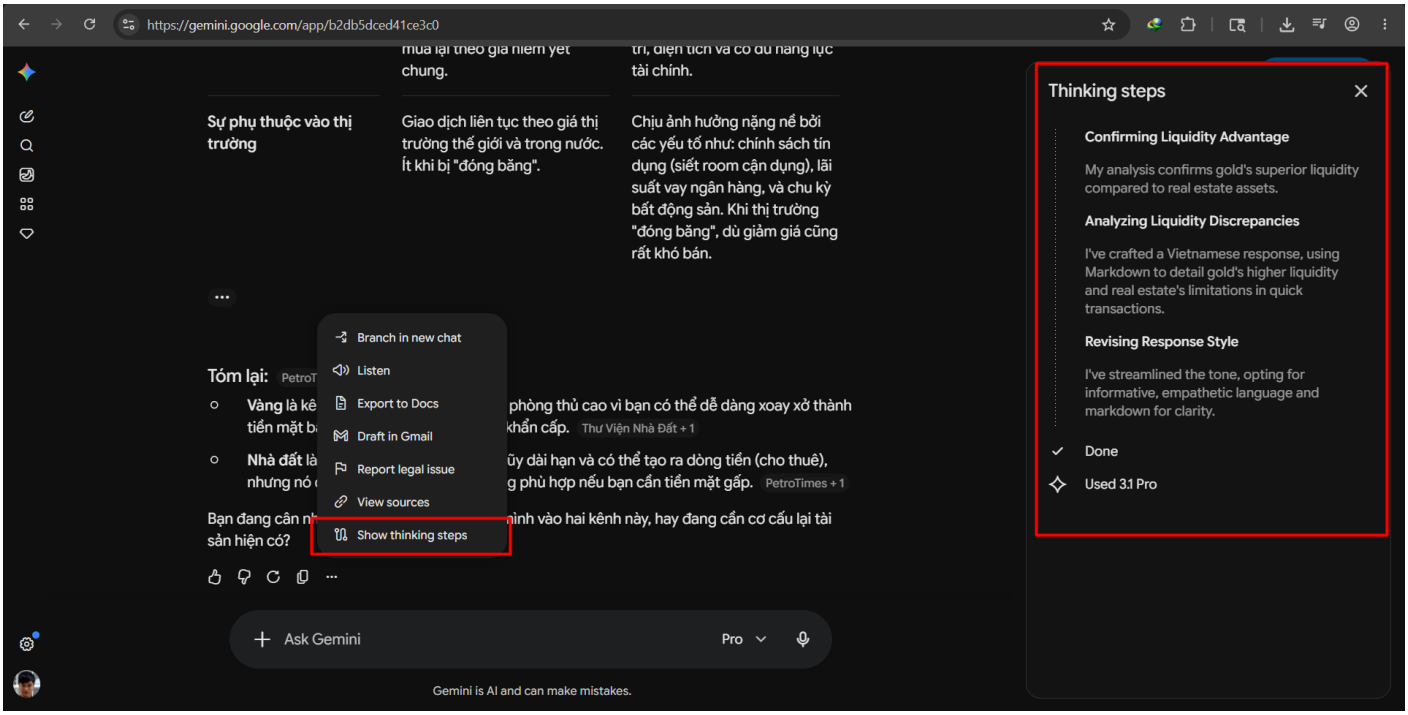
🤔 AI trả lời bằng tiếng Việt tốn token hơn !

🤔 Phân tích hoạt động của AI.

Đầu tiên, người dùng sẽ nhập câu hỏi vào khung chat.
Sau đó sẽ có bước thinking.
Đây là bước mà AI sẽ tự lẩm bẩm, tự trả lời, tự kiểm chứng để tăng tính chính xác.

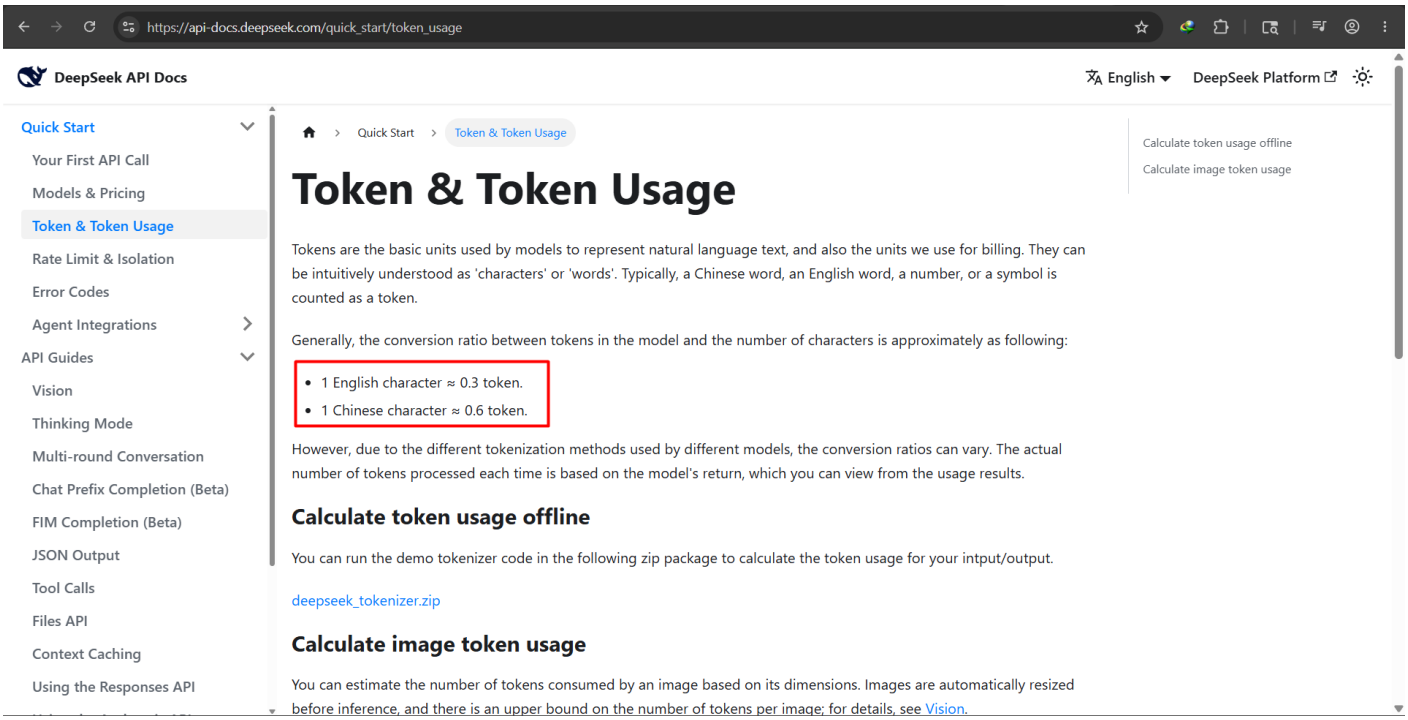


Như đã thấy, bên trái hỏi bằng tiếng Việt, bên phải lại hiển thị thinking bằng tiếng Anh. Tức là hoạt động của AI sẽ chạy như sau:

[Người dùng nhập prompt]
↓ [AI dịch sang tiếng Anh] **Lưu ý: Dịch ở đây là bước kích hoạt mạng thần kinh của AI, không phải công cụ như Google Translate**
↓ [AI Thinking]
↓ [AI dịch lại sang tiếng Việt]
↓ [AI trả lời tổng hợp dựa trên mấy điều đã đưa ra trong thinking]

Tại sao AI lại tốn token hơn ?

Đầu tiên là chữ tiếng Việt / Trung / Thái / ... dùng hệ mã hóa UTF8 / Unicode, nó dài gấp đôi so với ký tự tiếng Anh trong bộ nhớ. Các bên phát hành AI cũng nói rõ, ký tự tiếng Anh tốn 0.3 token thôi, nhưng tiếng khác thì tốn 0.6 token.



Tiếp theo là bước dịch để cho ra kết quả.
Mặc dù đã làm rất tối ưu, nhưng thêm bước vô tình cũng tăng lượng token sử dụng.
Và chi phí này cộng dồn không hề rẻ.

Giải pháp đỡ tốn token hơn

1. Dùng Google Dịch bỏ vô rồi mới lấy kết quả tiếng Anh để bỏ vô AI.
2. Xây dựng AI trung gian. Ví dụ Gemini giá cao, cho nó suy luận, còn AI giá thấp thì cho nó dịch (thậm chí là miễn phí, các model 4B hiện nay chạy ở máy người dùng dịch đã rất tốt).

Revision #5
Created 2026-08-30 00:17:09 UTC by sandichhuu
Updated 2026-08-30 07:54:26 UTC by sandichhuu