
To Be (lieve)
or
Not To Be (lieve)
RambleGPT
Claudia
Twins
DeepFind
SouthWIND
CroK
Clot
Alpaca
in
Artificial Intelligence

# 2B or more than 2B ?

that is the scientific question


Dr Alan Whitfield
U3A Lymington
6 Nov 2025
alanhwhitfield@aol.co.uk

**Outcomes**

Medical Breakthroughs
- Cancer Diagnosis & Cures
- Antibiotics, Drugs, Speech

Productivity
- Professional Services

Fake & Fraud
- news, information, persona, dissertations

Jobs

Education

Culture & Faith

Environmental

**Application Areas**

AI Search

Medical

Legal

Robotics
- Humanoids
- Drones
- Detectors

IoThings

Computer Vision

Human Devices & Interfaces

Autonomous Vehicles

**Models & Approaches**

Intelligence
- Human
- Artificial


Machine Learning (ML)
Neural Networks
Large Language Models (LLM)
Generative AI
Transformer Models
Active and total parameters vs tokens
Agents; RAG; MoE

**Infrastructure**

Data Centres

Specific Chips
- CPU, GPU

Electrical Power

**Resources**

Data
- www.....
- "private" information

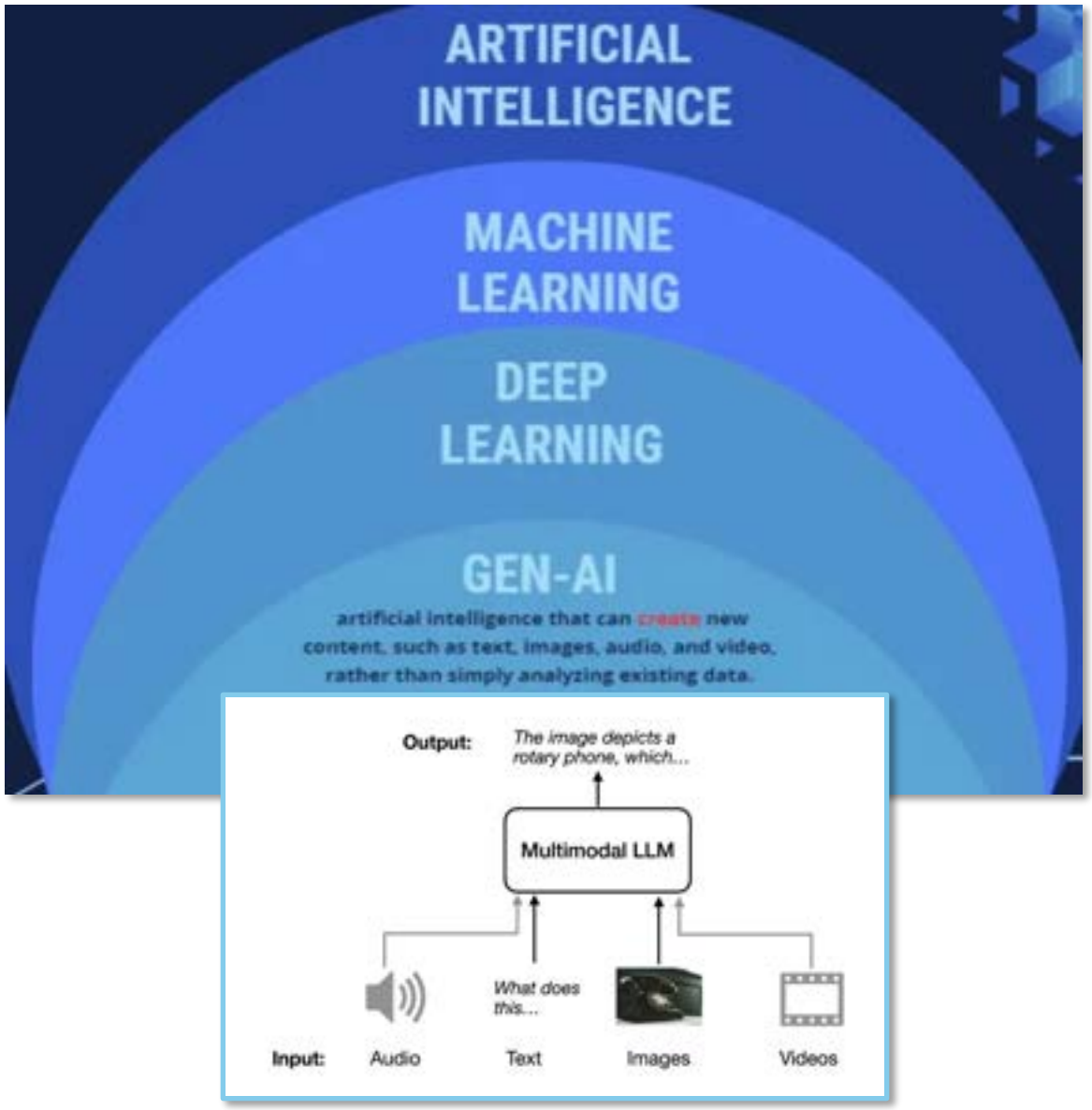
Finance
- Infrastructure
- Salaries
- Investment

People
- Model Architects, Model Developers,
- Data Scientists, Mathematicians
- Prompt Engineers

https://youtu.be/fr8FcFtkDK4?t=18

# What is AI ?

**Artificial Intelligence (AI)** is a broad field that aims to simulate human intelligence and behavior. Under its umbrella are machine learning, deep learning, and generative AI. All three concepts share a common foundation: learning from data.
**Machine Learning (ML)** is a subset of AI that involves training algorithms to recognize patterns and make predictions based on data.
**Deep Learning** is a specialized type of machine learning that utilizes *neural networks*, inspired by the structure of the human brain. These networks can process complex patterns and learn from large datasets.
**Generative (Gen-) AI** is a branch of AI that can create new content, such as text, images, or audio, by learning from existing data.




**Large Language Models (LLMs)**, like GPT, are a subset of generative AI initially designed to ingest, process and generate text.
Discovered and proposed by Google scientists in 2017, solving the problem of massive scale training data consumption, https://arxiv.org/abs/1706.03762 , LLMs use "**transformer**" architectures to analyze and "understand" vast amounts of text data. This enables them to generate human-quality text, even for tasks they haven't been explicitly trained on (known as zero-shot learning).

Many LLMs are capable of processing both input, correlation and output of multiple forms of media. They are sometimes called **multimodal LLM** (though since many most LLMs now handle multimedia the term multimodal is often dropped)

**Data Sources**

- Web Pages (Common Crawl)
- Books & Literature
- Online Publications & Research Articles
- Video Transcripts
- Code Archives
- *Proprietary/Private Datasets (for specific enterprise models - inputs to RAG)*

## Data Collection & Preparation

**Manipulation/Processing** (aka "**wrangling**"): The raw data undergoes extensive cleaning, filtering, and structuring.

- Extraction & Cleaning: Removing irrelevant content, HTML tags, and errors
- Filtering: Ensuring data quality and safety
- Tokenization: Converting text into numerical tokens that the model can process
- Structuring: Organizing data (e.g., creating knowledge graphs, handling tabular data)

**Training Data**

## Model Design & Training

Key **design factors of an LLM** include the tokenization scheme, number of internal (a billion or more) and input ("context window") parameters; the number of layers and heads of attention; degree of reinforcement learning.

The processed, massive dataset is used to **train the LLM**, enabling it to learn **patterns, grammar, and knowledge embedded** within the data. This creates the model's internal parameters (often in the billions) and knowledge representations.

**An LLM**

**User Query (or "prompt")**

## Inference & Application

- Prompt Tokenisation: The system tokenises a user "prompt" (text + any additional files).
- [*Retrieval Augmentation: If required (for privacy) further information (usually private files/databases) can be added to generate an enriched prompt (RAG) which is sent to the LLM*]
- LLM Response: The LLM processes the input, "reasons" over the provided "context", and generates a response. The final answer can include text, images, instructions, suggestions or code.

**Time**
- Highly time-consuming, iterative and labour-intensive
- 2-8 weeks for simpler AI projects
- 4+ months for a “frontier” LLM

**Time**
- LLM model design phase for a frontier model involves expert human effort over a period of several weeks to a few months

**Time**
- Training runs for state-of-the-art models typically take up to a few months.

## Data Collection & Preparation

Trainin g Data

## Model Design

## Model Training

An LLM

**People**
- Requires a specialized team including data scientists, data engineers, and domain experts
- Significant part of the human effort involves establishing "ground truth" and evaluation metrics to objectively measure the model's accuracy, which initially involves manual reviews of hundreds or thousands of test cases before automation can be fully implemented.
- Human data is a major cost component. Some suggest the cost of acquiring and preparing high-quality human data can be more expensive than the actual compute used for training the model.

**People**
- This phase requires a small, highly specialized team of top-tier ML researchers and data scientists with PhD-level knowledge and practical experience in deep learning, distributed systems, and AI architecture.
- Close collaboration with the data engineering teams and the infrastructure teams to ensure the proposed design can be trained efficiently on the available hardware (e.g., thousands of GPUs/TPUs)

**People**
- The training phase is not entirely automatic. A team of skilled ML engineers and researchers continuously monitors the process to detect/troubleshoot issues
- Overall R&D staff costs are a substantial part of the total cost, often 30-50%

**Compute**
- Non-stop massive clusters of thousands of high-performance GPUs (like Nvidia A100s or H100s)
- $10^{25}$ FLOPs … growing around 5x a year

**Energy**
- 40+ GigaWatt-hours (GWh) total
- 100+MW non-stop electricity for a small city

**User Query (or "prompt")**

**Inference & Application**

**An LLM**

**People**
- Millions of people – both casual and professional - search, comparison and "discovery"
- **ChatGPT** 700-800 million weekly, 190 million daily, 5.7 billion visits each month, serves around 3 billion queries a day
- **Google Gemini** 650 million monthly, 35 million daily, 2 billion monthly users of "AI Overviews"

**Compute**
- Non-stop massive clusters of thousands of high-performance GPUs (like Nvidia A100s or H100s)
- **80-90% of the total computational effort for AI is now used for inference !**
- cost per query is a critical competitive metric

**Energy**
- Per Query on Cloud: 0.24Wh Google Gemini 0.43Wh GPT-4o (simple prompt)
- 1,000 GWh Total Annual Energy …. the power consumption of a small city population of 100,000 (in USA)
- On-demand energy hundreds of MW

**Some Use Cases**
- Business: Companies favour ChatGPT for brainstorming, report automation, and technical support. Gemini's strong integration with Google Workspace streamlines collaborative document editing and productivity tools.
- Creative Projects: ChatGPT's coding, storytelling, and content generation lead; Gemini excels at image editing and multimodal project assistance.
- Fitness/Sports: Gemini is leveraged for workout tracking and health summaries via Google Health; ChatGPT assists with personalized routines and sport analytics.
- On-the-go (Outdoor): Gemini on Android edge devices provides local AI inference, while ChatGPT's mobile app offers versatile travel or planning support.
- Sleep/Wellness: Both are used to generate sleep stories, meditation scripts, or curated wellness content; Gemini's multimedia support expands options.

# Choosing an LLM - Key Principles

1. **Functionality** - The model must be suited to the user's specific needs.
    - Task Suitability: Identify the primary tasks the model needs to perform effectively (e.g., creative writing, coding assistance, quick fact-checking, language translation, or basic summarization).
    - Accuracy and Quality: The output must be factually correct and coherent enough for the intended use.
    - Context Handling (if needed): The ability to process and remember longer conversations or documents is crucial for complex interactions (a large "context window").
    - Multimodal (if needed): Check if the model can handle inputs beyond text, such as images or voice, if required.

2. **User Experience and Performance** - A good consumer LLM should be intuitive and fast enough for real-time interaction.
    - Speed and Latency: For an interactive chat experience, the response time (latency) should be low. Local models (e.g. MS Copilot; ChatGPT Atlas) generally have very low latency but require fairly powerful hardware, while cloud-based models depend on network connectivity and utilisation.
    - Offline Access: For users who need functionality without a constant internet connection, on-device models or those with a hybrid approach that allows for offline operation are a key consideration.

3. **Data Privacy and Security** - For personal use, privacy is a paramount concern.
    - Data Handling: Understand how user data is processed. Cloud-based models send data to external servers, which can raise privacy concerns for sensitive information.
    - Local Processing: Models that run entirely on the device (on-device LLMs) ensure data never leaves the user's control, which is ideal for privacy-sensitive data.
    - Compliance/Trust: Look for clear policies on data retention, use, and security certifications from the provider, if available.

# Choosing an LLM - Key Principles

4. **Cost and Value** - Evaluate the financial implications, whether through direct costs or indirect requirements.
    - Pricing Structure: Compare different cost models: free access (often with limitations), subscription fees (monthly/annual), or per-token API usage charges.
    - Hardware Requirements: If using a locally run model, consider the upfront cost of high-performance hardware (e.g., a powerful GPU).
    - Value for Money: Determine if the quality and features justify the cost for your particular usage patterns.

5. **Customization and Ecosystem** - Future-proofing and integration capabilities enhance long-term value.
    - Flexibility: Choose a model that allows for easy switching or upgrading as newer, better models are released.
    - Ecosystem: Consider the range of tools, integrations, and community support available. An active community or robust developer support can be invaluable for troubleshooting or expanding use cases.

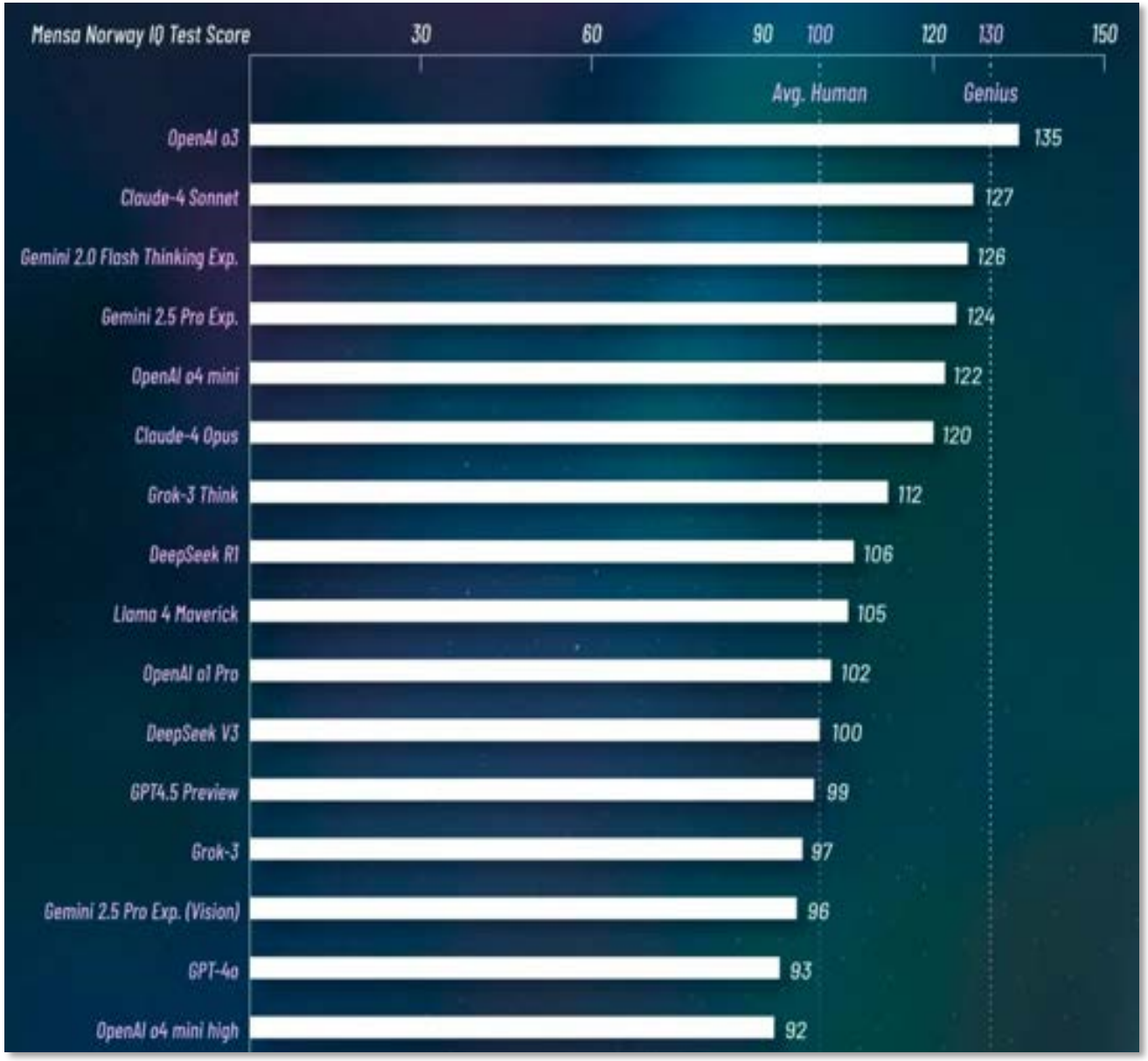
# What is Intelligence – and how to measure it ?

Wikipedia on “Intelligence”:

*the capacity for* ….
abstraction, logic, understanding,
self-awareness, learning,
emotional knowledge, reasoning,
planning, creativity,
critical thinking, and problem-solving.

For an extensive consideration of IQ …
https://en.wikipedia.org/wiki/Intelligence_quotient

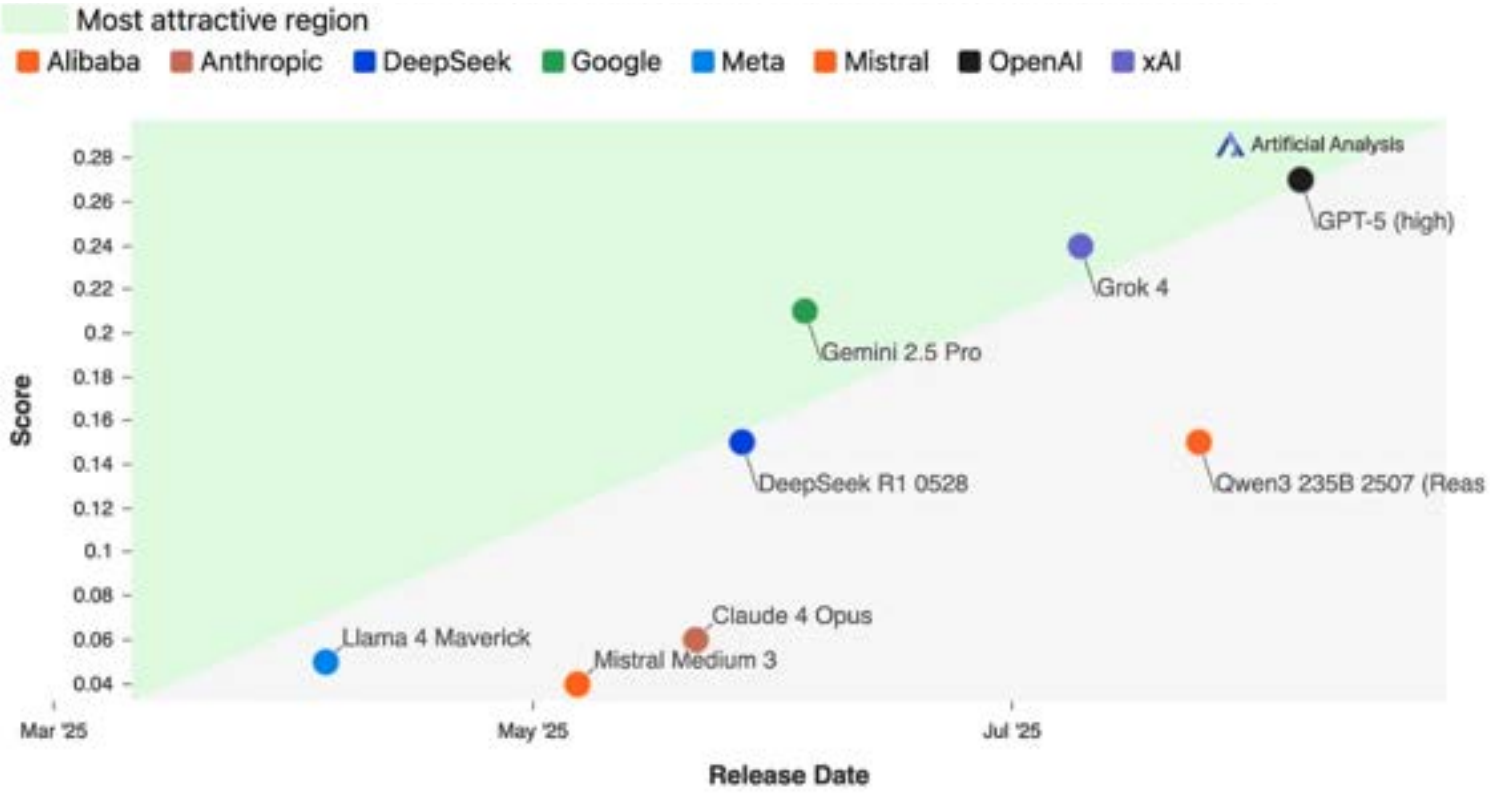
The test results to the right look persuasive
BUT the test was the Mensa Norway IQ Test
This test is based on various image recognitions and conclusions …. There are many such images that a strong LLM has “learned” and it has the ability to correlate them accurately.



## Top LLMs as of November 2025

| Model | Major Strengths | Major Weaknesses |
|---|---|---|
| OpenAI GPT-5 | General-purpose intelligence, top-tier reasoning, multimodal understanding, robust safety features, and vast ecosystem integration. | High cost for extensive usage; closed-source with limited transparency. |
| Anthropic Claude 4.5 Sonnet/Opus | State-of-the-art coding and agentic tasks performance, strong ethical alignment and safety, and long context window (up to 1M tokens in beta). | Can be conservative in creative tasks; higher pricing for the top Opus model. |
| Google Gemini 2.5 Pro | Excellent multimodal capabilities (text, images, audio, video), deep integration with Google services, and strong performance in complex reasoning. | Proprietary nature limits customization compared to open models; high computational requirements. |
| xAI Grok 4 | Real-time information access/web integration, strong performance in coding benchmarks, and freely available globally with usage limits. | Newer to the market, user feedback still mixed; can be less predictable in general dialogue. |
| DeepSeek-V3 (Open-source) | Top open-weight model for logic and math tasks, cost-efficient, fast inference speed, and open-source for full customization. | Smaller ecosystem and community support compared to Llama; safety layers must be added by the user. |
| Meta Llama 4 (Maverick/Scout) | Open-source framework, highly tunable and flexible, multimodal support, with ultra-long context options (Scout). | Base models lack pre-built safety/moderation layers, requiring custom implementation for sensitive applications. |
| Mistral Medium 3 / Mixtral | High efficiency and speed (low latency), strong multilingual capabilities, and competitive performance at a lower cost. | Slightly behind top proprietary models on the most difficult reasoning benchmarks. |

## Humanity's Last Exam (HLE) – a Rigorous Test of Knowledge with Reasoning




Humanity's Last Exam (HLE) is a rigorous AI benchmark consisting of 2,500-3,000 questions across more than 100 academic disciplines.

Each question has a clear-cut answer—multiple-choice or exact-match short answer—the model either gets it right or wrong. Since answers aren't available online, AI can't just search its way to success; it must genuinely **reason**.

The best AI models score below 30% (0.3 above), while human experts reach nearly 90% on the same questions

# Choosing an LLM – Where To Go

| LLM Name | Developer | Release Date | Context Length | License | Active Parameters | |
|---|---|---|---|---|---|---|
| GPT-5 | OpenAI | Aug-25 | 272 k | Proprietary | 2 to 5+ trillion | https://platform.openai.com/docs/models |
| GPT-4.1 | OpenAI | Apr-25 | 1 M | Proprietary | 1.8 trillion | |
| GPT-4o | OpenAI | Mar-25 | 128 k | Proprietary | | |
| o3-pro | OpenAI | Apr-25 | 200 k | Proprietary | | |
| o3 | OpenAI | Apr-25 | 200 k | Proprietary | | |
| o4-mini | OpenAI | Apr-25 | 200 k | Proprietary | | |
| o3-mini | OpenAI | Jan-25 | 200 k | Proprietary | | |
| Gemini 2.5 Pro | Google | Mar-25 | 1 M | Proprietary | 20+ billion | https://gemini.google.com/app |
| Gemini 2.5 Flash | Google | Apr-25 | 1 M | Proprietary | 5 billion | |
| Llama 4 Scout | Meta AI | Apr-25 | 10 M | Open Source | 17 billion | https://www.llama.com |
| Llama Nemotron Ultra | NVIDIA | Apr-25 | 128 k | Open Source | 70 billion | https://build.nvidia.com/search/models?q=Nemotron&ncid=no-ncid |
| Grok 4 | xAI | Jul-25 | 256 k | Proprietary | 1.7 trillion | https://grok.com |
| Grok 3 Mini | xAI | Feb-25 | 1 M | Proprietary | | |
| Mistral Medium 3 | Mistral AI | May-25 | 128 k | Proprietary | | https://docs.mistral.ai/getting-started/models |
| Claude Opus 4 | Anthropic | May-25 | 200 k | Proprietary | 300 to 500 billion | https://docs.claude.com/en/docs/about-claude/models/overview |
| Claude Sonnet 4 | Anthropic | May-25 | 200 k | Proprietary | | |
| DeepSeek-R1 | DeepSeek | Jan-25 | 128 k | Open Source | 671 billion (37B active) | https://www.deepseek.com |
| DeepSeek-R1-0528 | DeepSeek | May-25 | 128 k | Open Source | 671 billion (37B active) | |
| Qwen3-235B-A22B-Thinking-2507 | Alibaba | Jul-25 | 262 k | Open Source | 235 billion (22B active) | https://novita.ai/models/llm/qwen-qwen3-32b-fp8 |
| Perplexity Sonar | Perplexity | Oct-25 | 128 k | Open Source | 70 billion | https://www.perplexity.ai |
| MiniMax-Text-01 | MiniMax | Jan-25 | 4 M | Open Source | 45.9 billion | https://www.minimax.io |
| | | | | | | |
| | | | | | | |
| **Browsers** | | | | | | |
| ChatGPT Atlas | OpenAI | | | | | https://openai.com/index/introducing-chatgpt-atlas/ |
| Google Chrome - Gemini AI | Google | | | | | https://www.google.com/chrome/ai-innovations/ |
| Perplexity Comet | Perplexity | | | | | best ask perplexity.ai !!!! |
| | | | | | | |
| | | | | | | |
| **Multiple Choice** | | | | | | |
| Abacus.ai | multiple above | | | | | https://chatllm.abacus.ai |
| Poe by Quora | multiple above | | | | | https://poe.com/login |
| | | | | | | |
| **Embedded In Office Automation** | | | | | | |
| Copilot | Microsoft | Sep-25 | 128 k | Proprietary | ? 7 to 200+ billion | https://www.microsoft.com/en-gb/microsoft-365-copilot |
| Claude for Excel (and Slack!) beta | Anthropic | Oct-25 | 200 k | Proprietary | 300 to 500 billion | https://claude.com/claude-for-excel (claude-and-slack) |

# Using an LLM - Key Prompting Principles

**Be Direct and Clear**: Use straightforward, natural language. Avoid vague language or overly complex sentences.

**Define the Goal**: Clearly state what you want the model to do (e.g., "Summarize", "Write", "Explain", "Classify", "Translate").

**Provide Context**: Give the necessary background information so the model understands the situation. Perhaps give the model a role (e.g. act as a “professional graphic designer”; act as a ”accountant”; …)

**Set Constraints**: Define limitations on length, tone, or audience (e.g., "in 3 sentences," "for a general audience," "using a professional tone").

**Avoid "Don'ts":** Focus on telling the model what to do rather than what not to do (e.g., say "Write in an active voice" instead of "Do not use passive voice").

**Specify the Output Format**: If you need a specific format (e.g., a list, a paragraph, a table, or a file type (e.g. jpg or Gif for an image; mp4 or FLAC for audio), state it explicitly. This helps structure the response.

**Use Examples (Few-Shot Prompting):** For more complex tasks or to ensure a specific style, provide one or two examples of the desired input and output to guide the model.

**Upload one or more files**: when a picture paints a thousand words (*but remember any personal or company privacy concerns*)

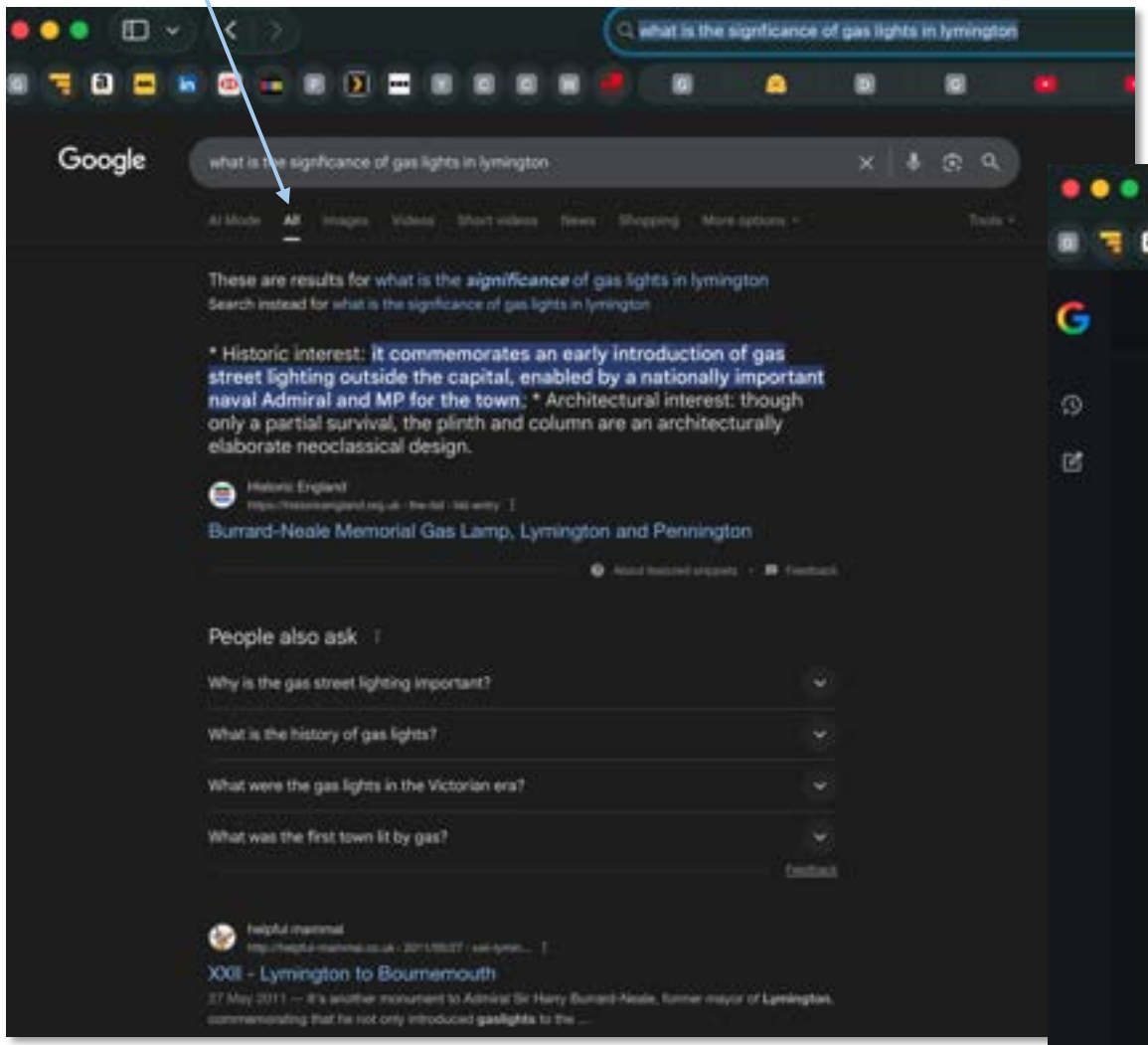
**Iterate and Refine**: Start with a simple prompt and add details based on the model's initial responses. It's an iterative process. Each LLM has a memory - so you seldom need to repeat the context as a whole .. simply adjust (be more specific) the bits you want to improve

**Push the LLM to think more deeply**: Add the phrase(s) …. “**Take you time” “Explain your reasoning”**…. To your prompt

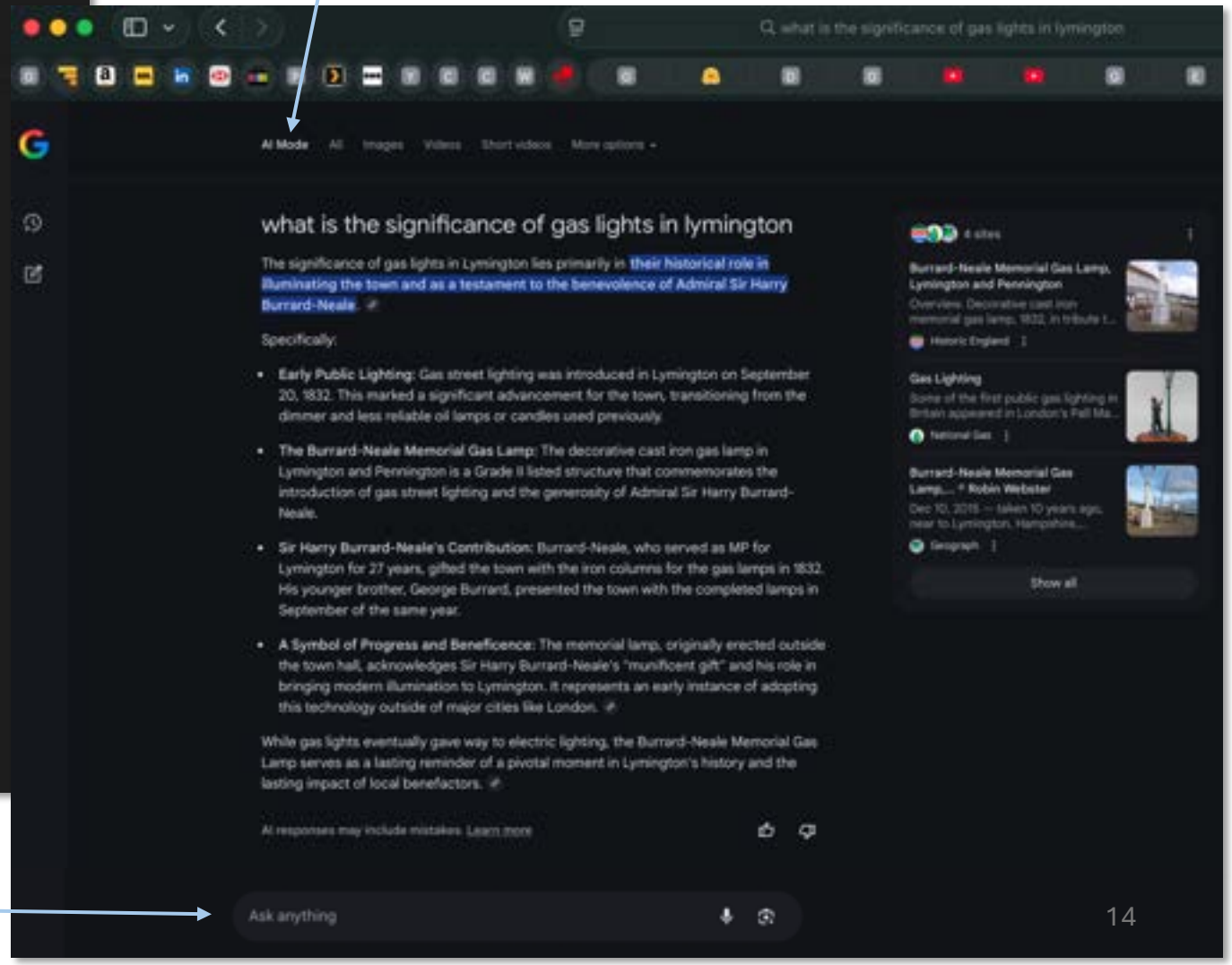
**Break a complex request task into smaller steps**: One LLM weakness is successfully completing a long chain-of-thought

# Google Search AI - an LLM (Google's Gemini 2.5) in action

1. Google Search in "All" mode




2. Google Search in "AI Mode"



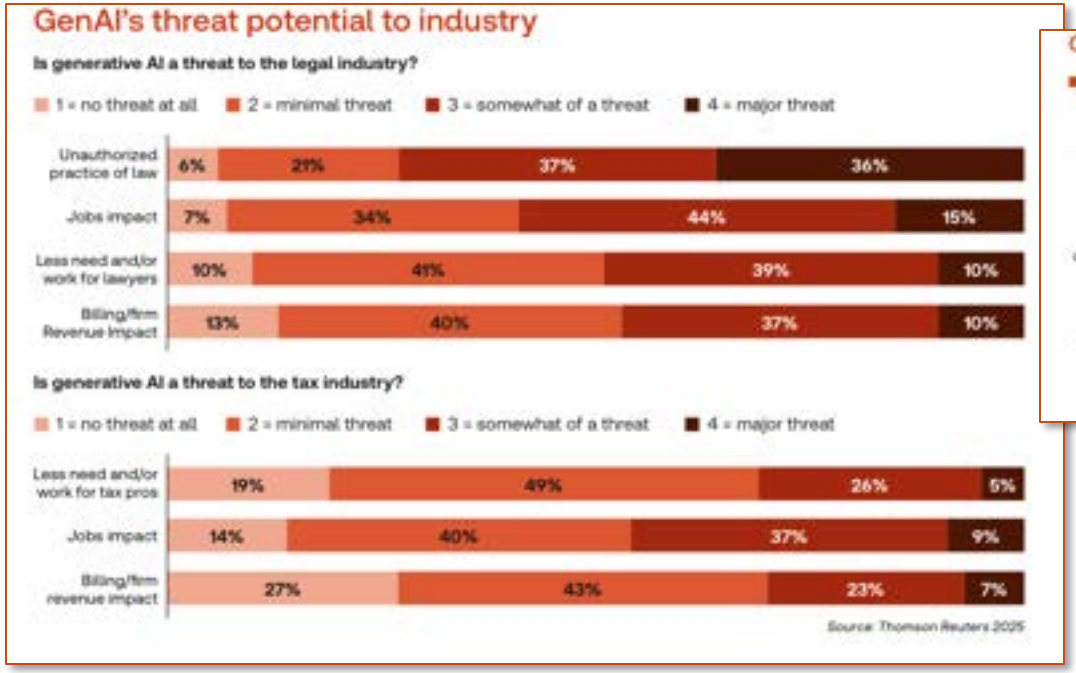

In AI Mode, your question ("prompt") is at the bottom of the page

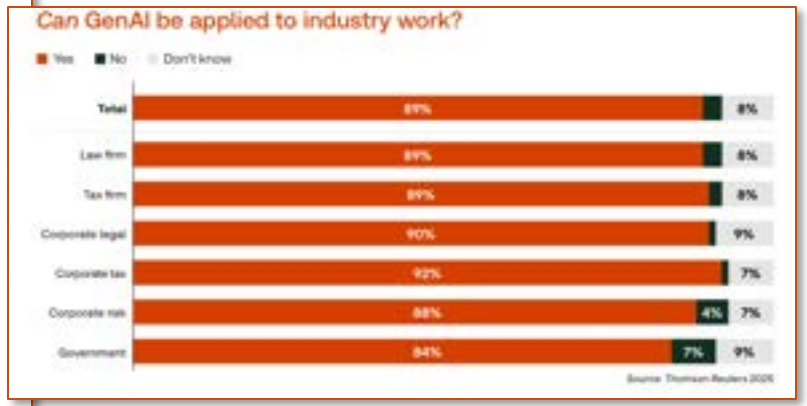
# AI is fundamental to Business – Professional Services

https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/future-of-professionals-report-2025.pdf

https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/2025-generative-ai-in-professional-services-report-tr5433489-rgb.pdf

https://www.thomsonreuters.com/en/c/future-of-professionals




# AI is fundamental to Business – Industry

Boston Dynamics (owned by Hyundai) : Atlas

https://youtu.be/HYwekersccY
https://youtu.be/I44_zbEwz_w

The Robot Takeover is Here !
https://youtu.be/WcHd1zSofcA?t=4

… and China may well be in the lead
https://youtu.be/giyl27gKvS4

# AI is fundamental to Research – Proteins and Material Science

**AlphaFold - The Most Useful Thing AI Has Ever Done**

https://www.youtube.com/watch?v=P_fHJIYENdI

**How could AI transform Materials Science research?**

https://www.youtube.com/watch?v=wxt9_e-Agnk

## AI is fundamental to Defence

The war in Ukraine has become a testing ground for AI-powered warfare, demonstrating the potential of these technologies to transform the battlefield.

### Intelligence and Targeting

*Enhanced Intelligence Gathering*: AI is being used to analyze vast amounts of data from various sources, including satellite imagery, social media, and intercepted communications, to identify enemy positions, movements, and intentions.

*Accelerated Targeting*: AI-powered systems are enabling faster target identification and engagement, potentially reducing the time available for human decision-making in critical moments.

*Real-time Battle Damage Assessment (BDA)*: AI algorithms are being used to assess the effectiveness of strikes and adapt strategies in real-time.

*Facial Recognition*: Ukrainian forces have been using facial recognition technology, including Clearview AI, to identify dead soldiers, potential assailants, and to combat misinformation.

*Counter-Disinformation*: AI is being employed to detect and counter propaganda and disinformation campaigns.

### Autonomous Systems

*Drones*: Both sides are utilizing AI-powered drones for reconnaissance, surveillance, and attack missions.

*Autonomous Weapon Systems*: There are reports of Russia deploying AI-powered drones that can autonomously identify and strike targets, raising concerns about the potential for unintended civilian casualties and the erosion of traditional warfare norms.

*Countering Electronic Warfare*: Ukraine has developed AI tools to help its drones evade Russian jamming and maintain target lock, demonstrating the ongoing technological arms race in this domain.

https://youtu.be/RLVBFuTVmGc

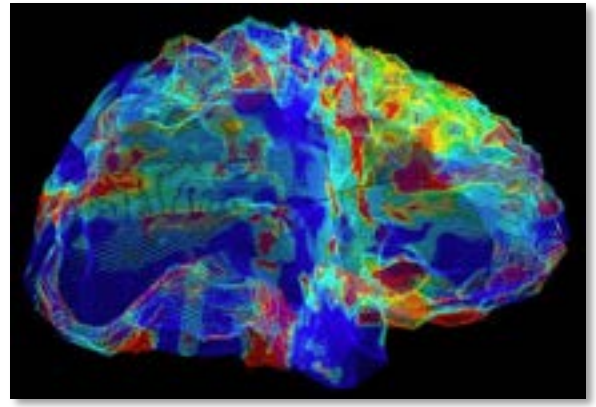
# AI is fundamental to Health

14 August: Artificial intelligence has invented two new potential antibiotics that could kill drug-resistant gonorrhoea and MRSA. The drugs were designed atom-by-atom by the AI and killed the superbugs in laboratory and animal tests.
https://www.bbc.co.uk/news/articles/cgr94xxye2lo

20 August 2025: MND left her without a voice. Eight seconds of scratchy audio gave it back to her
MND left her without a voice. Eight seconds of scratchy audio gave it back to her ... using AI
https://www.youtube.com/watch?v=JErn_9nUHrQ

20 December 2024: New AI software from Imperial College London and Edinburgh University is "twice as accurate" as professionals at examining the brain scans of stroke patients. Two UK universities trained the software on a dataset of 800 brain scans of stroke patients and then trialled it on 2,000 patients.

22 October 2024: NICE says broken bones are missed in 3-10% of cases - it is the most common diagnostic error in emergency departments. So using AI to do the initial scan could potentially avoid both unnecessary X-rays and missed fractures.



10 November 2023: Machine-learning approach detects Alzheimer's disease with an accuracy of more than 90% — a potential boon for clinicians and scientists developing treatments
https://www.nature.com/articles/d41586-023-03482-9

https://www.weforum.org/stories/2025/08/ai-transforming-global-health/

# Two final thoughts

Caveat utentis

&

Operam dare

Thank You

# Several Annexes

A : Types of Reasoning
B: Stanford University : AI Index Report 2025 Top Takeaways
C: How Big is the World of AI – just a few measures
D: Refining Data …. Data starts out dirty and caked in “sludge” … ends clean and usable
E: Data Tokenisation
F: Training Data Resources for LLM
G: Will we run out of human-generated data ?
H: Crash Course in Size Acronyms – large and larger
I: Crash Course in Size Acronyms – small and smaller

## Annex A - 7 types of reasoning [https://www.indeed.com/career-advice/career-development/types-of-reasoning]

**1. Deductive reasoning**

Deductive reasoning is a type of reasoning that uses formal logic and observations to prove a theory or hypothesis. In deductive reasoning, you start with an assumption and then make observations or rational thoughts to validate or refute the assumption. You can use deductive reasoning to apply a general law to a specific case or test an induction. The results of deductive reasoning typically have a logical certainty.

*For example, a marketing division evaluates data and reaffirms that their company's biggest demographic is young parents. Based on this information, they decide to allocate more of the marketing budget to social media platforms that target that group.*

**2. Inductive reasoning**

Inductive reasoning uses theories and assumptions to validate observations. In some ways it's the opposite of deductive reasoning, as it involves reasoning from a specific case or cases to derive a general rule. The results of inductive reasoning are not always certain because it uses conclusions from observations to make generalizations. Inductive reasoning is helpful for extrapolation, predictions and part-to-whole arguments.

*For instance, a kindergarten teacher has struggled to hold the attention of her class throughout the morning. She tries adding an extra five-minute activity break one hour after school starts. After a week of mood improvements and attention gains, she decides to permanently add the extra activity break.*

**3. Analogical reasoning**

Analogical reasoning is a form of thinking that finds similarities between two or more things and then uses those characteristics to find other qualities common to them. It's based on the brain's tendency to notice patterns and make associations. Once the brain recognizes a pattern, it can associate the pattern with specific things, and this leads to analogous reasoning. Analogous thinking can help you expand your understanding by looking for similarities between different things.

*A supermarket has served as an analogical source for many businesses. When planning a new business, evaluating how to serve customers better, or planning a new line, many business strategists reach for a supermarket analogy to ask if they can provide everything a customer may need when shopping for items in their category.*

**4. Abductive reasoning**

Abductive reasoning is a type of reasoning that uses an observation or set of observations to reach a logical conclusion. It's similar to inductive reasoning, however, abductive reasoning permits making best guesses to arrive at the simplest conclusions. Abduction has applications in troubleshooting and decision-making, especially when dealing with uncertainties. Abductive reasoning is especially useful when explaining an observation or phenomenon that the observer has very little or no existing knowledge about. The conclusion of abductive reasoning may not always be certain and may require further verification.

*For example, salespeople may use this type of reasoning when they receive a short correspondence from a client, asking them to reply quickly about an issue. When the client doesn't give enough information to understand before responding, a salesperson can use abductive reasoning to narrow down possible concerns. It's sometimes better to prepare answers for a few best guesses.*

*the operation of the whole.*

**5. Cause-and-effect reasoning**

Cause-and-effect reasoning is a type of thinking in which you show the linkage between two events. This reasoning is used to explain what may happen if an action takes place or why things happen when some conditions are present. This type of reasoning commonly guides everyday decision-making, in cases where people draw on personal experience and a desire to improve. Businesses and professionals also use prediction and forecast modeling. This type of reasoning can help people trust your arguments, especially if you are accurate most of the time.

*For instance, a marketing agency may use cause-and-effect reasoning to prove the value of their campaigns and request an increase in budget. They may show how in the first year they launched an advertising campaign for a product line before the holidays, and the product sales increased by 10%. The following year, they increased the advertising budget 15%, and the product sales increased 25%. Therefore, with a budget increase of 20%, they're expecting a sales increase of 30%.*

**6. Critical thinking**

Critical thinking involves extensive rational thought about a specific subject in order to come to a definitive conclusion. It's helpful in fields such as computing, engineering, social sciences and logic. Critical thinking plays a vital role in problem-solving, especially when troubleshooting technical issues. It's used to assess the authenticity of works of art, literature, films and other artistic expressions. Critical thinking also plays a vital role in mental and emotional matters, grey areas and other fields that deal with subjects less understood.

*For example, the general manager of a family restaurant learns that a bakery important to its supply chain is about to go on strike. They order extra baked goods to freeze and then plan a distributor they can use during the strike.*

**7. Decompositional reasoning**

Decompositional reasoning is the process of breaking things into constituent parts to understand the function of each component and how it contributes to the operation of the item as a whole. By analyzing each part independently, decompositional reasoning allows an observer to draw powerful conclusions about the whole. You find this approach in several disciplines, including science, engineering, marketing, product development, game development and software development.

*Project management utilizes decompositional reasoning in its division of a project into components. A manager assigns each component to an individual, who is responsible for completion and communication about integration into the project. This division ensures the success of each component and contributes to the operation of the whole.*

Why You're Thinking About "Reasoning" All Wrong …..
https://www.wordrake.com/blog/youre-thinking-about-reasoning-wrong#_ftn1

Annex B - Top Takeaways

Stanford University : AI Index Report 2025 https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

1. *AI performance on demanding benchmarks continues to improve*. In 2023, researchers introduced new benchmarks—MMMU, GPQA, and SWE-bench—to test the limits of advanced AI systems. Just a year later, performance sharply increased: scores rose by 18.8, 48.9, and 67.3 percentage points on MMMU, GPQA, and SWE-bench, respectively. Beyond benchmarks, AI systems made major strides in generating high-quality video, and in some settings, language model agents even outperformed humans in programming tasks with limited time budgets.

2. *AI is increasingly embedded in everyday life*. From healthcare to transportation, AI is rapidly moving from the lab to daily life. In 2023, the FDA approved 223 AI-enabled medical devices, up from just six in 2015. On the roads, self-driving cars are no longer experimental: Waymo, one of the largest U.S. operators, provides over 150,000 autonomous rides each week, while Baidu's affordable Apollo Go robotaxi fleet now serves numerous cities across China.

3. *Business is all in on AI*, fueling record investment and usage, as research continues to show strong productivity impacts. In 2024, U.S. private AI investment grew to $109.1 billion—nearly 12 times China's $9.3 billion and24 times the U.K.'s $4.5 billion. Generative AI saw particularly strong momentum, attracting $33.9 billion globally in private investment—an 18.7% increase from 2023. AI business usage is also accelerating: 78% of organizations reported using AI in2024, up from 55% the year before. Meanwhile, a growing body of research confirms that AI boosts productivity and, in most cases, helps narrow skill gaps across the workforce.

4. *The U.S. still leads in producing top AI models—but China is closing the performance gap*. In 2024, U.S.-based institutions produced 40 notable AI models, compared to China's 15 and Europe's three. While the U.S. maintains its lead in quantity, Chinese models have rapidly closed the quality gap: performance differences on major benchmarks such as MMLU and Human Eval shrank from double digits in 2023 to near parity in 2024. China continues to lead in AI publications and patents. Model development is increasingly global, with notable launches from the Middle East, Latin America, and Southeast Asia.

5. *The responsible AI ecosystem evolves*—unevenly. AI-related incidents are rising sharply, yet standardized RAI evaluations remain rare among major industrial model developers. However, new benchmarks like HELM Safety, AIR-Bench, and FACTS offer promising tools for assessing factuality and safety. Among companies, a gap persists between recognizing RAI risks and taking meaningful action. In contrast, governments are showing increased urgency: In 2024, global cooperation on AI governance intensified, with organizations including the OECD, EU, U.N., and African Union releasing frameworks focused on transparency, trustworthiness, and other core responsible AI principles.

6. *Global AI optimism is rising*—but deep regional divides remain. In countries like China (83%), Indonesia (80%),and Thailand (77%), strong majorities see AI products and services as more beneficial than harmful. In contrast, optimism remains far lower in places like Canada (40%), the United States (39%), and the Netherlands (36%). Still, sentiment is shifting: Since 2022,optimism has grown significantly in several previously skeptical countries, including Germany (+10%), France (+10%), Canada(+8%), Great Britain (+8%), and the United States (+4%).

7. *AI becomes more efficient, affordable, and accessible*. Driven by increasingly capable small models, the inference cost for a system performing at the level of GPT-3.5 dropped over 280-fold between November 2022 and October 2024. At the hardware level, costs have declined by 30% annually, while energy efficiency has improved by 40% each year. Open-weight models are closing the gap with closed models, reducing the performance difference from 8% to just 1.7% on some benchmarks in a single year. Together, these trends are rapidly lowering the barriers to advanced AI.

8. *Governments are stepping up on AI*—with regulation and investment. In 2024, U.S. federal agencies introduced59 AI-related regulations—more than double the number in 2023—and issued by twice as many agencies. Globally, legislative mentions of AI rose 21.3% across 75 countries since 2023, marking a ninefold increase since 2016. Alongside growing attention, governments are investing at scale: Canada pledged $2.4 billion, China launched a $47.5 billion semiconductor fund, France committed €109 billion, India pledged $1.25 billion, and Saudi Arabia’s Project Transcendence represents a $100 billion initiative.

9. *AI and computer science education is expanding*—but gaps in access and readiness persist. Two-thirds of countries now offer or plan to offer K–12 CS education—twice as many as in 2019—with Africa and Latin America making the most progress. In the U.S., the number of graduates with bachelor's degrees in computing has increased 22% over the last10 years. Yet access remains limited in many African countries due to basic infrastructure gaps like electricity. In the U.S., 81% of K–12 CS teachers say AI should be part of foundational CS education, but less than half feel equipped to teach it.

10. *Industry is racing ahead in AI*—but the frontier is tightening. Nearly 90% of notable AI models in 2024 came from industry, up from 60% in 2023, while academia remains the top source of highly cited research. Model scale continues to grow rapidly—training compute doubles every five months, datasets every eight, and power use annually. Yet performance gaps are shrinking: the Elo skill score difference between the top and 10th-ranked models fell from 11.9% to 5.4% in a year, and the top two are now separated by just 0.7%. The frontier is increasingly competitive—and increasingly crowded.

11. *AI earns top honors for its impact on science*. AI's growing importance is reflected in major scientific awards: Two Nobel Prizes recognized work that led to deep learning (physics) and to its application to protein folding (chemistry),while the Turing Award honored groundbreaking contributions to reinforcement learning.

12. *Complex reasoning remains a challenge*. AI models excel at tasks like International Mathematical Olympiad problems but still struggle with complex reasoning benchmarks like Plan Bench. They often fail to reliably solve logic tasks even when provably correct solutions exist, limiting their effectiveness in high-stakes settings where precision is critical.

## C. How Big is the World of AI – just a few measures !

Size and shape of the AI Marketplace

- The current global AI market is valued at approximately $391 billion. [Oil & Gas market is $4Tn]
- The AI industry is projected to increase in value by around 5x over the next 5 years.
- The AI market is expanding at a CAGR of 35.9%. *(spurious accuracy ?!)*
- Amazon Web Services (AWS) generates US$117Bn annual revenue run rate from AI services

Costs of AI "Large Language Model (LLM)" technology (plus electricity) and people are huge – but so is the investment pouring in

- Mark Zuckerberg commits up to US$65bn by 2025 to construct data centres housing 1.3 million GPUs
- Amazon Web Services (AWS) commits US$100Bn in capital expenditures for 2025
- Microsoft is investing US$80bn to planned AI data centre and cloud infrastructure for 2025
- Industry wide spending on data centres and related infrastructure is forecast to be $3Tn by 2029
- Cost of new power plants $500Bn in the near future
- A four-year total pay package worth about $100m for a very senior leader is not inconceivable for Meta
- An AI researcher was offered - and turned down - an $18m job offer from Meta
- Thinking Machines Lab has closed a $2Bn seed round. The deal values the *6-month-old startup* (!) at $10Bn.

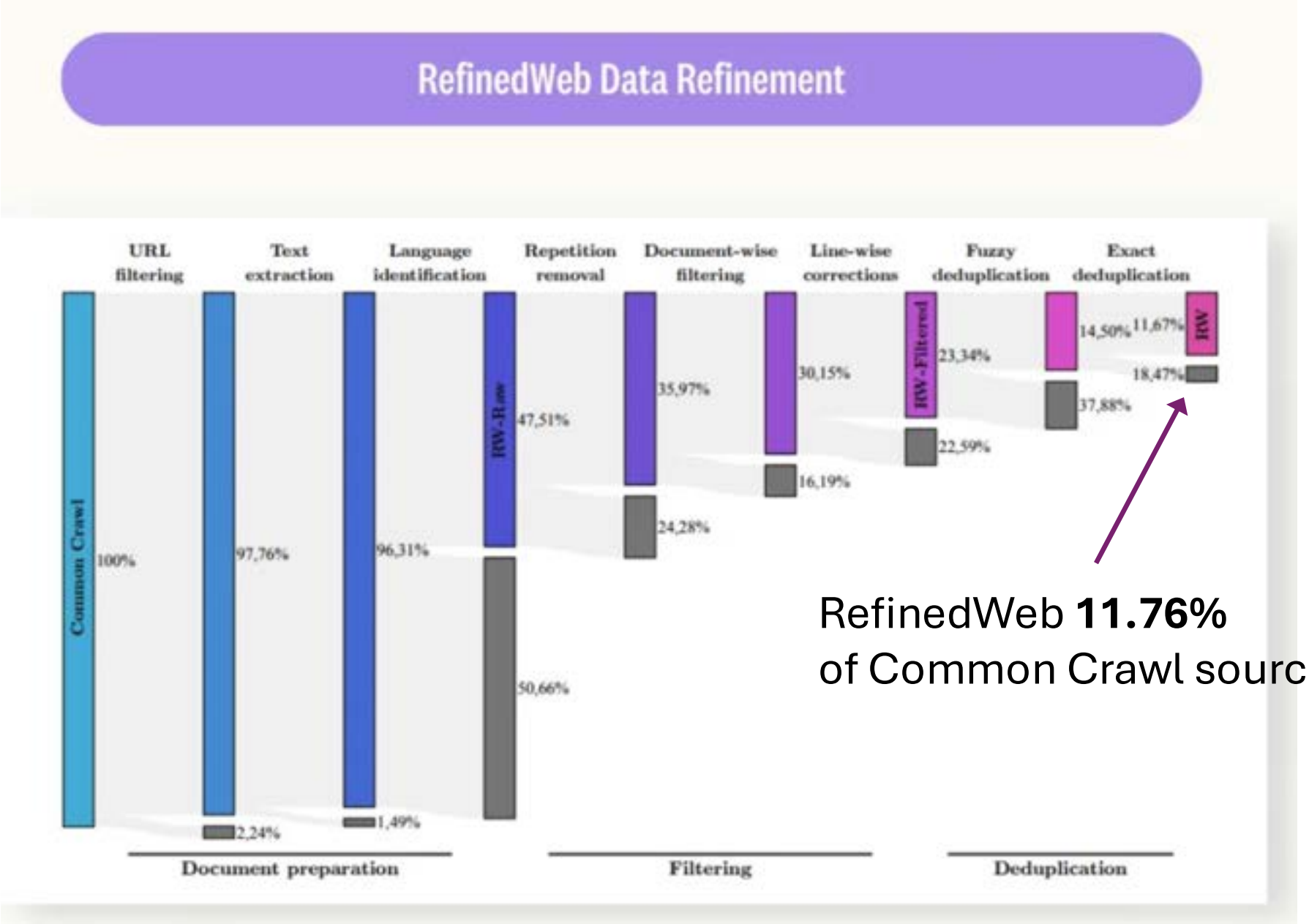
The World of AI at Work

- As of 2025, as many as 97 million people will work in the AI space.
- 83% of companies claim that AI is a top priority in their business plans.
- 65% of AI users are Millennials or Gen Z
- Netflix makes $1 billion annually from automated personalized recommendations.

… and Images (indeed Multimedia) are a fundamental part of the future landscape

- 34 million AI images are created every day.
- Over 15 billion AI images have been created since 2022.
- Generative AI reached 15 billion images in about 1.5 years – a feat that took traditional photography roughly 150 years to achieve.

D: Refining Data …. Data starts out dirty and caked in “sludge” … ends clean and usable




RefinedWeb **11.76%**
of Common Crawl source

**Data Wrangling -** is the process of cleaning, structuring, and transforming raw, messy data into a clean and usable format for analysis, machine learning, or other data-driven applications. It involves tasks like correcting errors, handling missing values, and merging different datasets to ensure data quality and usefulness for downstream purposes.

Tokens vs Words
The ratio of tokens per word depends on the tokenizer and the language. I have assumed 0.75 words per token, which is about right for English text using OpenAI tiktoken.

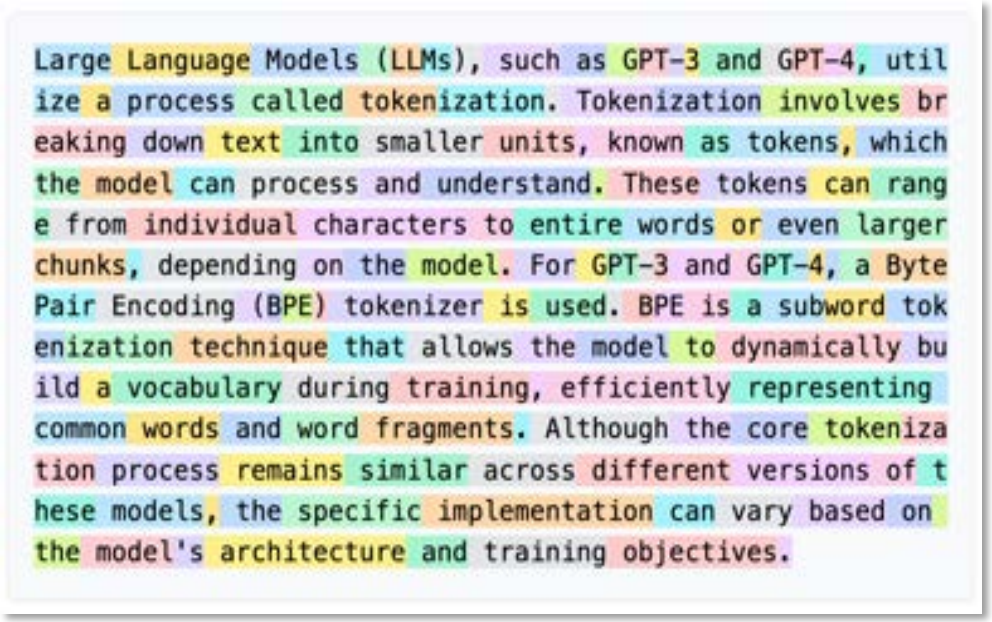
Common Crawl
Common Crawl is the basis for many large models, as it's a convenient source of massive-scale web data. It's sometimes referred to as being "the whole web". This isn't true, but it does cover a substantial fraction of all public HTML content. It misses dynamically rendered websites, PDF content, anything behind a login, etc. Google certainly has something much more comprehensive internally, and OpenAI and Anthropic also run their own custom crawls.

FineWeb
FineWeb comes in at 15 trillion tokens. It's a filtered English subset of Common Crawl dumps since 2013, and serves as a reasonable proxy for all useful English web text from Common Crawl.

# E: Data Tokenisation

Data Collection and Preprocessing: LLMs are trained on MASSIVE datasets of text, media and code. This data undergoes cleaning, formatting, and potentially annotation (labeling for supervised learning) and is represented by hundreds of millions of "*tokens*". As an example of scale … the Oxford English Dictionary contains around 600,000 words and tokenizers produce very broadly 1.3 tokens per word, 30 per sentence and 100 per paragraph.

Given a choice and design of Model Architecture, which involves choosing the number of layers, attention mechanisms, and other parameters, the process entails further sub-phases of Pre-training, Fine-tuning, and Evaluation and Optimization each of which apple various numerical algorithms to tens or hundreds of billion of "*parameters*" organized into small collections known as "vectors"……*For a "frontier" LLM this Training Phase can take 3 or more months of non-stop massive computing.*




In simple terms: Imagine tokens as the LEGO bricks used to build a structure, and parameters as the instructions that tell you how to connect those bricks to create a specific shape or design. The model uses the parameters to understand how to arrange the tokens (bricks) to generate meaningful text.

# F: Training Data Resources for LLM

How much data

- Pleias' "Common Corpus" is the largest open and permissibly licensed dataset for training LLM's – and comprises over **2 Trillion tokens.**
- Common Corpus covers multiple languages (see top 10 below) and is composed of five main sub-corpora: OpenGovernment, OpenCulture, OpenScience, OpenWeb, and OpenSource (i.e. GitHub source code). Each of these is distinct in style, content, and quality and comes from varied sources. The result is a diverse dataset suitable for training a general-purpose model.

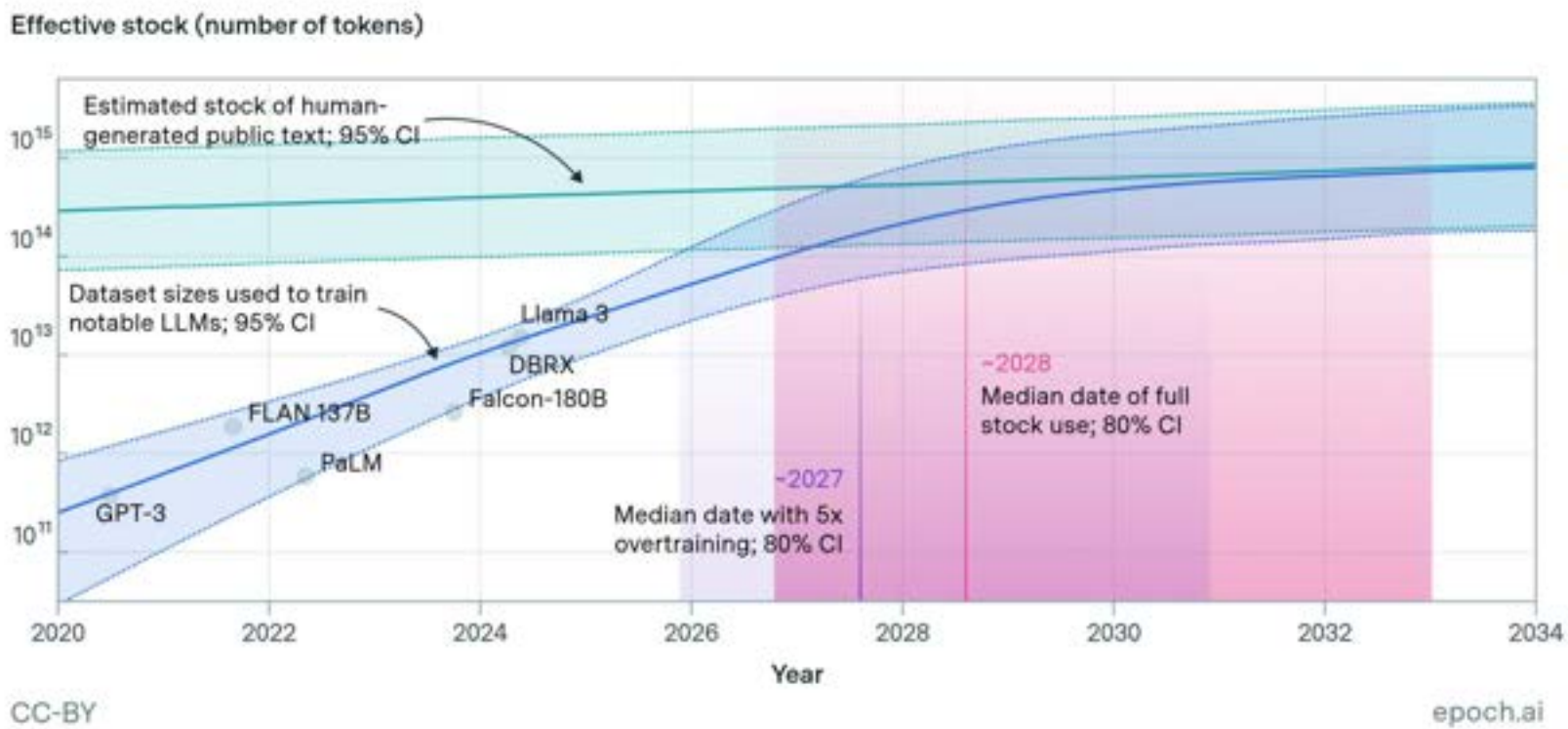
| Language | Percent | Tokens |
|---|---|---|
| English | 40.4 | 808 B |
| French | 15.8 | 316 B |
| German | 8.06 | 161.2 B |
| Spanish | 2.57 | 51.4 B |
| Latin | 2.27 | 45.4 B |
| Italian | 1.44 | 28.8 B |
| Dutch | 1.42 | 28.4 B |
| Portuguese | 1.15 | 23 B |
| Greek | 0.89 | 17.8 B |
| Polish | 0.78 | 15.6 B |

| Collection | Domain | Sources | Tokens |
|---|---|---|---|
| OpenGovernment | legal and administrative | Finance Commons (e.g. SEC, WTO) and Legal Commons (e.g. Europarl, Caselaw Access Project) | 468 B |
| OpenCulture | cultural heritage | public domain books and newspapers, wikisource | 1 T |
| OpenScience | academic | Open Alex, French theses | 98 B |
| OpenWeb | web text | YouTube Commons, Stack Exchange | 48 B |
| OpenSource | code | GitHub | 378 B |

https://thealliance.ai/blog/pleias-releases-common-corpus-open-multilingual-dataset-for-llm-training

## G: Will we run out of human-generated data ?

An estimate of the effective stock of quality and repetition adjusted **human-generated public text for AI training at around 300 trillion tokens**. If trends continue, language models will fully utilize this stock between 2026 and 2032, or even earlier if intensely overtrained.



# H: Crash Course in Size Acronyms – large and larger

| Acronym (Symbol) | Name | Value | Name of Value (Short Scale) |
|---|---|---|---|
| k | kilo | $10^3$ (1,000) | Thousand |
| M | mega | $10^6$ (1,000,000) | Million |
| G | giga | $10^9$ (1,000,000,000) | Billion |
| T | tera | $10^{12}$ (1,000,000,000,000) | Trillion |
| P | peta | $10^{15}$ | Quadrillion |
| E | exa | $10^{18}$ | Quintillion |
| Z | zetta | $10^{21}$ | Sextillion |
| Y | yotta | $10^{24}$ | Septillion |

- The prefixes for multiples of 1,000 and greater generally use **uppercase** symbols (e.g., M, G, T), with the exception of **kilo (k)**, hecto (h), and deka (da).
- These prefixes are attached directly to the unit symbol without a space (e.g., km for kilometer, MW for megawatt).
- While these are the official SI prefixes, in some specific fields (like computing), the capital letter abbreviations K, M, G, T might informally represent powers of two (specifically 1,024, 1,048,576, etc.) instead of powers of ten. However, the official standard recommends using the new binary prefixes (KiB, MiB, GiB, etc.) for powers of two to avoid confusion.

# I: Crash Course in Size Acronyms – small and smaller

| Prefix Name | Symbol (Acronym) | Factor (Scientific Notation) | Meaning (Fraction/Decimal) |
|---|---|---|---|
| deci | d | $10^{-1}$ | One tenth (0.1) |
| centi | c | $10^{-2}$ | One hundredth (0.01) |
| milli | m | $10^{-3}$ | One thousandth (0.001) |
| micro | μ (or u in plain text) | $10^{-6}$ | One millionth (0.000001) |
| nano | n | $10^{-9}$ | One billionth (0.000000001) |
| pico | p | $10^{-12}$ | One trillionth (0.000000000001) |
| femto | f | $10^{-15}$ | One quadrillionth |
| atto | a | $10^{-18}$ | One quintillionth |
| zepto | z | $10^{-21}$ | One sextillionth |
| yocto | y | $10^{-24}$ | One septillionth |
| ronto | r | $10^{-27}$ | One octillionth |
| quecto | q | $10^{-30}$ | One nonillionth |

- **Symbols are case-sensitive**: For example, `m` stands for milli ($10^{-3}$), while `M` stands for Mega ($10^{6}$).
- **The micro symbol** is the Greek letter mu ( `μ` ).
- **Prefixes are never combined** (e.g., use `nm` for nanometer, not `mμm` for millimicrometer).
- **The kilogram (kg)** is a unique base unit that already contains a prefix ( `k` for kilo, meaning 1000); for mass measurements, prefixes are attached to the gram (g), not the kilogram (e.g., milligram, `mg` , instead of microkilogram, `μkg` ).