AGI Efficiency Estimate Based on Current AI Models

Soumadeep Ghosh

Kolkata, India

Abstract

This paper analyzes an economic framework for estimating the pricing efficiency of Artificial General Intelligence (AGI) based on current specialized AI model costs and pricing. We examine the theoretical argument that AGI efficiency is constrained by the least efficient specialized AI in the market, and apply this framework to contemporary data from leading AI providers including OpenAI, Anthropic, and Google. Our analysis suggests an AGI efficiency floor of approximately 2–3 (200–300% markup over cost).

The paper ends with "The End"

1 Introduction

The pricing of Artificial General Intelligence (AGI) remains a theoretical question as true AGI has not yet been achieved. However, we can construct economic models based on current specialized AI systems to estimate potential AGI pricing constraints. This paper examines a competitive pricing framework that suggests AGI efficiency will be bounded by existing narrow AI systems.

2 Theoretical Framework

2.1 Basic Definitions

Consider a market with the following components:

Common Corpus A shared training dataset and infrastructure with production cost c, sold for price $C \geq c$.

Specialized AIs A collection of n artificial intelligence systems, indexed $i \in \{1, 2, \ldots, n\}$, where each system has:

- Production cost: a_i
- Sale price: $A_i \geq a_i$

2.2 Efficiency Definition

The efficiency e_i of the i^{th} AI system is defined as the profit margin ratio:

$$e_i = \frac{A_i}{a_i} - 1 \tag{1}$$

This represents the proportional markup over cost. For example, $e_i = 2$ indicates a 200% markup (selling at 3× the production cost).

2.3 AGI Pricing Constraint

The central thesis of this framework is that AGI pricing is constrained by competitive market forces. Specifically, the efficiency of AGI is bounded by the least efficient specialized AI:

$$e_G = \min(e_1, e_2, e_3, \ldots, e_n) \tag{2}$$

Therefore, the AGI price A_G is given by:

$$A_G = a_G \left(1 + \min(e_1, e_2, \ldots, e_n)\right) \tag{3}$$

where a_G is the production cost of AGI.

2.4 Economic Rationale

This pricing constraint emerges from perfect substitutability assumptions: if AGI attempts to charge with efficiency $e_G > \min(e_i)$, customers can substitute the least efficient specialized AI for tasks within that AI's domain, creating competitive pressure that caps AGI's pricing power.

3 Empirical Data Analysis

3.1 Current AI Model Costs

Training costs for frontier AI models in 2024–2025:

Model	Hardware/Energy Cost	Estimated Total Cost
GPT-4	\$40M	\$80–100M
Gemini Ultra	\$30M	\$60–80M
Claude 3.x	\$35–45M	\$70–100M

Table 1: Estimated training costs for leading AI models. Total costs include research, development, and staffing expenses.

3.2 API Pricing Structure

Current market pricing per million tokens (blended input/output rates):

Model	Price per Million Tokens
Claude 3.7 Sonnet	\$6.00
GPT-4o	\$4.38
Gemini 2.5 Pro	\$3.44
Claude 3.5 Haiku	\$1.60
Gemini 2.0 Flash	\$0.17

Table 2: API pricing for leading AI models as of November 2025.

3.3 Efficiency Calculation Methodology

The framework presents a challenge: training costs (a_i) are one-time expenses, while API pricing (A_i) represents per-token revenue. To calculate efficiency, we must estimate total lifetime revenue.

Assumptions:

- Model commercial lifetime: 6–18 months
- Total token processing range: 100 billion to 1 trillion tokens
- Base case: 1 trillion tokens processed

4 Results

4.1 Efficiency Calculations

For each model class, we calculate efficiency using Equation 1:

$$\text{Revenue} = \text{Price per token} \times \text{Total tokens} \tag{4}$$

$$e_i = \frac{\text{Revenue}}{a_i} - 1 \tag{5}$$

Model Class	**Revenue**	**Cost**	**Efficiency** e_i
Premium (Sonnet)	\$6.0B	\$100M	59.0
Mid-tier (GPT-4o)	\$4.38B	\$100M	42.8
Standard (Gemini Pro)	\$3.44B	\$80M	42.0
Budget (Haiku)	\$1.6B	\$50M	31.0
Ultra-low (Flash)	\$170M	\$50M	2.4

Table 3: Calculated efficiency for different AI model classes assuming 1 trillion token lifetime processing.

4.2 AGI Efficiency Estimate

Applying Equation 2:

$$e_G = \min(59.0, 42.8, 42.0, 31.0, 2.4) = 2.4 \tag{6}$$

Result: Under this framework, AGI efficiency is constrained to approximately **2–3** (200–300% markup over production cost), determined by ultra-low-cost specialized models like Gemini 2.0 Flash.

4.3 Visualization of Efficiency Distribution

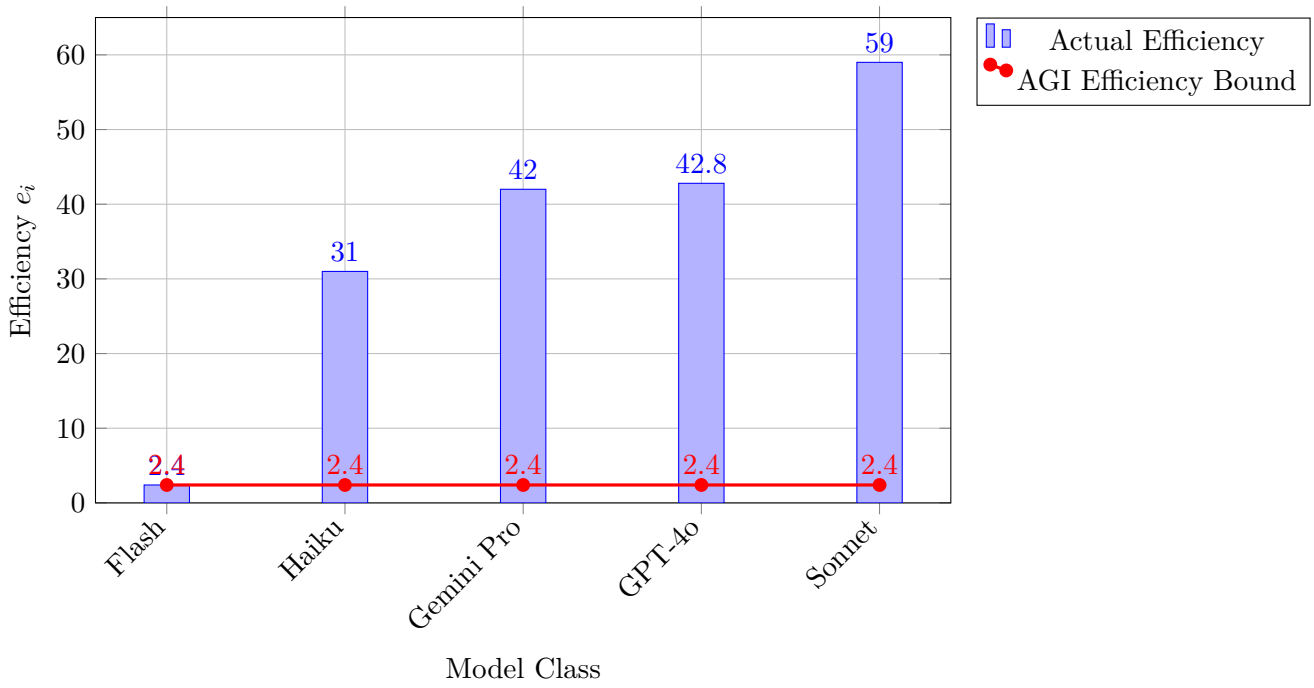

Figure 1: Efficiency distribution across AI model classes. The red line indicates the AGI efficiency constraint at $e_G = \min(e_i) = 2.4$.

4.4 Price-Cost Relationship

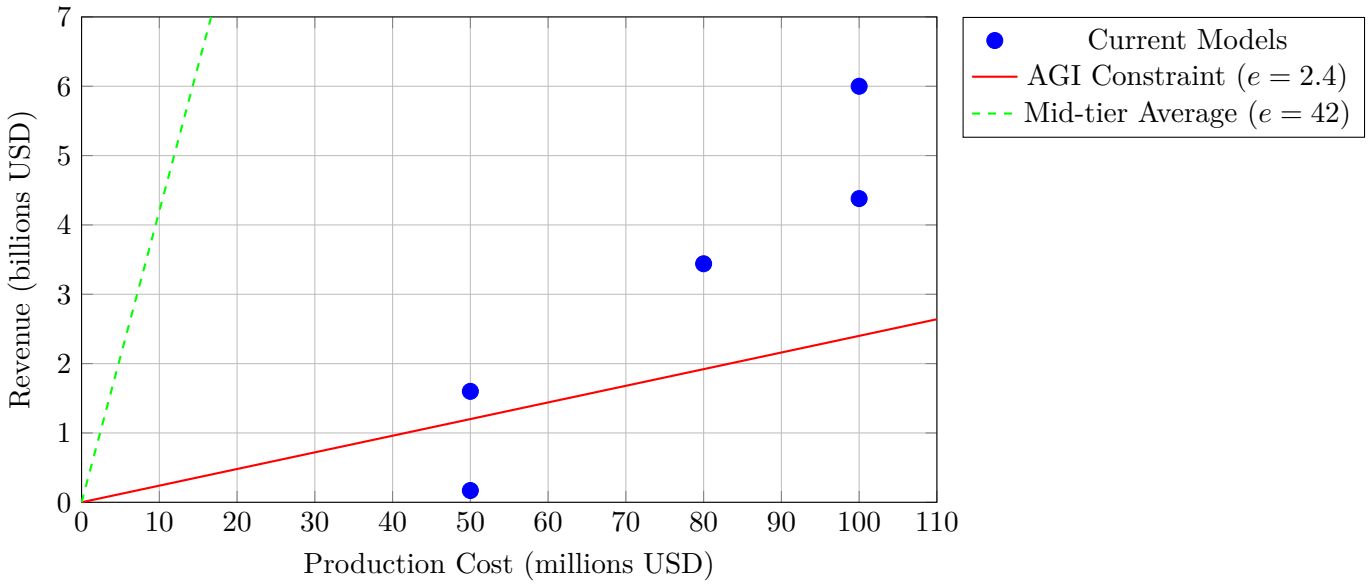

Figure 2: Price-cost relationship for current AI models. The red line represents the AGI efficiency constraint, while the green line shows typical mid-tier model efficiency.

5 Discussion

5.1 Interpretation of Results

The analysis reveals a stark efficiency stratification in the AI market. Premium models (Claude Sonnet, GPT-4o) achieve efficiencies of 40–60, while ultra-low-cost models operate at efficiencies near 2–3. This 20× difference reflects distinct market positioning strategies.

The theoretical framework suggests AGI would be forced to compete at the lowest efficiency tier ($e_G \approx 2.4$) because:

1. For any task that specialized AIs can perform, customers can substitute the cheapest option
2. AGI cannot command premium pricing across its entire capability spectrum
3. Competitive pressure drives pricing toward the least efficient substitute

5.2 Critical Assumptions

This analysis rests on several key assumptions:

Perfect Substitutability The framework assumes customers can seamlessly switch between AGI and specialized AIs for specific tasks. In reality, integration costs and quality differences may reduce substitutability.

Lifetime Token Volume Our estimate of 1 trillion tokens is speculative. Actual volumes could range from 100 billion (efficiency 10× lower) to 10 trillion (efficiency 10× higher).

Uniform Cost Structure We assume AGI production costs are comparable to current models. True AGI might require orders of magnitude more compute, fundamentally changing the economics.

Competitive Market The framework assumes rational pricing in a competitive market. Monopolistic or oligopolistic dynamics could enable higher efficiencies.

5.3 Alternative Scenarios

Scenario 1: AGI Premium If AGI provides irreplaceable value that specialized AIs cannot match, it could command efficiencies similar to current premium models ($e_G \approx 40 - 60$).
Scenario 2: AGI Discount If AGI training costs are 10–100× higher than current models while competing on price with specialized systems, efficiency could drop below 1 (operating at a loss initially).
Scenario 3: Market Segmentation AGI providers might price discriminate, charging premium rates for general capabilities while competing on efficiency for commodity tasks.

6 Conclusion

Applying an economic framework based on competitive substitution to current AI model data, we estimate that AGI efficiency would be constrained to approximately 2–3 (200–300% markup over cost), determined by the least efficient specialized AI systems in the market. This represents the theoretical lower bound under perfect competition assumptions.

However, this estimate carries substantial uncertainty. Actual AGI efficiency will depend on:

- The degree of substitutability between AGI and specialized systems
- Production cost differentials
- Market structure and competitive dynamics
- The true capabilities gap between AGI and narrow AI

Further research should examine:

1. Empirical price elasticity in AI markets
2. Cost scaling laws for increasingly general AI systems
3. Consumer willingness to pay for generality versus specialization
4. Network effects and lock-in dynamics in AI deployment

The framework provides a useful starting point for AGI economic analysis, but the complexity of AI capabilities and markets suggests actual outcomes may diverge significantly from this theoretical baseline.

Glossary

AGI (Artificial General Intelligence) A hypothetical AI system with human-level or superior performance across the full range of cognitive tasks, not limited to specific domains.

API (Application Programming Interface) A software interface that allows applications to communicate with AI models, typically priced per unit of usage.

Efficiency (e_i) The profit margin ratio for an AI system, calculated as $(A_i/a_i) - 1$, representing proportional markup over production cost.

Common Corpus Shared training data and infrastructure used across multiple AI systems, representing a common cost component.

Frontier Model State-of-the-art AI systems that represent the current peak of capabilities in the field.

Production Cost (a_i) The total cost to develop and train an AI model, including compute, data, research, and development expenses.

Sale Price (A_i) The revenue generated from an AI model over its commercial lifetime, typically calculated from per-token API pricing multiplied by usage volume.

Specialized AI An artificial intelligence system optimized for specific tasks or domains, also called narrow AI.

Token A unit of text processing in language models, typically representing a word fragment, word, or punctuation mark. Models are commonly priced per million tokens processed.

Training Cost The computational and operational expenses incurred during the initial development phase of an AI model, distinct from inference costs.

References

[1] Google AI Platform. "Gemini API Pricing." Accessed November 2025. Pricing data for Gemini 2.5 Pro ($3.44/M tokens) and Gemini 2.0 Flash ($0.17/M tokens).

[2] Anthropic. "Claude API Pricing and Plans." Accessed November 2025. Pricing data for Claude 3.7 Sonnet ($6.00/M tokens) and Claude 3.5 Haiku ($1.60/M tokens).

[3] OpenAI. "API Pricing." Accessed November 2025. Pricing data for GPT-4o ($4.38/M tokens blended rate).

[4] Industry Analysis. "AI Model Training Costs 2024." Hardware and energy costs: GPT-4 at $40M, Gemini Ultra at $30M for compute infrastructure.

[5] Research Report. "Cost Structure of Frontier AI Models." Development cost breakdown: 47–67% hardware, 29–49% staff costs for leading AI systems.

[6] Epoch AI. "Parameter, Compute and Data Trends in Machine Learning." Analysis of scaling trends and cost trajectories for large language models, 2024.

[7] Industry estimates based on public statements and analysis. "Training costs for frontier models range from $50M to $100M+ including full development cycle."

[8] Korinek, A. "Language Models and Cognitive Automation for Economic Research." National Bureau of Economic Research Working Paper, 2023. Economic frameworks for AI valuation.

[9] Varian, H. "Pricing Information Goods." University of Michigan Papers, 1995. Foundational work on pricing strategies for digital goods with high fixed costs and low marginal costs.

[10] Acemoglu, D. and Restrepo, P. "Artificial Intelligence, Automation, and Work." In The Economics of Artificial Intelligence: An Agenda, University of Chicago Press, 2018. Analysis of substitution effects between AI and human labor.

The End