involve 100 interactions monthly at 500 tokens per interaction (250 input, 250 output), reaching 50,000 tokens. High usage might entail 300 interactions per month at 1,000 tokens per interaction (500 input, 500 output), totaling 300,000 tokens. Agentic behavior, where the AI performs multiple internal steps or uses tools, can significantly increase these numbers [39]. Techniques like Retrieval-Augmented Generation (RAG) with efficient context management can help optimize token usage, especially for information retrieval, by providing relevant context without sending the entire conversation history with each query [26].

Table 4: Estimated Monthly Token Usage Scenarios

Usage Scenario	Monthly Interactions	Avg. Input Tokens/In	Avg. Output Tokens/In	Total Monthly Input	Total Monthly Output	Total Monthly Tokens

		teraction	teraction	Tokens	Tokens	
Low	30	100	100	3,000	3,000	6,000
Medium	100	250	250	25,000	25,000	50,000
High	300	500	500	150,000	150,000	300,000

Projected Monthly Costs for Paid LLM Usage:

Based on the estimated token usage, the projected monthly costs for different paid LLMs can be calculated. For the low-usage scenario (6,000 total tokens, 3,000 input and 3,000 output), using OpenAI's gpt-4o would cost approximately $0.0075 for input and $0.03 for output, totaling $0.0375. gpt-4o-mini would cost $0.00045 for input and $0.0018 for output, totaling $0.00225. gpt-3.5-turbo would cost $0.0015 for input and $0.0045 for output, totaling $0.006. Gemini 1.5 Pro would cost $0.00375 for input and $0.015 for output, totaling $0.01875. Gemini 1.5 Flash would cost $0.000225 for input and $0.0009 for output, totaling $0.001125. Claude 3.5 Haiku would cost $0.0024 for input and $0.012 for output, totaling $0.0144.

For the medium-usage scenario (50,000 total tokens, 25,000 input and 25,000 output), gpt-4o would cost $0.0625 for input and $0.25 for output, totaling $0.3125. gpt-4o-mini would cost $0.00375 for input and $0.015 for output, totaling $0.01875. gpt-3.5-turbo would cost $0.0125 for input and $0.0375 for output, totaling $0.05. Gemini 1.5 Pro would cost $0.03125 for input and $0.125 for output, totaling $0.15625. Gemini 1.5 Flash would cost $0.001875 for input and $0.0075 for output, totaling $0.009375. Claude 3.5 Haiku would cost $0.02 for input and $0.10 for output, totaling $0.12.

For the high-usage scenario (300,000 total tokens, 150,000 input and 150,000 output), gpt-4o would cost $0.375 for input and $1.50 for output, totaling $1.875. gpt-4o-mini would cost $0.0225 for input and $0.09 for output, totaling $0.1125. gpt-3.5-turbo would cost $0.075 for input and $0.225 for output, totaling $0.30. Gemini 1.5 Pro would cost $0.1875 for input and $0.75 for output, totaling $0.9375. Gemini 1.5 Flash would cost $0.01125 for input and $0.045 for output, totaling $0.05625. Claude 3.5 Haiku would cost $0.12 for input and $0.60 for output, totaling $0.72.

These calculations demonstrate that the monthly cost can vary significantly based on the chosen LLM model, even for the same token usage [14]. Lower-cost models like

gpt-4o-mini and Gemini 1.5 Flash offer substantially more affordable options compared to higher-performance models. Additionally, the ratio of input to output tokens in the usage pattern will influence the overall cost, as some models have different pricing for input and output tokens [39].

Table 5: Estimated Monthly Costs for Paid LLMs

Usage Scenario	OpenAI gpt-4o	OpenAI gpt-4o-mini	OpenAI gpt-3.5-turbo	Gemini 1.5 Pro	Gemini 1.5 Flash	Claude 3.5 Haiku
Low	$0.0375	$0.00225	$0.006	$0.01875	$0.001125	$0.0144
Medium	$0.3125	$0.01875	$0.05	$0.15625	$0.009375	$0.12
High	$1.875	$0.1125	$0.30	$0.9375	$0.05625	$0.72

Factors Contributing to the Cost of Paid LLMs:

Several factors contribute to the overall cost of using paid LLMs. The most direct factor is the **number of tokens processed** [39]. Since most LLM providers charge based on the volume of input and output tokens, higher usage directly translates to higher costs. The **specific model chosen** also plays a crucial role, as per-token costs can vary dramatically between different models from the same provider [14]. For example, using a more powerful model like gpt-4o is significantly more expensive per token than using a smaller, less capable model like gpt-4o-mini. The **context window size** can indirectly affect costs, as larger context windows might lead to processing more tokens, especially if the model charges differently for longer prompts [20]. Additionally, using **additional features or higher API tiers**, such as built-in tools within the Assistants API, can incur extra charges [1]. The **frequency of usage** is another obvious contributor; more frequent interactions will naturally lead to higher token consumption and costs. Finally, **prompt optimization and efficiency** are important considerations. Well-crafted and concise prompts can reduce the number of input tokens required, thereby lowering costs [23]. Techniques like iterative refinement of responses can also help in minimizing overall token usage [33]. Understanding the

tokenization process and the impact of prompt length and complexity is therefore crucial for effectively managing expenses [20].

Exploring the Capabilities of Open-Source LLMs for Agentic AI:

The landscape of open-source LLMs is vibrant and rapidly expanding, offering a compelling alternative to proprietary models [13]. Prominent projects include Meta's LLaMA family, Google's Gemma, Mistral AI's models, Falcon from the Technology Innovation Institute, BLOOM from BigScience, Alibaba's Qwen series, Microsoft's Phi family, DeepSeek, and many others. These models vary in size (number of parameters), context window length, and specific strengths, such as multilingual capabilities or coding proficiency [13]. Resources like Hugging Face Transformers serve as a central platform for accessing and utilizing a vast array of these open-source models [54].

Current open-source LLMs exhibit strong capabilities in natural language understanding, generation, and increasingly in reasoning, making them suitable for various tasks relevant to a personal assistant [51]. Many can perform text generation, language translation, and question answering effectively, with larger models demonstrating more complex reasoning abilities [52]. Furthermore, a growing number of open-source AI agent frameworks are available, providing the necessary infrastructure to build agentic AI applications [55]. These frameworks, such as Auto-GPT, BabyAGI, AgentGPT, Langchain, and Llama Index, offer functionalities like autonomous task execution, internet browsing, tool integration, and support for multi-agent systems [55]. The rapid evolution of the open-source LLM ecosystem, with frequent releases of new, more capable models featuring longer context windows, makes it an increasingly viable option [13]. For technically skilled users, these frameworks provide the tools to construct sophisticated agentic AI solutions [61].

Challenges Faced by Open-Source Initiatives in Building Free/Freemium Agentic AI:

Despite the advancements in open-source LLMs, several challenges hinder the development and maintenance of free or freemium agentic AI services. The **computational resources required for training and running** these models are substantial [47]. Training large language models demands significant investment in powerful GPUs and incurs considerable energy costs. Similarly, running inference, especially for a service intended for multiple users, necessitates ongoing computational resources. **Data acquisition, curation, and licensing** also pose significant hurdles [76]. Training LLMs requires vast amounts of high-quality, diverse

data, which can be difficult and expensive to acquire and curate. Legal and ethical considerations surrounding training data, including copyright and transparency requirements like those in the EU AI Act, add further complexity [76].

The **costs of ongoing development, maintenance, and community support** are also considerable [78]. Even open-source projects require continuous effort for development, bug fixes, security patches, and providing support to the community. Unlike commercial offerings, open-source projects often lack guaranteed professional support and rely heavily on volunteer contributions [78]. **Security, bias, and ethical implications** represent further challenges [66]. The public availability of code in open-source projects can expose security vulnerabilities. Additionally, biases present in the training data can lead to skewed or harmful outputs. The potential for misuse of freely available AI models also raises ethical concerns. The significant costs associated with training and running powerful LLMs make it economically challenging for open-source entities to offer free or freemium agentic AI services at scale [47]. Moreover, ensuring data quality, addressing biases, and mitigating security risks demand substantial resources and sustained community effort, which can be difficult to maintain consistently [66].

Potential Business Models and Sustainability Strategies for Open-Source Agentic AI:

To overcome these challenges, open-source agentic AI projects can adopt various business models and sustainability strategies. **Freemium models**, offering a basic version for free with paid tiers for advanced features or higher usage, are a common approach in the software industry [83]. This allows users with basic needs to benefit from a free version while generating revenue from those requiring more extensive capabilities. **Open-core strategies** involve providing the core LLM and basic agent functionalities as open source, while offering proprietary extensions, tools, or services for a fee [87]. **Donations and grants** from individuals, corporations, and foundations supporting open-source initiatives can provide crucial funding [89]. **Corporate sponsorships** from companies that benefit from the technology can also contribute resources [89]. Offering **paid services and support**, such as consulting, training, or custom development, can generate revenue while supporting the open-source project [88]. Building a strong and active **community-led sustainability** model, where community members contribute to development, testing, and support, is essential for long-term viability [90]. Given the costs associated with large-scale AI, a freemium or open-core model where advanced features or higher usage levels are paid for appears to be a particularly plausible strategy for open-source agentic AI [83]. Relying solely on donations might not be sufficient for sustaining large-scale projects,

necessitating a more robust business model involving services or premium features [89].

Comparing Paid LLMs and Open-Source LLMs for a Personal AI Assistant:

When comparing paid and open-source LLMs for a personal AI assistant, cost-effectiveness is a key consideration. While open-source LLMs themselves might be free of licensing fees, the **total cost of ownership** can be significant [74]. This includes the expenses of infrastructure (hardware or cloud services) required to run the models, the development time needed for setup and integration, and the ongoing effort for maintenance and updates. In contrast, paid LLMs operate on a pay-as-you-go model, with costs directly tied to usage [47]. For a technically proficient user, starting with paid LLM APIs might offer ease of use and a quicker entry point, allowing for understanding usage patterns and associated costs [64]. Subsequently, transitioning to open-source LLMs could provide greater control and potentially lower long-term costs, excluding the user's time investment in managing the solution.

Regarding **capabilities, performance, and ease of use**, top-tier paid models often hold an edge in cutting-edge performance and the availability of managed services, simplifying integration [13]. However, open-source LLMs offer significant **flexibility and customizability**, allowing users to fine-tune models for specific needs [52]. The "free" nature of open-source LLMs needs to be balanced against the infrastructure and maintenance costs, as well as the potential performance gap compared to leading paid models [47]. For a data scientist with experience in machine learning and transformers, leveraging open-source LLMs for a personal AI assistant is a viable path, offering greater control and alignment with a preference for open technologies. The decision ultimately depends on the criticality of top-tier performance, budget constraints, and the willingness to invest time in managing an open-source solution.

Conclusion:

The monthly cost of using paid LLMs for a personal AI assistant can range from under a dollar to several dollars based on the chosen model and usage intensity, as illustrated by the estimated costs across different scenarios. Key factors contributing to these costs include the number of tokens processed, the specific LLM model selected, and the frequency of interactions. While open-source LLMs and agent frameworks offer a compelling alternative, particularly for technically skilled users, they come with their own challenges related to computational resources, data management, and ongoing maintenance. For a data scientist with a preference for open-source, a pragmatic approach might involve initially utilizing paid LLM APIs to establish usage patterns and then exploring the capabilities of open-source LLMs for

a more customized and potentially cost-effective long-term solution. The choice between paid and open-source LLMs for a personal AI assistant involves a trade-off between cost, performance, ease of use, and the level of control desired by the user.

Works cited

1. Pricing - OpenAI API, accessed on March 22, 2025, https://platform.openai.com/docs/pricing?product=WM
2. Pricing | OpenAI, accessed on March 22, 2025, https://openai.com/api/pricing/
3. Azure OpenAI Service - Pricing, accessed on March 22, 2025, https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/
4. Pricing - ChatGPT - OpenAI, accessed on March 22, 2025, https://openai.com/chatgpt/pricing/
5. Claude Pricing: In-Depth Guide [2025] | Team-GPT, accessed on March 22, 2025, https://team-gpt.com/blog/claude-pricing/
6. Claude, accessed on March 22, 2025, https://claude.ai/
7. Those of you using OpenAI as your LLM, how much is it costing you each month? - Reddit, accessed on March 22, 2025, https://www.reddit.com/r/homeassistant/comments/1j9pa3u/those_of_you_using_openai_as_your_llm_how_much_is/
8. Gemini Pricing: Everything You'll Pay for Google Gemini - UC Today, accessed on March 22, 2025, https://www.uctoday.com/collaboration/gemini-pricing-everything-youll-pay-for-google-gemini/
9. Gemini Developer API Pricing | Gemini API | Google AI for Developers, accessed on March 22, 2025, https://ai.google.dev/gemini-api/docs/pricing
10. Gemini for Google Cloud pricing, accessed on March 22, 2025, https://cloud.google.com/products/gemini/pricing
11. Need help: Gemini API and their stupid pricing : r/GeminiAI - Reddit, accessed on March 22, 2025, https://www.reddit.com/r/GeminiAI/comments/1g4lz3b/need_help_gemini_api_and_their_stupid_pricing/
12. Billing | Gemini API | Google AI for Developers, accessed on March 22, 2025, https://ai.google.dev/gemini-api/docs/billing
13. Best 39 Large Language Models (LLMs) in 2025 - Exploding Topics, accessed on March 22, 2025, https://explodingtopics.com/blog/list-of-llms
14. Estimate API Cost of LLMs,GPT-4o, GPT-4o-mini, Gemini 1.5 Pro, Gemini 1.5 Flash and more - YouTube, accessed on March 22, 2025, https://www.youtube.com/watch?v=ijFyUwkVmAU
15. Claude 3.5 Haiku \ Anthropic, accessed on March 22, 2025, https://www.anthropic.com/claude/haiku
16. How much does Claude Pro cost? | Anthropic Help Center, accessed on March 22, 2025, https://support.anthropic.com/en/articles/8325610-how-much-does-claude-pro-

cost
17. What is the pricing for the Team plan? | Anthropic Help Center, accessed on March 22, 2025, https://support.anthropic.com/en/articles/9267305-what-is-the-pricing-for-the-team-plan
18. Efficient LLM Scheduling by Learning to Rank | Hao AI Lab @ UCSD, accessed on March 22, 2025, https://hao-ai-lab.github.io/blogs/vllm-ltr/
19. LLMs can Schedule - arXiv, accessed on March 22, 2025, https://arxiv.org/html/2408.06993v1
20. Solving LLM Token Limit Issues: Understanding and Approaches - RagaAI- Blog, accessed on March 22, 2025, https://raga.ai/blogs/error-reading-tokens-from-llm
21. Efficient LLM Scheduling by Learning to Rank - OpenReview, accessed on March 22, 2025, https://openreview.net/pdf?id=wlLjYl0Gi6
22. Maximizer: Hebbia's Distributed System for High-Scale LLM Request Scheduling, accessed on March 22, 2025, https://www.hebbia.com/blog/maximizer-hebbias-distributed-system-for-high-scale-llm-request-scheduling
23. How LLMs Use Tokens - Mehmet Ozkaya - Medium, accessed on March 22, 2025, https://mehmetozkaya.medium.com/how-llms-use-tokens-ec5916ee321a
24. Large Language Models for Information Retrieval: A Survey - arXiv, accessed on March 22, 2025, https://arxiv.org/html/2308.07107v3
25. Calculating LLM Token Counts: A Practical Guide - Winder.AI, accessed on March 22, 2025, https://winder.ai/calculating-token-counts-llm-context-windows-practical-guide/
26. Training an LLM to effectively use information retrieval - YouTube, accessed on March 22, 2025, https://www.youtube.com/watch?v=gu5ttnClB5g
27. How to track token usage for LLMs - LangChain, accessed on March 22, 2025, https://python.langchain.com/docs/how_to/llm_token_usage_tracking/
28. LLMs for Creative Writing API Pricing - BytePlus, accessed on March 22, 2025, https://www.byteplus.com/en/topic/381257
29. LLMs for Creative Writing: A Beginner's Guide - BytePlus, accessed on March 22, 2025, https://www.byteplus.com/en/topic/381249
30. What Is an LLM Token: Beginner-Friendly Guide for Developers - The New Stack, accessed on March 22, 2025, https://thenewstack.io/what-is-an-llm-token-beginner-friendly-guide-for-developers/
31. LLM Parameters Explained: A Practical Guide with Examples for OpenAI API in Python, accessed on March 22, 2025, https://learnprompting.org/blog/llm-parameters
32. Understanding tokens - .NET | Microsoft Learn, accessed on March 22, 2025, https://learn.microsoft.com/en-us/dotnet/ai/conceptual/understanding-tokens
33. Chain of Draft: Streamlining LLM Reasoning with Minimal Token Generation - Reddit, accessed on March 22, 2025, https://www.reddit.com/r/artificial/comments/1j04ezf/chain_of_draft_streamlining_llm_reasoning_with/

34. BEST RAG LLM for interraction with emails and files - Models - Hugging Face Forums, accessed on March 22, 2025, https://discuss.huggingface.co/t/best-rag-llm-for-interraction-with-emails-and-files/135662
35. Leveraging LLMs for Email Processing in Customer Centres - YouTube, accessed on March 22, 2025, https://www.youtube.com/watch?v=-sA7wwLK0Ic
36. Is it possible to use LLMs by manually picking the tokens?, accessed on March 22, 2025, https://ai.stackexchange.com/questions/46106/is-it-possible-to-use-llms-by-manually-picking-the-tokens
37. Spend Tokens to Make Tokens - Auro Tripathy, accessed on March 22, 2025, https://auro-227.medium.com/spend-tokens-to-make-tokens-5381302158ba
38. Guide to Free vs Paid ChatGPT Token Limits - Tactiq, accessed on March 22, 2025, https://tactiq.io/learn/free-vs-paid-chatgpt-token-limits-guide
39. Azure OpenAI: what are the real costs for prompts and responses? - ClearPeople, accessed on March 22, 2025, https://www.clearpeople.com/blog/what-are-the-real-costs-for-generating-prompts-and-responses-in-azure-openai
40. How to track token usage in ChatModels | 🦜 LangChain, accessed on March 22, 2025, https://python.langchain.com/docs/how_to/chat_token_usage_tracking/
41. Mastering Token Limits and Memory in ChatGPT and other Large Language Models | by Russell Kohn | Medium, accessed on March 22, 2025, https://medium.com/@russkohn/mastering-ai-token-limits-and-memory-ce920630349a
42. 42 Percent of US Smartphone Owners Use AI Personal Assistant Monthly - Voicebot.ai, accessed on March 22, 2025, https://voicebot.ai/2017/07/28/42-percent-us-smartphone-owners-use-ai-personal-assistant-monthly/
43. AI Statistics 2024 - AIPRM, accessed on March 22, 2025, https://www.aiprm.com/ai-statistics/
44. Assistants API token usage and pricing breakdown clarification - OpenAI Developer Forum, accessed on March 22, 2025, https://community.openai.com/t/assistants-api-token-usage-and-pricing-breakdown-clarification/508410
45. Managing Costs of Your LLM Application — Part 2 of 2 | by Chris Mann | Medium, accessed on March 22, 2025, https://productmann.medium.com/managing-costs-of-your-llm-application-part-2-of-2-896ac7cdfc33
46. OpenAI API Pricing Calculator - GPT for Work, accessed on March 22, 2025, https://gptforwork.com/tools/openai-chatgpt-api-pricing-calculator
47. Understanding the cost of Large Language Models (LLMs) - TensorOps, accessed on March 22, 2025, https://www.tensorops.ai/post/understanding-the-cost-of-large-language-models-llms
48. Complete Guide to AI Tokens: Understanding, Optimization, and Cost

Management, accessed on March 22, 2025, https://guptadeepak.com/complete-guide-to-ai-tokens-understanding-optimization-and-cost-management/
49. Tokens and tokenization in AI and LLMs - Marcus D. R. Klarqvist, accessed on March 22, 2025, https://www.mdrk.io/tokenizers-in-ai-and-llms/
50. Tokenization | Mistral AI Large Language Models, accessed on March 22, 2025, https://docs.mistral.ai/guides/tokenization/
51. 8 Top Open-Source LLMs for 2024 and Their Uses - DataCamp, accessed on March 22, 2025, https://www.datacamp.com/blog/top-open-source-llms
52. Top 10 open source LLMs for 2025 - Instaclustr, accessed on March 22, 2025, https://www.instaclustr.com/education/top-10-open-source-llms-for-2025/
53. What are Open Source Large Language Models? - IBM, accessed on March 22, 2025, https://www.ibm.com/think/topics/open-source-llms
54. Essential open source large language models to watch in 2025 - Pieces for developers, accessed on March 22, 2025, https://pieces.app/blog/open-source-llms
55. Top LLMs for AI Agents in 2025, accessed on March 22, 2025, https://tunehq.ai/blog/top-llms-for-ai-agents
56. Top 10 open source LLMs for 2024 - Instaclustr, accessed on March 22, 2025, https://www.instaclustr.com/education/top-10-open-source-llms-for-2024/
57. Best Open Source LLMs of 2024 - Klu.ai, accessed on March 22, 2025, https://klu.ai/blog/open-source-llm-models
58. The best large language models (LLMs) in 2025 - Zapier, accessed on March 22, 2025, https://zapier.com/blog/best-llm/
59. Top 12 Open Source Models on HuggingFace in 2025 - Analytics Vidhya, accessed on March 22, 2025, https://www.analyticsvidhya.com/blog/2024/12/top-open-source-models-on-hugging-face/
60. A list of open LLMs available for commercial use. - GitHub, accessed on March 22, 2025, https://github.com/eugeneyan/open-llms
61. Open Source AI Agents: Exploring Best AI Agents | Keploy Blog, accessed on March 22, 2025, https://keploy.io/blog/community/top-open-source-ai-agents
62. blog/open-source-llms-as-agents.md at main - GitHub, accessed on March 22, 2025, https://github.com/huggingface/blog/blob/main/open-source-llms-as-agents.md
63. #1-Getting Started Building Generative AI Using HuggingFace Open Source Models And Langchain - YouTube, accessed on March 22, 2025, https://www.youtube.com/watch?v=bFB4zqkcatU
64. Open-Source LLMs vs Closed: Unbiased Guide for Innovative Companies [2025], accessed on March 22, 2025, https://hatchworks.com/blog/gen-ai/open-source-vs-closed-llms-guide/
65. Open Source LLM Models: A Guide to Accessible AI Development - Medium, accessed on March 22, 2025, https://medium.com/@kanerika/open-source-llm-models-a-guide-to-accessible-ai-development-ec30892aca8f

66. Open Source LLMs: The Ultimate Guide - Lyzr AI, accessed on March 22, 2025, https://www.lyzr.ai/blog/open-source-llms-guide/
67. Introduction to Open-Source AI Agents - Botpress, accessed on March 22, 2025, https://botpress.com/blog/open-source-ai-agents
68. A curated list of awesome LLM agents frameworks. - GitHub, accessed on March 22, 2025, https://github.com/kaushikb11/awesome-llm-agents
69. Open-Source AI Agents: How to Use Them and Best Examples | by Springs | Medium, accessed on March 22, 2025, https://medium.com/@springs_apps/open-source-ai-agents-how-to-use-them-and-best-examples-e19560280df1
70. Open-source DeepResearch – Freeing our search agents - Hugging Face, accessed on March 22, 2025, https://huggingface.co/blog/open-deep-research
71. Open-Source AI Agents: How to Use Them and Best Examples - Springs, accessed on March 22, 2025, https://springsapps.com/knowledge/open-source-ai-agents-how-to-use-them-and-best-examples
72. Can an Open-Source LLM Compete with OpenAI? - Team-GPT, accessed on March 22, 2025, https://team-gpt.com/blog/open-source-llm-vs-openai/
73. What is the Cost of Training LLM Models? Key Factors Explained, accessed on March 22, 2025, https://botpenguin.com/blogs/what-is-the-cost-of-training-llm-models
74. Open-Source vs Proprietary LLMs: Cost Breakdown - Latitude.so, accessed on March 22, 2025, https://latitude.so/blog/open-source-vs-proprietary-llms-cost-breakdown/
75. What is the cost of training large language models? - CUDO Compute, accessed on March 22, 2025, https://www.cudocompute.com/blog/what-is-the-cost-of-training-large-language-models
76. EU AI Act: Open Source LLMs must disclose their training data - Modulos AI, accessed on March 22, 2025, https://www.modulos.ai/blog/eu-ai-act-open-source-llms-must-disclose-their-training-data/
77. Practical Data Protection Compliance for Open-Source LLMs - AWO Agency, accessed on March 22, 2025, https://www.awo.agency/blog/practical-dp-llms/
78. 5 Critical Limitations of Open Source LLMs: What AI Developers Need to Know - Galileo AI, accessed on March 22, 2025, https://www.galileo.ai/blog/disadvantages-open-source-llms
79. What are the limitations of open-source software? - Milvus, accessed on March 22, 2025, https://milvus.io/ai-quick-reference/what-are-the-limitations-of-opensource-software
80. Open-Source AI vs. Closed-Source AI: What's the Difference? - Multimodal, accessed on March 22, 2025, https://www.multimodal.dev/post/open-source-ai-vs-closed-source-ai
81. Open Source AI: Opportunities and Challenges - Linux Foundation, accessed on

March 22, 2025, https://www.linuxfoundation.org/blog/open-source-ai-opportunities-and-challenges
82. 8 Challenges Of Building Your Own Large Language Model - Labellerr, accessed on March 22, 2025, https://www.labellerr.com/blog/challenges-in-development-of-llms/
83. What is a freemium model in SaaS? - Milvus, accessed on March 22, 2025, https://milvus.io/ai-quick-reference/what-is-a-freemium-model-in-saas
84. Freemium Pricing: Examples, Models, and Strategies - High Alpha, accessed on March 22, 2025, https://www.highalpha.com/blog/freemium-pricing-examples-and-models
85. Freemium Models: Pros, Cons, and Best Practices for SaaS Companies | Maxio, accessed on March 22, 2025, https://www.maxio.com/blog/freemium-model
86. Understanding the Best AI Business Model for Success - Lomit Patel, accessed on March 22, 2025, https://www.lomitpatel.com/articles/ai-business-model/
87. A Framework for Freemium. With a touch of open-source | by Chase Roberts | Medium, accessed on March 22, 2025, https://chsrbrts.medium.com/a-framework-for-freemium-8f03a5195315
88. SaaS freemium business model: how does it work? - TinyMCE, accessed on March 22, 2025, https://www.tiny.cloud/blog/saas-freemium-business-model/
89. How to Make an Open Source Project Sustainable Financially? : r/opensource - Reddit, accessed on March 22, 2025, https://www.reddit.com/r/opensource/comments/1h37zp5/how_to_make_an_open_source_project_sustainable/
90. Financial Sustainability in Open Source Projects: A Guide to Thriving - DEV Community, accessed on March 22, 2025, https://dev.to/ashucommits/financial-sustainability-in-open-source-projects-a-guide-to-thriving-1mcn
91. AI for Math Fund – A brighter future for all through science, technology, and innovation, accessed on March 22, 2025, https://renaissancephilanthropy.org/initiatives/ai-for-math-fund/
92. Research Grants - Fal.ai, accessed on March 22, 2025, https://fal.ai/grants
93. 10 Grants to Support Open Source Technology Initiatives - fundsforNGOs, accessed on March 22, 2025, https://www2.fundsforngos.org/articles-listicles/10-grants-to-support-open-source-technology-initiatives/
94. Open Source Sustainability: Thriving in the Face of Challenges | by Configr Technologies, accessed on March 22, 2025, https://configr.medium.com/open-source-sustainability-thriving-in-the-face-of-challenges-a7a94cfd4061
95. Open Source Business Models Explained - YouTube, accessed on March 22, 2025, https://www.youtube.com/watch?v=t4loX5H29mk
96. Examples of Open Source Business Models - The Turing Way, accessed on March 22, 2025, https://book.the-turing-way.org/collaboration/oss-sustainability/oss-sustainability

-examples.html
97. What's Next for Open Source? Workshop Highlights and Calls to Action to Inspire Progress for Global Sustainability, accessed on March 22, 2025, https://openssf.org/blog/2024/08/08/whats-next-for-open-source-workshop-highlights-and-calls-to-action-to-inspire-progress-for-global-sustainability/
98. What is Open Source LLM? Benefits, Challenges & Considerations - Deepchecks, accessed on March 22, 2025, https://www.deepchecks.com/glossary/open-source-llm/