PROVIDER SELECTION & EVALUATION REPORT: WEB SEARCH

Executive Summary

The benchmarking results indicate that **Gemini 2.5 Pro** and **Gemini 2.5 Flash** are the top-performing models across almost all metrics, particularly when paired with the **Brave** search provider. While the **GPT-4 series** (notably GPT-4.1) remains highly competitive in relevance and usefulness, the Gemini models demonstrate superior freshness and coverage scores.

1. Best Performing Models

The models were evaluated on relevance, freshness, quality, usefulness, and coverage.

- **Overall Performance: Gemini 2.5 Pro:** Achieved a near-perfect overall score of **9.8 to 10.0** in advanced tech search scenarios. It consistently provides the highest depth of information and source coverage.
    - **Gemini 2.5 Flash:** The most efficient high-performer, maintaining a perfect **10.0 overall score** in basic searches while being significantly faster than Pro versions.
- **Strong Alternatives:**
    - **GPT-4.1:** The strongest performer within the OpenAI suite, often reaching an overall score of **9.4 to 9.6**. It is particularly noted for high "usefulness" and "quality" ratings.
    - **GPT-4o Mini:** Offers the best balance of speed and reliability for standard queries, consistently scoring around **8.6**.

2. Provider Benchmark: Tavily vs. Brave

The choice of search provider significantly impacts the quality of the model's output.

- **Brave (Recommended for Quality):** Models using Brave as a provider consistently reached the highest possible scores (10.0).
    - Brave-powered searches resulted in higher **freshness scores** compared to Tavily, making it superior for fast-moving "tech" breakthrough queries.
- **Tavily (Strong for Research):** Tavily performs exceptionally well in "advanced" depth modes, where it helps models like Gemini 2.5 Pro achieve higher coverage.
    - However, it showed occasional volatility; for instance, Gemini 2.5 Pro's quality dropped significantly in one specific Tavily test case (overall score 0.0), suggesting potential integration edge cases.

3. Detailed Performance Metrics Analysis

Metric	Leader	Finding
Relevance	Gemini 2.5 Pro / GPT-4.1	Both models consistently hit scores of 10/10.
Freshness	Gemini 2.5 series	Superior ability to capture developments from January 2026.
Search Time	GPT-4o Mini / Gemini Flash	Consistently lower search latency (~0.3s - 0.6s).
Coverage	Gemini 2.5 Pro	Best at synthesizing information from a wide variety of URLs.

4. Recommendations

1. **For Maximum Accuracy & Research Depth:** Use **Gemini 2.5 Pro** with the **Brave** provider. This configuration maximizes freshness and source coverage.
2. **For Speed & Efficiency:** Use **Gemini 2.5 Flash** or **GPT-4o Mini**. These models provide professional-grade results (Scores >8.5) with minimal latency.
3. **For General Usefulness: GPT-4.1** is the most reliable model if your primary goal is high-quality, actionable summaries.

5. Final Strategic Recommendations

To ensure long-term stability and peak performance, the following roadmap is recommended:

- **Phase-out OpenAI Series**: Given that the GPT-4 series is retiring in February 2026, standardizing on the Gemini ecosystem is recommended to avoid service interruptions.
- **Standardize on Gemini 2.5 Suite**: Proceed with **Gemini 2.5 Pro** and **Flash** as the primary models.
- **Implementation of Dynamic Model Routing**: Use an "Auto-Choosing" model logic based on the specific situation:
    - **Gemini 2.5 Flash** for high-speed, standard queries where latency is critical.
    - **Gemini 2.5 Pro** for complex, advanced searches requiring maximum research depth and coverage.
- **Provider Optimization**: Utilize the **Brave** provider for all time-sensitive tech queries to maximize freshness and accuracy.

View live interactive dashboard: https://palanisuhas.github.io/benchmark_analysis/visualizations/