General Comparison Report: AI Model Evaluation

This report summarizes observations from benchmark tests across two models:

MoonshotAI Kimi-K2 (free) and **OpenAI GPT-OSS-20B (free)**

1. Latency

- **Kimi-K2** demonstrated **faster responses**, averaging ~4.2s per request.
- **GPT-OSS-20B** was **slower**, averaging ~5.6s per request.
- Latency differences were consistent across tasks, with Kimi generally 20–25% quicker.

2. JSON Reliability

- **Kimi-K2** achieved a **75% JSON compliance rate**, meaning most responses adhered to strict JSON requirements.
- **GPT-OSS-20B** lagged significantly at **25% compliance**, often producing malformed or extra text around JSON.
- This makes Kimi more suitable for structured automation tasks.

3. Translation Fidelity

- **Kimi-K2** maintained simpler, more literal translations with fewer deviations.
- **GPT-OSS-20B** tended to generate longer completions (~1036 chars vs. ~210), sometimes embellishing content, which reduced faithfulness to the original.

4. Style Control

- **Kimi-K2** respected temperature and prompt instructions more consistently, leading to more controlled outputs.
- **GPT-OSS-20B** often drifted stylistically, adding verbosity even under constrained settings.

5. Seed Stability

- Runs with identical seeds were more reproducible in **Kimi-K2**, showing minor variations.
- **GPT-OSS-20B** outputs fluctuated heavily despite seed settings, limiting deterministic use cases.

6. Rate Limits

- Both models in the free tier exhibited occasional throttling, but Kimi maintained steadier throughput.
- GPT-OSS-20B showed higher latency spikes, possibly due to stricter backend rate caps.

Summary Table

Model	Avg Latency (ms)	JSON Pass Rate	Avg Output Length
MoonshotAI Kimi-K2	**4163**	**75%**	210 chars
GPT-OSS-20B	5626	25%	1036 chars