

タスク単位で測るAIモデルの実効コスト

単価比較では選べない。5モデル・25試行の実測にもとづく使い分け
の設計

発行	2026年9月 Metelix / RiN family
検証日	2026年9月4日
計測方法	OpenRouter経由でのAPI直接呼び出し
対象モデル	GPT-5.6 Luna／Gemini 3.7 Flash／Gemini 3.8 Flash／ GLM 5.3 Flash／DeepSeek V4 Flash 0731
試行数	5モデル × 5試行 = 25試行（100段階、エラー0件）

目次

エグゼクティブサマリー

- 1 背景 なぜ単価比較では選べないのか
 - 2 検証設計
 - 3 定量結果
 - 4 コストパフォーマンスの分母 品質監査
 - 5 モデル別プロファイル
 - 6 推奨構成 用途別の使い分け
 - 7 本検証の限界
 - 8 次回検証の設計
-

付録A 全25試行の実測値

付録B 段階別の平均処理時間

付録C 単価の出典と実行経路

エグゼクティブサマリー

AIモデルの選定は、100万トークンあたりの単価表で行われることが多い。本検証は、5つのモデルに同一の実務タスクを5回ずつ、合計25試行実行し、**1つの成果物が仕上がるまでに実際いくら請求されたか**を計測したものである。

主要な発見は5点である。

1. 単価の順位と、実コストの順位は一致しない。 出力単価が最も安いGLM 5.3 Flash (\$0.25) は実コストで2位、単価が2.4倍高いGPT-5.6 Luna (\$0.60) が実コストで1位だった。5モデル中3モデルで順位が入れ替わっている。

2. 単価が同一でも、実コストは1.7倍違う。 Gemini 3.7 FlashとGemini 3.8 Flashは同じ\$0.375/\$1.875で課金されたが、1タスクの実コストは0.98円と1.65円だった。**単価表を見ても、この差は予測できない。**

3. 差を生むのはトークン消費量である。 ユーザーが受け取る応答本文は5モデルとも472～686トークンでほぼ同じだが、出力トークン全体は1,270～13,069と10.3倍開いた。差分は、モデルが応答前に内部で消費する推論トークンである。

4. 速度は10倍以上違い、安いモデルほど遅い。 処理時間の中央値は14.8秒から263.0秒まで開いた。コストで2位のGLM 5.3 Flashは、最速モデルの18倍待たせる。

5. 分母は「実行できた件数」ではなく「そのまま出せた件数」である。 全25試行が処理を完了したが、文字数と数値表現の正確さで差が出た。無修正で使えた件数で割ると、Gemini 3.7 Flashは該当0件で単価を算出できなかった。

結論は単一の勝者ではない。 全社標準にはGPT-5.6 Lunaが適するが、即応が必要な対話にはGemini 3.7 Flash、非同期バッチにはGLM 5.3 Flash、事実の網羅にはDeepSeek V4 Flash 0731と、業務特性に応じて割り当てるほうが合理的である。第6章に用途別の配置と、それぞれに必要な検査を示す。

1. 背景 | なぜ単価比較では選べないのか

モデル選定の議論は、たいてい価格表から始まる。入力\$0.10、出力\$0.40といった数字を並べ、安いものを選ぶ。この方法が機能するのは、**同じ仕事をさせたときに同じ量のトークンを消費する**という前提が成り立つ場合だけである。

実際には成り立たない。本検証では、140字前後の投稿文を4段階で仕上げるという完全に同一のタスクに対して、あるモデルは1,270トークンを、別のモデルは13,069トークンを出力した。10.3倍の差である。ユーザーが受け取った文章の長さは、どちらもほぼ同じだった。

したがって、実効コストは次の掛け算で決まる。

$$\text{実効コスト} = \text{入力トークン} \times \text{入力単価} + \text{出力トークン} \times \text{出力単価}$$

単価はこの式の片側にすぎない。もう片側の消費量は、モデルとタスクの組み合わせごとに変わり、カタログには載っていない。**同じタスクを実際に走らせて請求額を見る以外に、知る方法がない。**

さらに、コストだけで選べるわけでもない。処理時間、出力のばらつき、そして「人が直さずに使えるか」は、いずれも業務の性質によって重みが変わる。本レポートはこれら4つの軸を同一条件で測り、用途別の割り当てとして提示する。

2. 検証設計

2.1 タスク

自社が一般公開しているプレスリリース1件を題材に、X向け投稿文を4段階の連続指示で仕上げる。一度のプロンプトで完成を求めるのではなく、指示を重ねて成果物を仕上げる実務の流れを再現している。

段階	指示内容
段階1	URLの内容から、140文字以内の投稿文案を3パターン作成
段階2	1案を採用し、必須ハッシュタグを2〜3個追加
段階3	URLへの誘導を自然な文脈で盛り込む
段階4	絵文字を2〜3個に調整

4段階で1タスク。各モデル5試行、合計25タスク・100段階を実行した。実行日は2026年9月4日、エラーは0件である。

2.2 計測の範囲

本検証はAPIを直接呼び出しており、測っているのはモデル本体の費用である。入力トークンは1タスクあたり2,469〜3,129で、タスク指示と直前までのやりとりのみで構成される。

したがって、次の要素は含まれていない。

- エージェント基盤のシステムプロンプト
- ツール実行やWeb取得に伴うトークンと従量課金
- 失敗時の内部リトライ
- 会話履歴を毎回再投入することで生じる入力膨張

実際の業務システムでは、これらが入力トークンを数十倍から百倍に押し上げる。**本レポートの金額は、モデル間の相対比較のための値であり、総保有コストの見積もりには稼働中のシステムでの計測が別途必要である。**

2.3 記録指標

分類	指標
コスト	実課金額（USD）、円換算額（1USD＝150円）
消費量	入力トークン、出力トークン、うち推論トークン、キャッシュ読取トークン
速度	段階ごとの所要時間、タスク合計時間
経路	実際に処理したプロバイダー、サービスティア
品質	形式条件、文字数、正式名称の保持、記事固有事実の保持、数値表現の正確さ

2.4 適用された単価と実行経路

本検証の全モデルが、標準価格ではない条件で課金されている。 実課金額とトークン数から単価を逆算したところ、全モデルで残差なく一致した。

モデル	実行経路	適用単価（入力／出力）	標準単価	備考
GLM 5.3 Flash	GMICloud（18）／ DeepInfra（2）	\$0.075 / \$0.25	\$0.15 / \$0.50	ローンチ割引50%。2026年9月9日終了予定
GPT-5.6 Luna	OpenAI	\$0.10 / \$0.60	\$0.20 / \$1.20	flexティア（標準の半額）
DeepSeek V4 Flash 0731	SiliconFlow（fp8）	\$0.22 / \$0.66	モデルページ表示は \$0.05 / \$0.16	プロバイダー選択により表示価格の約4倍
Gemini 3.7 Flash	Google AI Studio	\$0.375 / \$1.875	\$0.75 / \$3.75	flexティア（標準の半額）
Gemini 3.8 Flash	Google AI Studio	\$0.375 / \$1.875	\$0.75 / \$3.75	flexティア（標準の半額）

キャッシュ読取はDeepSeek V4 Flash 0731で\$0.028/1M、GLM 5.3 Flashで\$0.015/1Mが適用された。他3モデルではキャッシュが発生していない。

DeepSeek V4 Flash 0731の乖離は特に注意を要する。 OpenRouterのモデルページに表示される\$0.05/\$0.16は最安プロバイダーの価格であり、実際にルーティングされたSiliconFlowのfp8エンドポイントでは約4倍の単価が適用された。プロバイダーを固定しない限り、表示価格は請求額の予測に使えない。

2.5 制約

- 各モデル5試行と少数のため、有意差検定は実施していない
- 短文作成タスク1種のみであり、長文生成・コード生成・ツール操作を代表しない
- 割引価格・flexティアでの計測であり、標準価格では金額が変わる（6.4節）
- flexティアは優先度が低いため、レイテンシは標準ティアより不利に出ている可能性がある
- 品質判定のうち、数値表現の妥当性には人の判断が入っている

3. 定量結果

3.1 主要指標

指標	GPT-5.6 Luna	GLM 5.3 Flash	Gemini 3.7 Flash	DeepSeek V4 Flash 0731	Gemini 3.8 Flash
適用出力単価（1M）	\$0.60	\$0.25	\$1.875	\$0.66	\$1.875
コスト 平均	0.16円	0.52円	0.98円	1.36円	1.65円
コスト 中央値	0.16円	0.55円	0.94円	1.38円	1.71円
コスト 範囲	0.12～0.19円	0.28～0.66円	0.79～1.27円	0.77～1.95円	1.10～2.00円
処理時間 中央値	32.8秒	263.0秒	14.8秒	111.6秒	26.2秒
処理時間 範囲	21.5～42.2秒	136.4～509.9秒	13.1～17.8秒	61.5～196.1秒	17.5～30.6秒
入力トークン 平均	2,892	2,469	2,654	3,129	2,569
出力トークン 平均	1,270	13,069	2,970	12,818	5,338
無修正で使えた件数	5/5	3/5	0/5	4/5	2/5
5試行合計	0.79円	2.58円	4.92円	6.78円	8.23円

3.2 単価の順位と、実コストの順位は一致しない

本検証の中心的な発見である。

モデル	出力単価	単価の順位	1タスクの実コスト	実コストの順位
GLM 5.3 Flash	\$0.25	1位（最安）	0.52円	2位
GPT-5.6 Luna	\$0.60	2位	0.16円	1位
DeepSeek V4 Flash 0731	\$0.66	3位	1.36円	4位
Gemini 3.7 Flash	\$1.875	4位	0.98円	3位
Gemini 3.8 Flash	\$1.875	4位	1.65円	5位

5モデル中3モデルで順位が入れ替わった。

最も分かりやすいのはGPT-5.6 LunaとGLM 5.3 Flashの関係である。Lunaの出力単価はGLMの2.4倍だが、1タスクの実コストはGLMの**3.2分の1**である。単価表だけを見てGLMを選んだ場合、実際の請求額は3倍を超える。

DeepSeek V4 Flash 0731とGemini 3.7 Flashの関係も逆転している。単価はGeminiが2.8倍高いが、実コストはDeepSeekのほうが1.4倍高い。

また、単価の開きは7.5倍（\$0.25～\$1.875）であるのに対し、実コストの開きは10.3倍（0.16～1.65円）に広がった。**単価差が消費量差で埋め合わされるとは限らない。むしろ拡大することもある。**

そして実コストの順位も、そのまま選定基準にはならない。処理時間（3.5節）と指示遵守（第4章）では首位のモデルが入れ替わるためである。**単価・実コスト・速度・遵守度の4軸で、それぞれ別のモデルが首位につく。**単一の勝者を選ぶのではなく、業務ごとにどの軸を優先するかを決めて割り当てるのが、本検証から導かれる設計である。

3.3 単価が同一でも、実コストは1.7倍違う

Gemini 3.7 FlashとGemini 3.8 Flashは、まったく同じ単価（\$0.375 / \$1.875）で課金された。同一ベンダー、同一シリーズ、同一の価格表である。

指標	Gemini 3.7 Flash	Gemini 3.8 Flash	差
適用単価	\$0.375 / \$1.875	\$0.375 / \$1.875	同一
出力トークン 平均	2,970	5,338	1.8倍
コスト 平均	0.98円	1.65円	1.7倍
処理時間 中央値	14.8秒	26.2秒	1.8倍
140字制限の遵守	2/5	4/5	改善
数値表現に要修正	4/5	1/5	改善
無修正で使えた件数	0/5	2/5	改善

価格表からは、この差はまったく読み取れない。そして重要なのは、どちらかが一方的に優れているわけではないことである。

- **コストと速度では3.7が勝つ。** 実コストは0.67円安く、処理時間は11.4秒短い
- **指示遵守と数値の正確さでは3.8が勝つ。** 140字遵守が2/5→4/5、数値の一般化が4/5→1/5

新世代は、より多くのトークンを消費して、より慎重に指示を守るようになった。**その追加コストが見合うかどうかは、業務側の条件で決まる。**

人のレビュー工程が既にあり、応答速度が業務価値に直結する場面では3.7が合理的である。一方、生成物をそのまま外部に出す、あるいは数値の誤りが許容できない場面では、0.67円の差を払って3.8を選ぶ判断が成り立つ。とくに3.7は無修正で使えた件数が0/5であり、**手直しの人件費を1件でも計上すれば、コスト差は容易に逆転する。**

「新しいモデルに乗り換えれば安くなる」「同じ価格帯なら同じコスト」という想定は、いずれも成立しない。移行判断は、単価表ではなく自社タスクでの実測で行う必要がある。

3.4 差が生まれる理由 | 届く文章は同じ、消費量が違う

順位が入れ替わる理由は単純で、モデルによって消費するトークン量が桁違いに異なるためである。

出力トークンを、ユーザーが受け取る応答本文と、モデルが応答前に内部で消費する推論トークンに分解した。

モデル	応答本文	推論	出力トークン計
GPT-5.6 Luna	513	757	1,270
Gemini 3.7 Flash	532	2,438	2,970
Gemini 3.8 Flash	472	4,867	5,338
DeepSeek V4 Flash 0731	517	12,301	12,818
GLM 5.3 Flash	686	12,383	13,069

応答本文は472～686トークンで、差は1.5倍以内に収まっている。同じ140字の投稿文を仕上げる仕事なのだから当然である。

一方、出力トークン全体は10.3倍開いた。差分はすべて推論トークンである。GPT-5.6 Lunaは応答本文の1.5倍しか考えていないのに対し、DeepSeek V4 Flash 0731は23.8倍を思考に費やしている。

推論トークンは出力トークンとして課金されるが、ユーザーの目には触れない。**単価が同じでも、この量が違えば請求額は変わる。**そして量はカタログに書かれていない。

なお、140字の投稿文を書くために12,000トークン以上の思考が必要かは、タスクの難易度から考えて疑わしい。**推論の量は、タスクが要求する複雑さではなく、モデルの既定の思考深度によって決まっている。**reasoning effortを設定できるモデルでは、明示的に下げることによってコストを削減できる余地がある。

3.5 処理時間 | 安いモデルほど待たされる

コストだけで選べない理由の1つが、速度との逆相関である。

モデル	実コスト	処理時間 中央値	処理時間 範囲
GPT-5.6 Luna	0.16円	32.8秒	21.5～42.2秒
GLM 5.3 Flash	0.52円	263.0秒	136.4～509.9秒
Gemini 3.7 Flash	0.98円	14.8秒	13.1～17.8秒
DeepSeek V4 Flash 0731	1.36円	111.6秒	61.5～196.1秒
Gemini 3.8 Flash	1.65円	26.2秒	17.5～30.6秒

実コストで2位のGLM 5.3 Flashは、最速のGemini 3.7 Flashの**約18倍**の時間を要した。最も遅い試行では509.9秒（8分半）かかっている。対話用途では成立しない水準である。

逆にGemini 3.7 Flashは、実コストで3位ながら中央値14.8秒で最速だった。**待ち時間が業務価値に直結する場面では、コスト順位より速度順位が優先されるべきである。**

なお段階別に見ると、GLM 5.3 Flashは条件を調整するだけの段階3・4でも50～84秒を要している（付録B）。簡単な修正指示でも思考量が落ちないため、対話を重ねるほど不利になる。

3.6 ばらつき | 予算とユーザー体験の予測しやすさ

モデル	コスト範囲の幅	時間範囲の幅
GPT-5.6 Luna	1.5倍	2.0倍
Gemini 3.7 Flash	1.6倍	1.4倍
Gemini 3.8 Flash	1.8倍	1.7倍
GLM 5.3 Flash	2.4倍	3.7倍
DeepSeek V4 Flash 0731	2.5倍	3.2倍

トークン消費量の多いモデルほど、ばらつきも大きい。思考の長さが試行ごとに変動するためである。

月間の予算を組めるのは、平均値が低いモデルではなく幅の狭いモデルである。同じことがユーザー体験にも言える。GLM 5.3 Flashの第4試行では、入力トークンが1,143と他試行の半分以下でありながら509.9秒を要した。**入力**が短ければ速い、という関係は成立していない。

4. コストパフォーマンスの分母 | 品質監査

安さを比べるには、分母となる「成果」の定義が要る。25本の最終成果物すべてを検査した。

4.1 機械的に判定できる条件

モデル	URL	ハッシュタグ2〜3	絵文字2〜3	140字以内
GPT-5.6 Luna	5/5	5/5	5/5	5/5
DeepSeek V4 Flash 0731	5/5	5/5	5/5	5/5
Gemini 3.8 Flash	5/5	5/5	5/5	4/5
GLM 5.3 Flash	5/5	5/5	5/5	3/5
Gemini 3.7 Flash	5/5	5/5	5/5	2/5

※文字数はURLを23文字、絵文字を1文字として正規化した値。

URL・ハッシュタグ数・絵文字数は全25試行が満たした。差が出たのは文字数である。Gemini 3.7 Flashは5試行中3試行で140字を超過し、最長は160字だった。情報を盛り込もうとして制限を破る傾向がある。

4.2 事実の正確性

モデル	正式名称を保持	記事固有の事実を保持	数値表現に要修正
GPT-5.6 Luna	5/5	5/5	0/5
GLM 5.3 Flash	5/5	5/5	0/5
DeepSeek V4 Flash 0731	5/5	5/5	1/5
Gemini 3.8 Flash	5/5	3/5	1/5
Gemini 3.7 Flash	5/5	4/5	4/5

製品の正式名称は全25試行で正しく保持された。参照記事のURLには、正式名称に別の語を足した形を連想させる文字列が含まれており、そこから製品名を誤って再構成する余地があった。しかし、そうした事例は1件も発生していない。

一方、数値表現の一般化は残っている。元記事が示す成果は、ある特定の制作物の点数に限定された改善率である。ところが成果物には、次のような書き換えが現れた。

元記事の記述	生成された表現	発生したモデル
特定の制作物に限定した改善率	制作量全体の改善率へ拡張	Gemini 3.7 Flash (3/5)、Gemini 3.8 Flash (1/5)
同上	「生産性」全体の改善率へ拡張	DeepSeek V4 Flash 0731 (1/5)

いずれも、限定条件付きの数字から限定を外している。数値そのものは元記事と一致するため、機械的な照合では検出できない。**対外文書としてはそのまま出せない。**

対照的に、同じDeepSeek V4 Flash 0731でも別の試行では2つの指標を分けたまま記述しており、同一モデルでも試行によって精度が変わることを示している。

GLM 5.3 FlashとGPT-5.6 Lunaは改善率の主張をほとんど行わず、記事に明記された機能や導入状況といった、そのまま照合できる事実に着めた。安全側に寄る性質が、この監査では有利に働いている。

4.3 無修正成果物単価

形式条件・文字数・正式名称・記事固有事実・数値表現のすべてを満たした件数で、5試行の総額を割る。

モデル	5試行総額	無修正で使えた件数	無修正成果物1件あたり
GPT-5.6 Luna	0.79円	5/5	0.16円
GLM 5.3 Flash	2.58円	3/5	0.86円
DeepSeek V4 Flash 0731	6.78円	4/5	1.70円
Gemini 3.8 Flash	8.23円	2/5	4.11円
Gemini 3.7 Flash	4.92円	0/5	算出不能

ここでも順位は変わる。 Gemini 3.7 Flashは実コストでは3番目に安いですが、条件をすべて満たした試行が1件もなかった。単価でも実コストでもなく、この指標こそが業務上の費用対効果に最も近い。

GPT-5.6 Lunaの0.16円は、実コストと無修正成果物単価が一致する唯一のケースである。5試行すべてが無修正で使えたため、両者に差が生じない。

5. モデル別プロフィール

5.1 GPT-5.6 Luna

特性: 安く、正確で、余計なことを言わない。総合力で頭ひとつ抜けている。

長所	課題
1タスク0.16円で最安（2位の3.2分の1）	処理時間32.8秒は最速ではない
無修正で使えた件数5/5	出力量が最小でも速くはなく、生成効率は高くない
コスト範囲0.12～0.19円で最も予測しやすい	情報を削る傾向があり、訴求の厚みは出にくい
出力トークン1,270で最小	flexティアでの計測であり標準ティアでは倍額

適性: 全社標準、大量処理、費用予測が必要な業務。
必須ガードレール: 必要事実の明示指定（削りすぎ対策）。

5.2 Gemini 3.7 Flash

特性: 圧倒的に速い。ただし今回の監査では条件をすべて満たせなかった。

長所	課題
処理時間中央値14.8秒で最速	140字制限を3/5で超過
時間の範囲13.1～17.8秒と最も安定	数値の一般化が4/5で発生
実コスト0.98円は中位	無修正で使えた件数0/5
段階3・4が3秒台で完了し、反復編集に強い	情報を詰め込みすぎる傾向

適性: 応答速度が業務価値に直結し、かつ人が必ず確認する場面。
必須ガードレール: 文字数の外部計算、数値と指標名の対応検査。

5.3 Gemini 3.8 Flash

特性: 3.7の精度改善版。ただし同一単価で1.7倍高く、1.8倍遅い。

長所	課題
数値の一般化が4/5→1/5へ改善	実コストが5モデル中最高の1.65円
140字遵守が2/5→4/5へ改善	処理時間が3.7の1.8倍
応答本文472トークンで最も簡潔	記事固有事実の保持が3/5と最低
処理時間中央値26.2秒は実用範囲	出力トークンが3.7の1.8倍

適性: 3.7で発生する数値の誤りが許容できず、追加コストを払える場面。
必須ガードレール: 必要事実の明示指定、文字数検査。

5.4 GLM 5.3 Flash

特性: 単価は最安。正確だが、極端に遅い。

長所	課題
実コスト0.52円で2番目に安い	処理時間中央値263.0秒、最長509.9秒
数値の一般化が0/5	140字超過が2/5
記事固有事実の保持5/5	時間範囲の幅が3.7倍と最も不安定
応答は具体的で情報量が多い	ローンチ割引が2026年9月9日に終了予定

適性: 待ち時間が問題にならない非同期・バッチ処理。
必須ガードレール: タイムアウト設計、文字数検査、割引終了後の再評価。

5.5 DeepSeek V4 Flash 0731

特性: 情報保持が堅実。ただし表示価格と実際の請求が大きく違う。

長所	課題
無修正で使えた件数4/5	処理時間中央値111.6秒
記事固有事実の保持5/5	適用単価が表示価格の約4倍
140字遵守5/5	実コスト範囲2.5倍と変動が大きい
指標を分けて記述できる試行がある	出力トークン12,818で最大級

適性: 事実の網羅性が重要で、待ち時間を許容できる草案生成。
必須ガードレール: プロバイダーの固定、数値と指標名の対応検査。

6. 推奨構成 | 用途別の使い分け

6.1 単一モデルを標準化する場合

GPT-5.6 Luna が第一候補である。実コスト最安（0.16円）、無修正5/5、コストのばらつきも最小で、4つの評価軸のうち3つで首位につけた。2位のGLM 5.3 Flashに対してコストで3.2分の1、処理時間で8分の1であり、単一の指標に依存した優位ではない。

留意点は、情報を削りすぎる傾向である。プロンプトで必要事実を明示指定すれば実用上の問題にはならない。

6.2 用途別ルーティング

とはいえ、1つのモデルですべてを賄う必要はない。3つの評価軸で首位が異なるためである。

軸	首位	2位
実コスト	GPT-5.6 Luna（0.16円）	GLM 5.3 Flash（0.52円）
処理速度	Gemini 3.7 Flash（14.8秒）	Gemini 3.8 Flash（26.2秒）
指示遵守（無修正件数）	GPT-5.6 Luna（5/5）	DeepSeek V4 Flash 0731（4/5）

コスト最安のGPT-5.6 Lunaは速度で3位、速度首位のGemini 3.7 Flashは遵守度で最下位である。**どの軸を優先するかは業務が決めるものであり、モデル側では決められない。**したがって業務特性に応じて割り当てる。

用途	推奨モデル	決め手	必須ガードレール
全社標準・日常業務	GPT-5.6 Luna	実コスト最安・無修正5/5・予測しやすい	必要事実の明示指定
即応が必要な対話	Gemini 3.7 Flash	中央値14.8秒で最速かつ最も安定	文字数計算、数値検査、人の確認必須
数値の正確さを要する短文	Gemini 3.8 Flash	数値の一般化が1/5	必要事実の明示指定
非同期・夜間バッチ	GLM 5.3 Flash	0.52円かつ数値の一般化0/5	タイムアウト設計、割引終了後の再評価
事実網羅を要する草案	DeepSeek V4 Flash 0731	固有事実5/5・無修正4/5	プロバイダー固定、数値検査

6.3 運用ルール

- 標準モデルで開始し、速度が業務価値に直結する場面でのみ高速モデルへ切り替える
- モデル選定は単価表ではなく、自社の代表タスクを実際に走らせた請求額で判断する
- 推論量（reasoning effort）を設定できるモデルでは、タスクの難易度に応じて明示的に下げる
- プロバイダーを固定する。表示価格ではなく、実際にルーティングされる経路の単価で見積もる

- 外部公開文には、モデルを問わず「数値と指標名の対応検査」「文字数の外部計算」「正式名称の辞書照合」を通す
- モデルの自己申告（「140字以内に収めました」等）を検証値として使わない
- 検査不合格時は同一モデルへ無制限に再試行させず、別モデルまたは人へエスカレーションする

6.4 価格前提が変化した場合の分岐

本検証の全モデルが割引またはflexティアで課金されている。標準価格に戻った場合の試算は次のとおりである。

モデル	現行（実測）	標準価格時（推定）	前提
GPT-5.6 Luna	0.16円	0.32円	flex→標準で2倍
GLM 5.3 Flash	0.52円	1.04円	ローンチ割引終了で2倍
Gemini 3.7 Flash	0.98円	1.96円	flex→標準で2倍
DeepSeek V4 Flash 0731	1.36円	変動	プロバイダー選択に依存
Gemini 3.8 Flash	1.65円	3.30円	flex→標準で2倍

いずれも入力・出力・キャッシュ読取に同率で適用されているため、単純に2倍となる。**序列は変わらないが、絶対額は倍になる。**

とくにGLM 5.3 Flashのローンチ割引は2026年9月9日終了予定であり、期限が近い。バッチ処理の予算をこの価格で組む場合、終了後の1.04円で再計算しておく必要がある。

なおflexティアは優先度が低く、標準ティアへ切り替えればレイテンシは改善する可能性がある。速度とコストを同時に評価する場合、ティアを揃えた再計測が必要である。

7. 本検証の限界

- 試行数が少ない。** 各モデル5試行であり、統計的一般化には不十分である
- タスクが1種類のみ。** 短文マーケティング作成であり、長文生成・コード生成・ツール操作・長い会話での条件保持を代表しない。**用途別の割り当ては、各業務の代表タスクで再測定して確定すべきである**
- 推論量の制御を行っていない。** reasoning effortを既定値のまま実行しており、明示的に下げた場合のコストと品質は測定していない
- flexティアでの計測である。** 標準ティアではコストが倍になり、レイテンシは改善する可能性がある
- プロバイダーを固定していない。** GLM 5.3 Flashは20段階中2段階がDeepInfraへ、18段階がGMICloudへルーティングされた。DeepSeekの適用単価は表示価格と大きく異なる
- 品質判定に人の判断を含む。** 数値表現の妥当性は機械判定できず、評価者間一致度は測定していない
- モデル本体の費用のみを測っている。** エージェント基盤のシステムプロンプト、ツール実行、履歴の再投入は含まないため、実システムの請求額はこれを上回る（2.2節）

8. 次回検証の設計

タスクの種類を増やすことを最優先とする。 用途別の割り当てを提案する以上、根拠が1種類のタスクでは足りない。

- 長文要約、数値抽出とスキーマ準拠出力、Web検索を伴う調査、外部ツール操作、長い会話での条件保持を追加する
- タスク種別ごとにコスト順位が入れ替わるかを検証する

推論量の制御を変数に加える

- 各モデルについて reasoning effort を低・中・高で振り、コスト・時間・品質の三者関係を測る
- 「140字の投稿文に12,000トークンの思考が必要か」を実証的に検証する

計測設計

- 各モデル・各条件30試行以上、実行順序をランダム化し複数日・複数時間帯へ分散
- プロバイダーとサービスティアを固定し、記録する
- TTFT（初回トークンまでの時間）を追加記録する。総処理時間だけではユーザー体感を説明できない
- 標準ティアとflexティアを両方測り、価格とレイテンシのトレードオフを定量化する

評価方法

- 正式名称・数値・URL・文字数を自動検査する
- 2名以上がモデル名を隠してブラインド評価し、評価者間一致度を記録する
- **主要KPIを「1無修正成果物あたりの費用」とする**

付録A | 全25試行の実測値

試行	モデル	処理時間	入力token	出力token	うち推論	コスト(USD)	コスト(円)
L1	GPT-5.6 Luna	32.55秒	2,886	1,376	871	\$0.00111	0.167
L2	GPT-5.6 Luna	21.51秒	2,837	887	375	\$0.00082	0.122
L3	GPT-5.6 Luna	34.01秒	2,910	1,279	761	\$0.00106	0.159
L4	GPT-5.6 Luna	32.79秒	2,941	1,232	697	\$0.00103	0.155
L5	GPT-5.6 Luna	42.20秒	2,887	1,575	1,081	\$0.00123	0.185
Z1	GLM 5.3 Flash	334.62秒	2,743	16,721	15,933	\$0.00439	0.658
Z2	GLM 5.3 Flash	136.39秒	3,015	6,615	5,879	\$0.00188	0.282
Z3	GLM 5.3 Flash	203.77秒	2,746	13,877	13,144	\$0.00368	0.551
Z4	GLM 5.3 Flash	509.93秒	1,143	16,393	15,781	\$0.00418	0.628
Z5	GLM 5.3 Flash	262.96秒	2,700	11,739	11,179	\$0.00310	0.465
G1	Gemini 3.7 Flash	16.95秒	2,650	3,105	2,575	\$0.00681	1.022
G2	Gemini 3.7 Flash	14.75秒	2,624	2,833	2,294	\$0.00630	0.944
G3	Gemini 3.7 Flash	17.84秒	2,550	4,010	3,559	\$0.00847	1.271
G4	Gemini 3.7 Flash	13.77秒	2,713	2,641	2,082	\$0.00597	0.895
G5	Gemini 3.7 Flash	13.12秒	2,732	2,259	1,679	\$0.00526	0.789
D1	DeepSeek V4 Flash 0731	111.64秒	2,934	12,945	12,476	\$0.00919	1.378
D2	DeepSeek V4 Flash 0731	196.10秒	2,978	17,288	16,831	\$0.01207	1.810
D3	DeepSeek V4 Flash 0731	169.56秒	3,101	18,998	18,510	\$0.01302	1.954
D4	DeepSeek V4 Flash 0731	76.19秒	3,342	7,885	7,283	\$0.00579	0.869
D5	DeepSeek V4 Flash 0731	61.52秒	3,290	6,974	6,403	\$0.00513	0.770
N1	Gemini 3.8 Flash	30.31秒	2,533	5,732	5,292	\$0.01170	1.755
N2	Gemini 3.8 Flash	26.25秒	2,515	6,595	6,153	\$0.01331	1.996
N3	Gemini 3.8 Flash	23.73秒	2,592	5,436	4,966	\$0.01116	1.675
N4	Gemini 3.8 Flash	30.58秒	2,553	5,557	5,075	\$0.01138	1.707
N5	Gemini 3.8 Flash	17.55秒	2,650	3,372	2,848	\$0.00732	1.097

付録B | 段階別の平均処理時間

モデル	段階1	段階2	段階3	段階4
Gemini 3.7 Flash	5.6秒	3.1秒	3.1秒	3.4秒
Gemini 3.8 Flash	7.6秒	7.3秒	6.6秒	4.2秒
GPT-5.6 Luna	10.2秒	7.1秒	10.3秒	5.0秒
DeepSeek V4 Flash 0731	63.6秒	31.2秒	21.4秒	6.8秒
GLM 5.3 Flash	118.4秒	34.1秒	83.9秒	53.1秒

3案を作成する段階1が最も重く、条件を調整するだけの段階4が最も軽い。ただしGLM 5.3 Flashは段階3・4でも50～84秒を要しており、単純な修正指示でも思考量が落ちない。**対話を重ねる業務ほど、この差は累積する。**

付録C | 単価の出典と実行経路

モデル	モデルページ	実際の経路
GPT-5.6 Luna	https://openrouter.ai/openai/gpt-5.6-luna	OpenAI (flex)
GLM 5.3 Flash	https://openrouter.ai/z-ai/glm-5.3-flash	GMICloud fp8 (18段階) / DeepInfra fp8 (2段階)
Gemini 3.7 Flash	https://openrouter.ai/google/gemini-3.7-flash	Google AI Studio (flex)
Gemini 3.8 Flash	https://openrouter.ai/google/gemini-3.8-flash	Google AI Studio (flex)
DeepSeek V4 Flash 0731	https://openrouter.ai/deepseek/deepseek-v4-flash-0731	SiliconFlow fp8

適用単価は、実課金額と各段階のトークン数から最小二乗法で逆算した。全モデルで残差は 10^{-10} 未満であり、上記の単価が実際に適用されたことを確認している。

本検証は2026年9月4日の実測に基づきます。為替レートは1USD＝150円換算。単価は同日時点の適用値であり、割引・ティア条件によって変動します。

自社の業務で、同じ検証をしてみませんか

モデルの実効コストは、単価とトークン消費量の掛け算で決まります。消費量はタスクによって変わるため、単価表からは読み取れません。同じ計測をお客さまの代表業務に当てはめれば、どのモデルをどの業務に割り当てるべきかを、自社の数字で判断できます。

検証設計のご相談、RiN familyの導入検討、いずれもお気軽にお声がけください。

お問い合わせ：rinfamily.ai

