QUÀ TẶNG · AI CHO MỌI NGƯỜI · TỪ CA 13

Phiếu "đo trước khi cài"

RAM đủ chưa có nghĩa là đủ nhanh. Đo máy của mình trước khi chạy model về máy.

Một thử nghiệm thật: chạy model 32 tỉ tham số trên máy không có card đồ họa. Tốc độ khoảng 1,35 token mỗi giây, tức 227 giây để viết một đoạn ngắn, và kết quả còn lẫn chữ nước ngoài. Model nhỏ hơn (7 tỉ tham số) vẫn làm tốt việc lọc và phân loại. **Máy chứa được model chưa có nghĩa là máy chạy nổi model.**

1 Trước khi cài

- [ ] Tôi ghi lại cấu hình máy: bộ xử lý, RAM, có card đồ họa không.
- [ ] Tôi biết việc mình cần: viết dài, hay chỉ lọc, phân loại, trả lời ngắn.
- [ ] Tôi đặt ngưỡng chấp nhận: chờ tối đa bao nhiêu giây cho một lần trả lời?

2 Đo (10 phút)

1. Chọn một yêu cầu cố định, ví dụ: "Viết 100 từ về một ngày đi chợ".
2. Bấm giờ từ lúc gửi đến lúc xong. Ghi số giây.
3. Đếm số token đầu ra nếu công cụ có hiển thị. Nếu không, đếm số từ và coi đó là ước lượng thô.
4. Tốc độ = số token ÷ số giây. Thử 3 lần, lấy số giữa.

Model	Lần 1 (giây)	Lần 2 (giây)	Lần 3 (giây)	Token/giây	Chất lượng (đạt hay không)

3 Quyết định

- Số giây một lần trả lời có nhỏ hơn ngưỡng chấp nhận không?
- Chất lượng có đủ cho việc đó không, hay có lỗi lạ (lẫn ngôn ngữ, bỏ ý)?
- Nếu chậm: thử model nhỏ hơn cho việc đơn giản, và để việc nặng cho dịch vụ khác.

Giới hạn. Số đo chỉ đúng cho máy và cách chạy của anh chị lúc đó. Phiếu này không so sánh các model với nhau.

Quà tặng từ Nhiệt Thành · nhietthanhreal.com · Mẫu trống, dùng được ngay, không cần đăng ký. Soạn từ việc đã làm thật, có sự hỗ trợ của AI và đã được người xem lại. Dùng, sửa và chia sẻ cho người cần.