

AI利用における
「APIコスト超過」と
「VRAM不足」を回避するための指標



1. AI導入時の「解像度」を上げる
2. 生成AIのAPIコストの考え方
3. 【ツール活用①】APIコスト試算
4. GPUのVRAM容量の考え方
5. 【ツール活用②】VRAM必要量試算
6. まとめ

※本資料に使用しているイラスト画像は全てNano Banana 2により作成

「なんとなく」ではなく

「根拠をもって」数値化したい

1.AI導入時の「解像度」を上げる

AI導入時に直面する「具体的なようで、実態が見えない数値」

APIコストの疑問

「月間500万トークンまで定額」と言われても、それが全社で使うのに十分なのか、あるいは1日で枯渇する量なのか判断がつかない。

日本語に関する注意事項：日本語は一般的に文字数よりトークンを多く消費(約1.1~1.3倍)するため文字数だけで計算すると予算がショートするリスクがある。

ハードウェアの疑問

「社内PCに搭載されているGPUのVRAM16GBで動くのか？」
「最新モデルを動かすには、どのランクのGPUを発注すべきか？」
などの基準が不明確。

1.AI導入時の「解像度」を上げる

検討段階で「リソース（予算・GPU）」と「性能（モデル）」の
相関をツールを用いて即時に試算できる状態にし

意思決定のスピードを上げる。

月500万トークンって
どのくらいで
消費する？



50名で
1日1回利用可能

**試算ツールでコストや
VRAM要件を可視化**

GPUのVRAM容量は
16GB？24GB？



4-bit量子化なら
16GBのVRAM容量でも
20BのLLMを載せられる

トークン消費構造

トークン消費構造を理解し、利用人数・頻度からコストを算出（以下はRAG構成時の例）

項目	トークン数の目安	補足
システムプロンプト	200~1,000トークン	指示文（「あなたは誠実な…」等）
ユーザーの質問	50~200トークン	ユーザーが入力した生の質問文
検索結果（コンテキスト）	1,500~4,000トークン	データベースから抽出した資料 （最大の消費源）
アウトプット	300~800トークン	生成される回答の長さ
合計	2,050~6,000トークン	1回あたりの総消費量

《試算例》

1回に5,000トークンを消費する場合
⇒500万トークンには「1,000回」で到達

例)
50名が月20日、1日1回利用すると月1000回となる

※日本語倍率(文字数の約1.1倍~1.3倍) に換算済み

※RAG = Retrieval-Augmented Generation(検索拡張生成)

3. 【ツール活用①】APIコスト試算

《トークン数とコスト試算ツール：デモ》

トークン数とコスト試算ツール 2026 Feb Edition
モデルを選択するか数値を変更して、月間費用をシミュレーション

月間試算コスト (30日) **¥169** 税込目安

月間総トークン **984,960 tk**

1日あたりの費用 **¥5.627**

1チャット消費 **5,472 tk**

主要モデル・プリセット

gpt-5.4 gpt-4o-mini gpt-5-nano Claude Haiku 4.5 Gemini 3 Flash

モデル単価 (\$/1M TK) 利用規模

INPUT	OUTPUT	ユーザー数	1日回数/人
<input type="text" value="1.0"/>	<input type="text" value="2"/>	<input type="text" value="6"/>	<input type="text" value="1"/>

平均出力文字数

RAG 内部パラメータ設定

換算設定: 150円 | 日本標準率: 1.2x | 計算式: (Input / 1M * Price) + (Output / 1M * Price) * 1.50
※ 算額は2026年2月時点の観測データに基づいた参考値です

・ローカルで動作

HTMLで作成しているため
オフラインで利用可能。

・入力パラメータ

- ・ 主要モデル
- ・ ユーザー数
- ・ 利用回数
- ・ 平均出力文字数

・内部パラメータ

プリセット項目(変更可)

- ・ 100万トークンあたりの金額
- ・ システムプロンプトの文字数
- ・ 質問文字数
- ・ RAGのチャンク文字数
- ・ TOP-Kの数

4. GPUのVRAM容量の考え方

ローカル環境でLLMを動かす際は「GPUのVRAM容量」が重要



VRAM不足のリスク
(OOM: Out of Memory)

容量が1MBでも不足すると
モデルは起動すらしないか
推論速度が極端に低下。



推論速度の維持

全てのデータがVRAM内に収まって
初めて、実用的でストレスの無い速
度での応答が可能となる。

必要なVRAM容量の計算ロジック

モデル重み + 会話メモリ(KVキャッシュ) + システム余白

※単純なモデルサイズ以上の容量が必要になる。

4.インフラ選定：VRAM容量の考え方

・モデル重み

土地(VRAM)に載る「建物」の大きさ。
パラメータ数(B)と量子化ビット数
(4-bit/8-bit等)で決まる
モデルそのもののサイズ。

「キャッシュ」と呼ばれる通り
高速化のために利用されている「資材」。
このKVキャッシュ分の土地(VRAM)の
空きも考慮する必要がある。

・KVキャッシュ



・システム余白

何も載っていない「庭」。
ディスプレイや推論時に一時的に使われる
作業領域で、一定容量を確保しておく必要
がある。
※設定等により変動するが、ツールでは1.5GBで設定

・GPUのVRAM容量

LLMモデルを動かすための「土地」。
足りなければ動作せず、余裕があれば
(KVキャッシュによる)高速な会話が可能。

5. 【ツール活用②】 VRAM必要量試算

- 「モデル起点」の選定

「高度な論理推論が必要 → 30Bクラスのモデル → 24GB以上のGPUが必要」というように使いたいモデルを起点としてGPUを選定する方法

- 「リソース起点」の選定

「手持ちのGPUが16GB → 量子化した12B~14Bクラスを使う」というように利用可能なGPUの容量が決まっており、それに合わせてモデルを選定する方法

どちらを起点とした場合でも試算可能な

モデルのパラメータやトークン数などから

必要なGPUのVRAM容量を試算するツールを作成

GPU VRAM要件シミュレーター
モデルサイズと利用状況から、必要なGPUスペックを算出します。
※全て理論値のよび概算となります。

1. モデル基本構成
パラメータ数 (BILLIONS) 20 精度 (PRECISION) 8-bit (標準的品質)
2. ワークロード設定
コンテキスト長 (RAG込み) (最大トークン) 8132 tk
同時推論 (BATCH SIZE) 1

必要VRAM容量
23.9 GB

モデル読み込み 20.0 GB
RAGキャッシュ 1.4 GB
システムメモリ 1.5 GB

推奨ハードウェア構成
GPU RTX 4060 Ti
GPU RTX 5080 / 4080
GPU RTX 5090

5. 【ツール活用②】 VRAM必要量試算

《GPU VRAM要件シミュレーター：デモ》

GPU VRAM要件シミュレーター
モデルサイズと利用状況から、必要なGPUスペックを算出します。
※全て理論値および概算となります。

1. モデル基本構成

パラメータ数 (BILLION) B 量子化 (PRECISION)

2. ワークロード設定

コンテキスト長 (RAG込/合計トークン)

同時接続数 (BATCH SIZE)

推奨ハードウェア構成

CONSUMER RTX 4060 Ti 対応可能 16GB	CONSUMER RTX 5080 / 4080 対応可能 16GB
CONSUMER FLAGSHIP RTX 5090 / 4090 対応可能 24GB	PROFESSIONAL RTX 6000 Ada 対応可能 48GB

必要総VRAM容量
6.5 GB

モデル重み 4.0 GB
KVキャッシュ 1.0 GB
システム余白 1.5 GB

※ 理論をVRAM内で完結させるための最小要件です。これ以下のVRAMでは推論速度が大幅に低下します。

・ ローカルで動作

トークン数とコストの試算ツールと同じくHTMLで作成しているためオフラインで利用可能。

・ 入力パラメータ

- ・ LLMモデルのパラメータ数(B)
- ・ 同時接続数(バッチサイズ)
- ・ 量子化
- ・ コンテキスト長

・ GPU選定

設定した数値から試算した“必要総VRAM容量”を元に推奨されるGPUを選定。

試算ツールの
利用



「なんとなく」ではなく「**根拠をもって**」
判断する基準が出来る



APIコストおよび
VRAM必要量を試算

⇒ **可視化**



「いくらかかるか分からない」
「動くか分からない」という不安を解消し
意思決定のスピードアップに繋がられる

承認プロセスの
スピード **UP**



現場の意思決定
スピード **UP**

 株式会社ブライエ