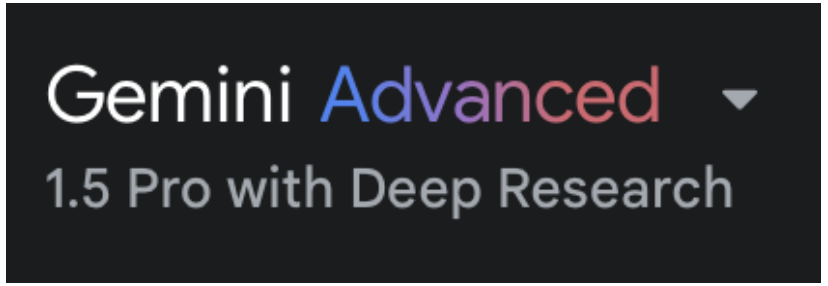
Deep Research

By: Jack Francis

AI Solutions Director -- Davidson Technologies

What is Deep Research?

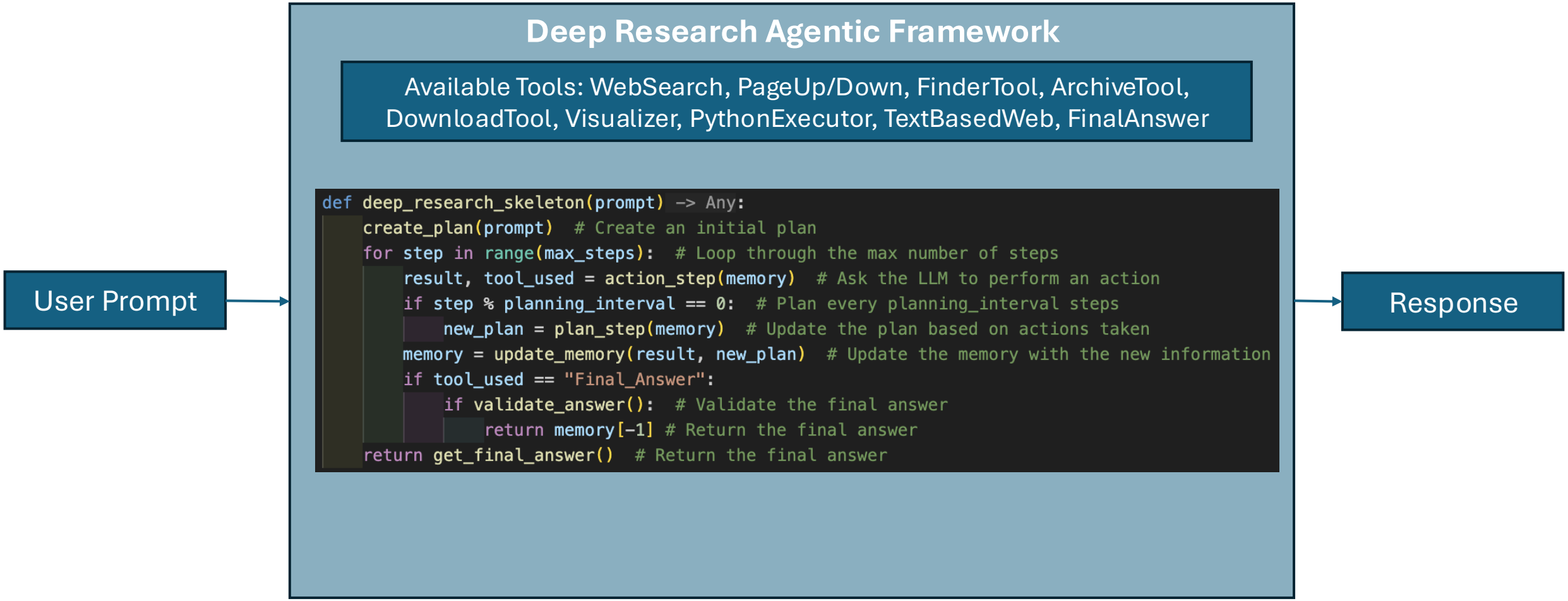


- Agentic, personal research assistant
- Analyze hundreds of sources and generate a comprehensive research report
- Timeline:
  - Google: 12/11/2024
  - OpenAI: 2/2/2025
  - Perplexity: 2/14/2025
  - Grok3: 2/17/2025
  - Open-Source recreations: Ongoing

Let’s start with an example

- Gemini.google.com

Deep Research -- How it works



Source: https://github.com/huggingface/smolagents/blob/main/examples/open_deep_research/run.py

Benchmarks

- GAIA (General AI Assistants)
  - OpenAI Deep Research: 67.36%,
  - Open-Source Deep Research: 55.15%
  - Gemini Deep Research: N/A
  - GPT-4o: 7%

Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?
Ground truth: 90

Level 2

Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.
Ground truth: +4.6

Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.
Ground truth: White; 5876

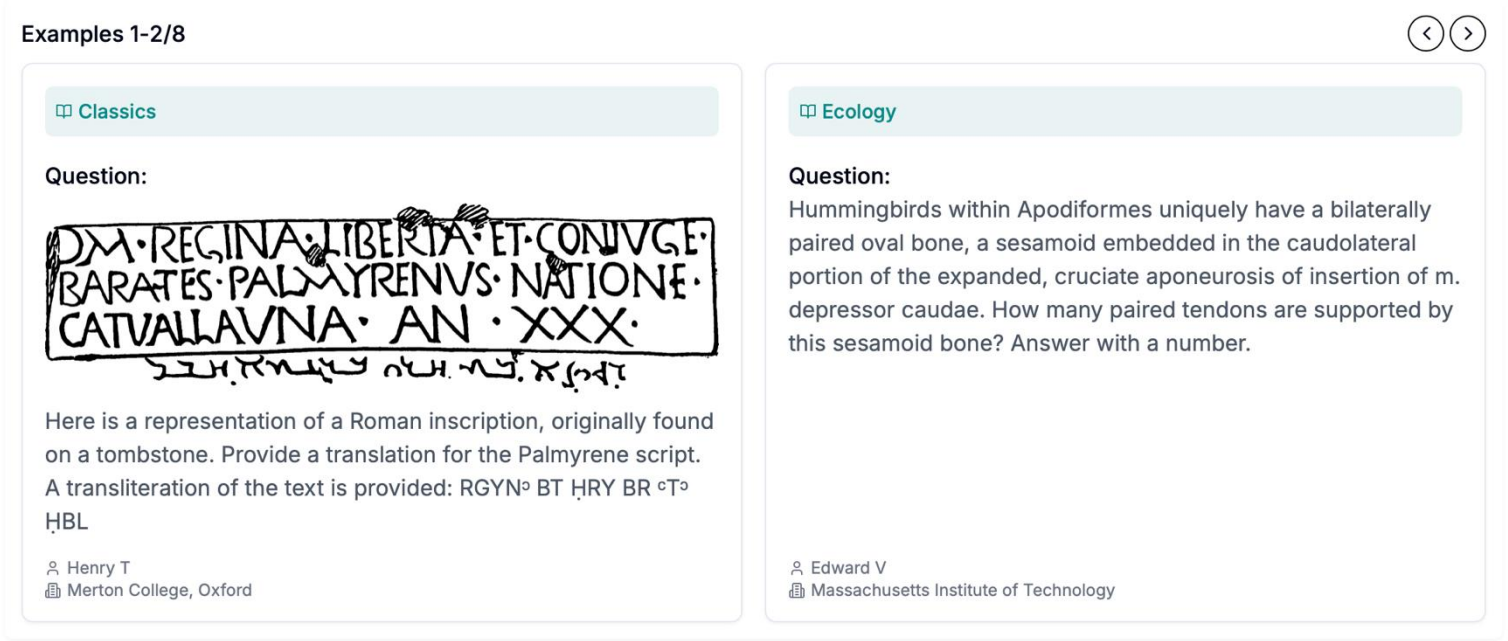
GAIA Benchmark Leaderboard: https://huggingface.co/spaces/gaia-benchmark/leaderboard

Benchmarks (Cont.)

- Humanity’s Last Exam

Judge Model: o3-mini-2025-01-31 | Last Updated: 02/11/2025

Model	Accuracy (%) ↑	Calibration Error (%) ↓
GPT-4o	3.1	92.3
Grok-2	3.9	90.8
Claude 3.5 Sonnet	4.8	88.5
Gemini Thinking	7.2	90.6
o1	8.8	92.8
DeepSeek-R1*	8.6	81.4
o3-mini (medium)*	11.1	91.5
o3-mini (high)*	14.0	92.8



Samples of the diverse and challenging questions submitted to Humanity's Last Exam.

OpenAI Deep Research: 26.6% accuracy

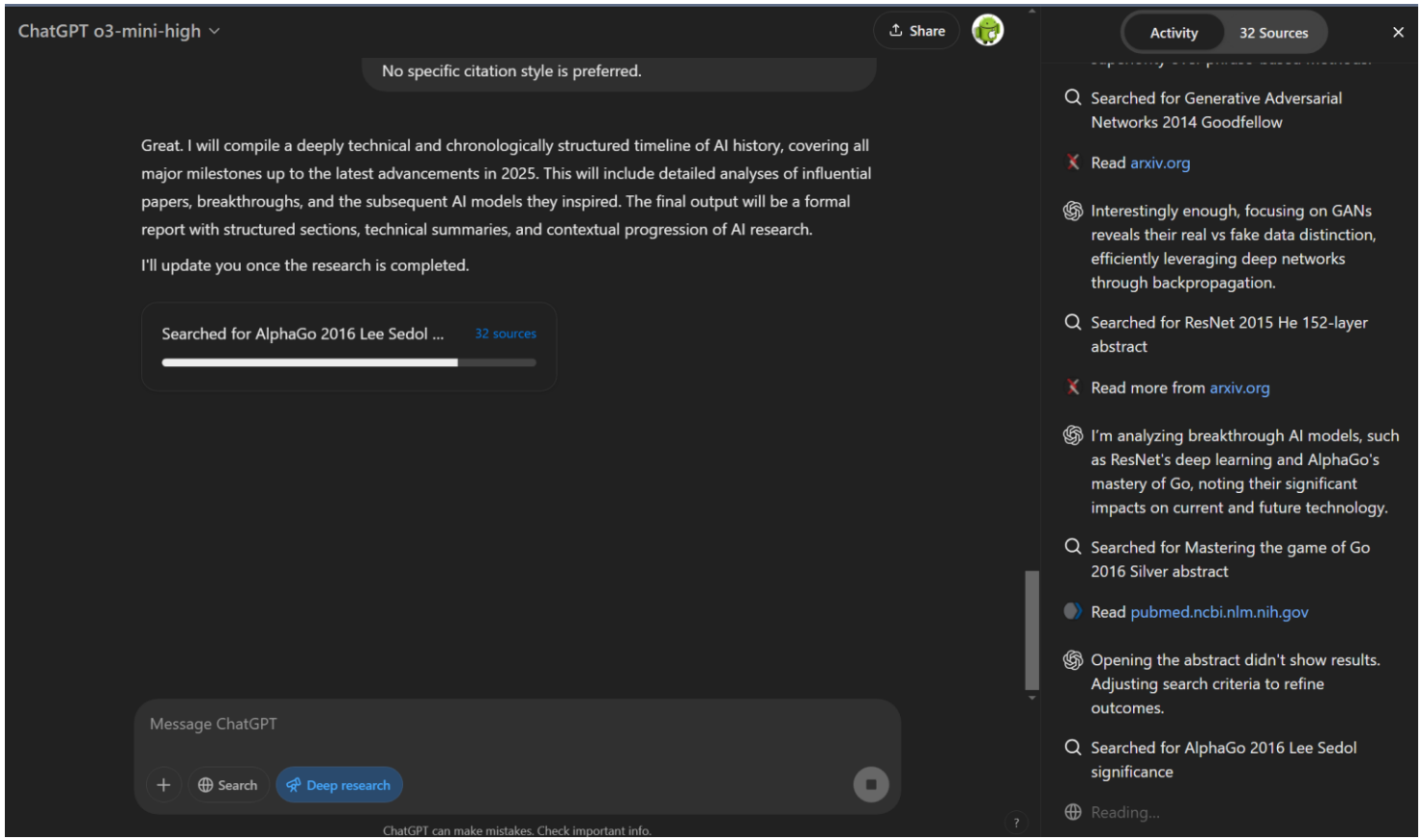
Humanity’s Last Exam: https://lastexam.ai

Google Gemini Deep Research

- Workflow: Provide a prompt, then edit a research plan
- Pricing: Included in $20/month Gemini Advanced
- Rate Limits: None, 6 concurrent research tasks
- Time to complete: Generally 3-5 minutes
- Output Format: Text in browser, can open in Google Docs. 4-6 pages of output
- Quirks/Bugs:
  - Roughly 5-10% of deep research prompts don’t finish and exit early
  - It can get confused on pages with a lot of content and assign things incorrectly
  - Currently uses the 1.5-Pro model instead of 2.0-Pro model

OpenAI Deep Research

- Workflow: Provide a prompt, OpenAI Deep Research will ask follow up questions that you answer to provide more context



OpenAI Deep Research

- Workflow: Provide a prompt, OpenAI Deep Research will ask follow up questions that you answer to provide more context
- Pricing: Included in $200/month OpenAI Pro Tier
- Rate Limits: 100 messages per month
- Time to complete: Generally 5-30 minutes (can be longer)
- Output Format: Text in browser, 8-40 pages of output.
- Quirks/Bugs:
  - Make sure to use the GPT Copy button for consistent formatting
  - Have yet to see one fail to finish, but will sometimes fail to start
  - Hallucinates more often than other models, but most of output is useful
  - Unclear on model, seems to be a custom o3

Grok Deep Search

- Workflow: Provide a prompt
- Pricing: Included in $40/month X Premium Tier
- Rate Limits: Currently unknown
- Time to complete: < 5 minutes
- Output Format: Text in browser
- Quirks/Bugs:
  - Too early to have any specifics here

Perplexity Deep Research

- Workflow: Provide a prompt
- Pricing: Free and $20/month variants
- Rate Limits: Free: 5 queries per day, Paid: 500 queries / month
- Time to complete: 2-4 minutes
- Output Format: Text in browser
- Quirks/Bugs:
  - Too early to have any specifics here

Open-Source Deep Research

- Workflow: Provide a prompt
- Pricing: Free and open-source
- Rate Limits: None
- Time to complete: 1-5 minutes
- Output Format: Text in browser, can integrate with apps locally
- Quirks/Bugs:
  - Poorer performance compared to Gemini / OpenAI
  - Quality is heavily dependent on the base model you use, will likely require significant compute to perform well (i.e., GAIA results used o1)

Comparison of Tools

- Outputs are extremely long, so manually reviewing is not feasible during this meeting.
- HSV.AI Github presentations repo has prompts and unedited outputs for each of these 10 prompts
  - https://github.com/HSV-AI/presentations/tree/master/2025

Prompts

Number	Prompt Summary	Prompt Length	OpenAI Length (words)	Gemini Length (words / websites searched)	Open-Source Length (words)
1	History of AI	604	16276	5166 / 125	397
2	Continuing Care Retirement Communities in the US	871	17511	3672 / 326	409
3	AI Job Replacement	750	11834	3062 / 95	140
4	Early Warning System Lit Review	417	12207	1410 / 12	532
5	Electronic Warfare (EW) Light Systems Market Analysis	595	20797	2577 / 203	746
6	AI Coding Agent Comparison	201	17908	2138 / 139	422
7	Neoliberalism	406	11189	1240 / 44	439
8	Physics-Informed Machine Learning vs. Kalman Filters for Autonomous Drone Systems	1034	20541	3121 / 382	560
9	Algorithmic Trading	78	12356	3017 / 59	365
10	AI in Secondary Schooling	396	11257	3089 / 75	230

Rule of Thumb: 500 words = 1 page

Comparison Findings – AI History Prompt

Objective:
Develop a highly detailed, chronologically structured timeline of the history of artificial intelligence (AI). This timeline should encompass every major milestone, influential paper, breakthrough advancement, and the subsequent public or private models derived from these contributions.

Instructions:

1. **Literature and Milestone Identification:**
   - **Scope:**
     Identify foundational and transformative AI contributions from the inception of the field (e.g., Alan Turing's work) through to modern advancements (e.g., deep learning, reinforcement learning, transformer architectures, and large-scale language models).
   - **Data Points:**
     For each milestone, include:
     - **Title and Year:** Clearly state the title of the paper, research contribution, or technological breakthrough along with the publication or release year.
     - **Authors/Institutions:** List key authors, institutions, or research groups involved.
     - **Abstract/Key Contributions:** Summarize the central ideas, innovations, and methodologies introduced.
2. **Chronological Structuring:**
   - **Timeline Format:**
     Organize the research milestones in a chronological timeline. Each entry should clearly indicate the temporal context (decade, year) and the progression of ideas over time.
   - **Contextual Linkage:**
     For each entry, include a brief explanation of how it built upon previous work or set the stage for future advancements.
3. **Technical Analysis and Impact Evaluation:**
   - **Deep Dive Analysis:**
     For every influential paper or breakthrough:
     - **Technical Contributions:** Describe the core technical innovations (algorithms, architectures, theories, etc.).
     - **Impact Assessment:** Explain its significance in advancing the field of AI, including any shifts in research paradigms, methodologies, or application areas.
     - **Subsequent Developments:** Detail how this work influenced later research, including direct citations or conceptual lineage in future models.
4. **Tracing Model Lineages:**
   - **Model References:**
     Identify and document each public or private AI model that was developed based on, or significantly influenced by, the influential papers or advancements. For each model, include:
     - **Model Name and Release Date:** Provide the name, version (if applicable), and year of release.
     - **Development Context:** Explain how the model was derived from or inspired by the original research.
     - **Public/Private Status:** Clarify whether the model is publicly accessible (open-source or published) or developed privately (commercial/industry-specific).
     - **Technical Summary:** Offer a concise description of the model's architecture, key features, and performance benchmarks if available.
5. **Comprehensive Coverage and Cross-Referencing:**
   - **Breadth of Coverage:**
     Ensure the timeline covers:
     - Early theoretical foundations (e.g., Turing's "Computing Machinery and Intelligence").
     - Early neural network research and the perceptron.
     - The rise and fall of symbolic AI.
     - The emergence of machine learning paradigms and statistical methods.
     - The advent of deep learning, including convolutional and recurrent neural networks.
     - The development and impact of transformer models and large-scale language models (e.g., GPT series, BERT).
   - **Cross-Reference System:**
     Include detailed references (DOIs, URLs, or archive links) for all mentioned papers, articles, and technical reports to facilitate further investigation.
6. **Contextual and Socio-Technical Commentary:**
   - **Broader Impact:**
     For each milestone, include commentary on the socio-technical implications, controversies, and shifts in the AI research landscape. Discuss how each advancement influenced not just technical domains but also ethical, economic, and regulatory aspects of AI.
   - **Evolution of Ideas:**
     Highlight how incremental improvements and paradigm shifts contributed to the evolution of AI, noting recurring themes or challenges that re-emerged over time.
7. **Output Format:**
   - **Structured Document:**
     Deliver the final output as a well-organized document.
   - **Clarity and Depth:**
     Ensure the narrative is both comprehensive and deeply analytical, avoiding superficial summaries.

Final Deliverable:
A meticulously detailed, chronologically ordered document that traces the complete history of AI—from its theoretical inception to its modern manifestations—linking influential papers to their resulting technological models, and providing extensive technical, contextual, and socio-technical insights.

High Level Summary – AI History

Tool	Response Length (words)	Number of AI Topics in Timeline	Number of Sources cited in final report	Prompt Adherence	Interesting Findings
OpenAI	16276 (~32 pages)	48	27	A+. Consistently followed the provided format for all topics. For general trends, specific papers/projects were provided along with overall impact.	In-text references point to specific paragraph used. When arxiv is cited, it appears to only reference the abstract
Gemini	5166 (~10 pages)	35	41	B. Followed the format for some topics but regularly missed tech contributions / impact / lineage. Explanations were 1-2 sentences for each subtopic.	Did not cite arxiv.
Grok	1794 (~3 pages) / 9916 (~ 20 pages)	15 / 43	6 / None*	B. Followed the format better than Gemini but provided fewer relevant results. 1/3 of results were 2017 or newer. “deeper response” had more topics, but very bullet and list heavy, information wasn’t synthesized very well.	Gave info on where paper was published, but not direct link. More focus on recent advancements. Each section was ~2000 words
Open-Source	397 (~0.5 pages)	12	N/A	D. Each topic was a “headline”. Lot of missing steps in the timeline for AI History	

2020 (GPT-3) – OpenAI's **GPT-3** (175 billion parameters) pushed scaling to the extreme and achieved surprising new abilities. Its landmark paper *"Language Models are Few-Shot Learners"* demonstrated that GPT-3 can achieve strong results on many NLP tasks *without fine-tuning*, simply by being given a few examples in the prompt (few-shot learning). For instance, given two examples of English to French translation, GPT-3 could translate a new sentence moderately well – essentially performing like a system trained for that task. GPT-3 also showed uncanny abilities in tasks like common-sense reasoning, arithmetic, and code generation, albeit not perfectly reliable. This was a qualitative leap: it suggested that *general linguistic intelligence* might be emerging from scaling. GPT-3's capabilities, accessible via an API, sparked enormous interest in building applications (e.g., copywriting, coding assistants, chatbots, etc.) on top of it, inaugurating the trend of large language model services. It also raised intense discussion on ethics (bias in large models, potential for misinformation, etc.). On the research front, GPT-3 provided evidence for **meta-learning** in large models – it effectively learned how to learn from prompts. This influenced subsequent work in prompt design and **in-context learning**.

Impact: GPT-2 and GPT-3 together marked the transition of language models from NLP tools to potential **general AI**components. They blurred distinctions between training and usage – the model itself contains so much knowledge that with the right prompt it can perform tasks it was never explicitly trained on. This has huge practical implications: solving tasks with little to no labeled data by leveraging a pre-trained model's knowledge. It also influences AI research directions: more focus on *scaling laws* (predictable improvements by scaling model/data) and *emergent behaviors*(unexpected abilities at certain scales). Moreover, GPT-3's success (and its API commercialization) spurred competitive efforts: e.g., EleutherAI's **GPT-J** and **GPT-Neo** open-source models, Microsoft/NVIDIA's **Megatron-Turing NLG**(530B), Google's **PaLM** (540B, 2022) – all Decoder transformers in the GPT style. These models further pushed performance and enabled experimentation at scale outside OpenAI. Another major impact has been on conceptions of "understanding" – while these models generate impressively, debates ensue whether they truly understand or just statistically emulate. Regardless, they've forced a rethinking of what tasks truly test understanding vs. recall.

Model Lineage: GPT-2/3 lineage is straightforward from **GPT-1 (2018)** which itself was a Transformer decoder adaptation of the OpenAI transformer-based LM that followed the original BERT and Transformer works. But GPT-3 was a departure in showing **few-shot prompting** as a competitive alternative to fine-tuning, reintroducing a way to use models without altering weights (a bit reminiscent of how humans learn new tasks without rewiring our brains, just by being shown examples). This has sparked new lines of research in **prompt engineering, prompt tuning (soft prompts)**, and even using language models as interactive agents (with systems like **Chain-of-Thought prompting** in 2022 giving better multi-step reasoning by prompting the model to explain its reasoning step by step). Essentially, GPT-3 demonstrated that with great scale comes a form of generality. This directly foreshadows the development of AI assistants like **ChatGPT (2022)**, which fine-tune GPT-3.5 on dialogue and human feedback to great effect, and **GPT-4 (2023)** which is even more capable and multi-modal. Thus, GPT-3's family of models are considered a step toward **AGI-like** systems by some, or at least extremely versatile AI. In conclusion, GPT-2 and GPT-3 showed that *"more is different"* in AI: enough scale can qualitatively change a model's behavior, enabling it to perform tasks in ways we hadn't seen before

OpenAI Deep Research

2020s: Large Language Models and Beyond

The 2020s have seen the rise of large language models (LLMs) with unprecedented scale and capabilities. These models, trained on massive datasets, have demonstrated remarkable abilities in generating human-like text, translating languages, and performing complex language-related tasks.

2020s: Large Language Models (LLMs) Emerge

- **Key Models:** GPT-3, BERT, LaMDA, PaLM 2
- **Capabilities:** Human-like text generation, language translation, question answering, code generation
- **Impact:** LLMs have pushed the boundaries of natural language processing and have led to the development of new applications in various domains, including chatbots, content creation, and code generation[38].

2022: ChatGPT

- **Authors/Institutions:** OpenAI
- **Abstract/Key Contributions:** OpenAI released ChatGPT, a conversational AI model based on the GPT-3.5 architecture. ChatGPT gained widespread attention for its ability to engage in human-like conversations, generate creative text formats, and answer questions in a comprehensive and informative way[40].

2023: GPT-4

- **Authors/Institutions:** OpenAI
- **Abstract/Key Contributions:** GPT-4, a more advanced version of GPT-3, was released, further improving the model's capabilities in language understanding and generation. GPT-4 demonstrated even greater fluency, coherence, and creativity in its text outputs, pushing the boundaries of what's possible with LLMs[38].

Model Name	Type	Key Contributions	Applications
GPT-3	Autoregressive Language Model	Human-like text generation, few-shot learning	Chatbots, content creation, code generation
BERT	Bidirectional Encoder Representations from Transformers	Contextualized word embeddings, improved language understanding	Search engines, sentiment analysis, question answering
LaMDA	Language Model for Dialogue Applications	Conversational AI, natural and engaging dialogue	Chatbots, virtual assistants
PaLM 2	Pathways Language Model 2	Improved reasoning and code generation abilities	Code generation, problem-solving, knowledge retrieval

Gemini Deep Research

2020 – "Language Models are Few-Shot Learners" (GPT-3)

- **Authors**: Tom B. Brown et al.
- **Institution**: OpenAI
- **Abstract/Key Contributions**: Introduced GPT-3, a 175B-parameter autoregressive language model excelling in few-shot learning without fine-tuning.
- **Technical Contributions**: Scaled Transformers with massive data (45TB), leveraging in-context learning.
- **Impact Assessment**: Demonstrated the power of scale, shifting AI toward general-purpose models.
- **Subsequent Developments**: Led to ChatGPT (2022), GPT-4 (2023).
- **Model Lineage**:
  - **Model**: GPT-3 (2020)
  - **Development Context**: Trained on Common Crawl, books, Wikipedia.
  - **Public/Private Status**: Private (API access).
  - **Technical Summary**: 175B parameters, autoregressive Transformer.

Initial Grok Response on GPT-3

2020 – "Language Models are Few-Shot Learners" (GPT-3)

- **Authors**: Tom B. Brown, Benjamin Mann, Nick Ryder, et al.
- **Institution**: OpenAI
- **Abstract/Key Contributions**: Introduced GPT-3, a 175B-parameter autoregressive language model excelling in few-shot learning, published in *NeurIPS*. Demonstrated emergent capabilities without fine-tuning.
- **Technical Contributions**:
  - **Architecture**: Scaled Transformer decoder with 96 layers, 12,288-dimensional embeddings, and 96 attention heads, trained on 45TB of text (Common Crawl, books, Wikipedia).
  - **Few-Shot Learning**: Prompted with examples (e.g., "Translate: 'Hello' → 'Hola'"), predicted via $p(x_t|x_{<t})$, leveraging in-context learning over 175B parameters.
  - **Training**: Used 10,000 V100 GPUs, ~3.14x10^23 FLOPs over weeks, optimized with mixed-precision.
- **Impact Assessment**:
  - Proved scale could yield general-purpose intelligence, shifting NLP to "foundation models."
  - Highlighted limitations: factual inaccuracies and high inference costs.
- **Subsequent Developments**:
  - Led to ChatGPT (2022) and GPT-4 (2023), refining usability and reasoning.
  - Inspired competitors (e.g., PaLM, 2022; LLaMA, 2023).
- **Model Lineage**:
  - **Model**: GPT-3 (2020)
  - **Development Context**: Trained on proprietary corpus, accessed via API.
  - **Public/Private Status**: Private (OpenAI API).
  - **Technical Summary**: 175B parameters, ~10^12 FLOPs/query, 96-layer Transformer decoder.
- **Contextual Linkage**: Extended GPT-2 (2019) and Transformer (2017), betting on scale over architecture.
- **Socio-Technical Commentary**:
  - Amazed users with versatility (e.g., code generation), fueling AI hype.
  - Raised concerns: energy use (~500 MWh), bias amplification, and API gatekeeping.

"Deeper" Grok Response on GPT-3

Practical Tips on Using Deep Research

- Deep Research approaches tend to do better with highly-structured, very detailed initial prompts
  - Recommendation: Paste your potential prompt into o3-mini, or other model and ask it to make it detailed, clear, and verbose. Provide specific sections/format you want your output to be in. Provide specific sources when available
- Deep Research tends to provide a lot of information
  - Recommendation: Let Deep Research provide as much information as possible, then summarize the output using another model (gpt-4o, o1, 2.0-Flash) – Claude seems to do a great job of breaking down the info

Deep Research Use Cases

- Market Research
  - Fantastic for identifying large number of relevant sources though initial “Googling”
- Technical Topic Deep Dive
  - Can narrow the focus of Deep Research to certain topics / websites / organizations / paper titles / etc.
- Personalized reports on making purchasing decisions
  - OpenAI demo using Deep Research to figure out best snowboard to purchase
- In general, if your problem would benefit from multi-step context from the internet, Deep Research is probably a good fit.

What’s next for Deep Research?

- Increased tool use effectiveness with tools like Anthropic’s Computer Use Agent, OpenAI’s Operator, and other browser manipulation tools
- As reasoning models (o1, o3-mini, gemini-2.0-pro, deepseek, etc.) improve, expect the outputs of Deep Research to continue improving.
- OpenAI plans to bring Deep Research to Plus and Free Tiers in coming months
- Expect to see different “levels” of intelligence for deep research based on pricing tiers (potentially tied to time spent)