Fragments.

Nearly Free, Hardly Cheap

Appendix & Sources — methodology, key concepts, glossary, and references

SIDHARTH RATH · THEFRAGMENTS.IN · COMPILED 2 JUNE 2026

This essay is data-driven, and several of its claims live inside interactive figures. This appendix records exactly how each figure was built — what is **real and sourced** versus **illustrative and clearly labelled** — defines the key terms, and lists every reference. The guiding principle, matching the essay's voice, is quantified honesty: numbers you can check, and assumptions you can see.

1 · Methodology — how each figure was built

"What actually happens when AI answers" ILLUSTRATIVE

A simplified, hand-built teaching demo — **not live model output**. The token splits, attention weights, and next-token probabilities are curated for a handful of example prompts to convey the mechanism. Temperature reshapes a fixed set of logits via a standard softmax; "small vs large model" scales those logits to illustrate confidence. Mechanism follows the Transformer architecture (Vaswani et al., 2017); next-token prediction is the training objective of GPT-family models.

"Who's actually ahead? Pick a test" CURATED · DATED

A snapshot in the style of the **Artificial Analysis** leaderboard (June 2026). Scores are illustrative-but-plausible and dated; frontier rankings shift week to week, so the durable point is the *pattern* — models bunch together on saturated tests (e.g. MMLU-Pro) and spread apart on hard ones (e.g. GPQA Diamond, SWE-bench). Model names follow the essay's near-future framing. "Price ($/1M)" is a lower-is-better axis included to tie capability back to cost.

"Cheaper than a person? Depends where you live" LIVE FX + SOURCED + ESTIMATED

Human pay is anchored to **US BLS Occupational Employment & Wage Statistics** (May 2024 median wages), converted to a monthly figure, then scaled to each market by an **average-wage proxy relative to the US** (Numbeo, 2025) — a wage-*level* approximation, not a per-role local salary. AI cost = tokens × a representative **~$1.50 / 1M tokens** blended price for a capable 2026 model. **Tokens-per-task and the agentic-intensity multiplier are illustrative estimates**, chosen to span a single naive call through a deep agent loop. **Exchange rates are live** (open.er-api.com, fetched server-side and cached daily), with a dated fallback. The takeaway is structural and robust to the exact numbers: the AI price is one global dollar figure, while local wages are not — so AI's advantage over a worker shrinks, and can invert, in lower-income markets.

"Augment or automate? It depends on the job" CURATED · DATED

A snapshot in the style of the **Anthropic Economic Index** (early 2026). For each sector, "augment" is the share of AI use where a person and the model work together; the remainder is "automate," where the model does the task alone. Per-sector splits and shares of total AI use are approximate. The overall ~57% augmentation figure is drawn from the Index.

2 · Key concepts

Inference price vs the price of a capability. A model's sticker price (dollars per million tokens) is not the same as what it costs to reach a fixed level of performance. The latter has fallen far faster — by 9× to 900× per year, depending on the task (Epoch AI).

Agentic workflows and tokens-per-task. When a model is asked to do real, multi-step work autonomously, it consumes far more tokens per task than a single question. A cheaper per-token price can still produce a larger bill if consumption grows faster — the paradox the essay is named for.

Purchasing power and the dollar. Frontier models are priced in US dollars and hosted in the US. For a non-dollar earner the real cost is the sticker price × an exchange rate, measured against local wages — which is why "getting cheaper" is true mainly for those who earn in dollars.

Augmentation vs automation. Augmentation is a person and a model working together; automation is the model doing a task alone. The mix, not a single headline, is what "AI and work" actually looks like.

Benchmark saturation. When a test gets so easy that top models all score near the ceiling (e.g. MMLU near 90%), it stops separating them — pushing evaluators toward harder tests where real gaps remain.

3 · Glossary

Token

A chunk of text a model reads or writes — often a word-fragment ("token" → "tok" + "en"). Pricing and consumption are measured in tokens; ~1M tokens ≈ 750,000 words.

Embedding / vector
The list of numbers a token is turned into, placing its meaning in space so related words sit near each other.

Attention
The mechanism by which a model weighs which earlier tokens matter for predicting the next one — the core idea of the 2017 Transformer.

Temperature
A setting that flattens or sharpens the next-token probabilities: low = safe and repetitive, high = varied and "creative."

MMLU / MMLU-Pro
A broad multiple-choice knowledge test; now largely "saturated" at the top.

GPQA Diamond
PhD-level science questions experts answer correctly only ~65% of the time — a genuine reasoning test.

SWE-bench
Real software bugs a model must fix, with the fix verified against tests.

AIME
A hard maths-olympiad exam used to test multi-step reasoning.

Intelligence Index
A composite that blends several hard evaluations into one overall score (Artificial Analysis).

PPP (purchasing power parity)
A conversion factor reflecting what a unit of currency actually buys locally, distinct from the market exchange rate.

OEWS
The US Bureau of Labor Statistics' Occupational Employment & Wage Statistics — the source of the US wage anchors.

4 · Caveats to read honestly

> Benchmark model names and scores drift weekly; the figure is timestamped "June 2026," and the durable point is the clustering pattern, not any exact ranking.

> The per-country wage proxy (Numbeo, 2025) is a wage-*level* approximation relative to the US, not a surveyed local salary for each specific role.

> Tokens-per-task figures in the cost interactive are informed estimates, not measured production traces; they are meant to be moved, not trusted to the decimal.

> Exchange rates are live as of page load; purchasing-power factors are World Bank 2024 figures. Re-pull both before relying on the numbers at a later date.

> Model names (e.g. “Claude Opus 4.8,” “GPT-5.5”) follow the essay’s near-future framing and are illustrative placeholders for the 2026 frontier.

5 · References

1. **Epoch AI**, “LLM inference price trends.” Price to reach a fixed capability fell 9×–900× per year by task; ~200×/yr median after January 2024.
2. **TechCrunch** (30 May 2026) on GitHub Copilot’s move to token-based billing (31 May 2026); developers reported 10–50× cost increases for agentic workflows; sign-ups paused; Opus-class models removed from the Pro tier.
3. **AI “winters”**: c. 1974–1980 (DARPA funding withdrawal; UK 1973 Lighthill Report) and late 1980s (collapse of expert systems / LISP-machine market). Standard AI history (Wikipedia, “AI winter”; ACM Communications).
4. **Vaswani et al.**, “Attention Is All You Need,” arXiv:1706.03762 (12 June 2017); NeurIPS 2017; 170,000+ citations as of 2025.
5. **Google Brain & DeepMind** merged to form Google DeepMind, April 2023.
6. Next-token prediction as the core training objective of GPT-style LLMs, built on the Transformer architecture (Vaswani et al., 2017).
7. OpenAI, Anthropic (founded by former OpenAI staff), and Google DeepMind all build on the 2017 Transformer.
8. **Artificial Analysis Intelligence Index** & **Stanford HAI 2026 AI Index**: top models cluster within ~7–8 points; MMLU saturated near 89–92%, pushing evaluators toward GPQA Diamond and Humanity’s Last Exam.
9. **Epoch AI**: OpenAI frontier input price fell from $30 / 1M tokens (GPT-4, March 2023) to a few dollars or less for GPT-4-class capability by 2026.
10. **TechCrunch** (30 May 2026): developer projections of monthly costs jumping from tens to hundreds of dollars; agentic sessions consume far more tokens than a single question.
11. **USD/INR** ≈ 95 in early June 2026, forecast ~100 by October 2026, down ~11% over twelve months (Trading Economics; US Federal Reserve H.10).
12. **DeepSeek** data-storage disclosures and subsequent bans/restrictions across Italy, Australia, India, South Korea, and 17+ US states (IAPP; US House Select Committee on the CCP).
13. **Anthropic Economic Index** (March 2026): ~57% augmentation vs ~43% automation; “deskilling” — average task handled needs ~14.4 years of education vs 13.2 for the average task in the economy.
14. **Challenger, Gray & Christmas** (December 2025): AI cited in 54,836 cuts (~5% of the 2025 total); total 2025 cuts 1,206,374, up 58% year over year.
15. **World Economic Forum**, Future of Jobs Report 2025: 40% of employers anticipate AI-driven workforce reduction; 170M roles created / 92M displaced / net +78M by 2030; 77% plan to reskill.
16. **Gartner** projects a broadly neutral near-term net employment impact through 2026; corroborated by Goldman Sachs, the US BLS, and the IMF’s automatable/augmentable task taxonomy.

Data sources used in the interactive figures

- **US wages:** US BLS Occupational Employment & Wage Statistics (OEWS), May 2024 medians.
- **Per-country wage levels:** Numbeo average net salary rankings, 2025 (wage-level proxy).
- **Purchasing power parity:** World Bank, PA.NUS.PPP, 2024.
- **Live exchange rates:** open.er-api.com (cached daily).

- **Benchmarks:** Artificial Analysis leaderboard style, June 2026 (curated snapshot).
- **Work patterns:** Anthropic Economic Index style, early 2026 (curated snapshot).
