

Lựa Chọn Mô Hình AI Phù Hợp: Hiệu Quả Hay Model Càng Lớn Càng Tốt

Mở đầu (2 phút)

Trong bối cảnh các mô hình ngôn ngữ lớn (LLM) ngày càng mạnh mẽ, nhiều người mặc định rằng “model càng lớn càng tốt” – kể cả với các tác vụ đơn giản như viết lại email hay tóm tắt văn bản ngắn. Tuy nhiên, điều này đang dẫn đến sự lãng phí lớn về tài nguyên và thời gian tính toán.

Bài trình bày này sẽ chia sẻ một góc nhìn thực tiễn: Vì sao việc chọn đúng mô hình cho đúng công việc không chỉ giúp tối ưu hiệu năng, mà còn giúp xây dựng các hệ thống AI hiệu quả, nhanh gọn và dễ triển khai hơn. Thông qua một số ví dụ cụ thể với LLaMA 3, Gemma, và LLaMA 4, bài nói sẽ minh họa cách những mô hình nhỏ có thể hoàn toàn đủ sức – nếu được dùng đúng chỗ.

“Có bao giờ bạn chờ 30 giây để một LLM ‘nghìn tỷ tham số’ viết lại dòng email, summary 1 cuộc họp hoặc chỉ là dịch sang một ngôn ngữ khác vì bạn không tin vào các phần mềm dịch thuật truyền thống?”

- Giới thiệu nhanh về xu hướng sử dụng LLM siêu lớn cho mọi tác vụ.
- Nhấn mạnh: mô hình càng to **không đồng nghĩa với đầu ra tốt hơn**, đặc biệt với các tác vụ nhỏ.

Show 3 kết quả email rewrite. Gợi ý: đầu ra không khác biệt lớn, nhưng thời gian và tài nguyên lại khác rõ.

“Bigger ≠ Better?”

- Nhiều cá nhân và startup mặc định dùng LLM lớn nhất có thể.
- Hệ quả:
 - Thời gian phản hồi chậm
 - Chi phí GPU tăng vọt
 - Khó triển khai trong môi trường thực tế (edge, mobile, nội bộ doanh nghiệp...)

Phần Live Demo + So sánh (8 phút)

- Tác vụ: viết lại email gốc: *"Help me write an title for my speak about select the right ai model instead of pick the most powerful model with big parameter just 5 ideas."*
- Kết quả với 3 model:
 - **GPT 4.1 mini:** phản hồi trong 2s
 1. "Choosing the Right AI Model: Quality Over Size"
 2. "Smart AI Selection: Why Bigger Isn't Always Better"
 3. "Beyond Big Parameters: Finding the Best AI Model for Your Needs"
 4. "Right Model, Right Results: Selecting AI for Purpose, Not Power"
 5. "Less Can Be More: How to Pick the Optimal AI Model"

O3: phản hồi trong 9s

1. "Fit-for-Purpose AI: Choosing Smart Models Over Massive Ones"
2. "Right-Sized Intelligence: Why Bigger Isn't Always Better in AI"
3. "Beyond Parameter Count: A Practical Guide to Picking the Perfect Model"
4. "From Gigabytes to Good Enough: Strategies for Selecting the Model That Works"
5. "Efficiency Over Excess: Mastering Model Selection Without the Megascale"

- **Gemini 2.5 Pro:** 13s, nhưng đầu ra chỉ khác... vài dấu chấm phẩy

Right-Sizing Your AI: Why Smarter Beats Bigger in Model Selection.

The Goldilocks AI: Finding the "Just Right" Model, Not the Most Powerful.

Beyond Big Parameters: A Practical Guide to Efficient AI Model Choice.

Precision Over Power: How to Select the Right AI Model for Your Needs.

Lean AI: Achieve More with Less – Smart Choices in Model Selection.

- 🎯 **Tác vụ: "Give me a list of title 5 ideas for a tech talk at an AI meetup. The topic is about choosing the right AI model for a task instead of just picking the biggest or most powerful model. The tone should be both technical and inspirational. Make the titles bilingual—English and Vietnamese. The audience includes AI engineers, students, and investors. Keep the titles sharp and meaningful"**

- ⚖️ **So sánh ba mô hình:**

Thông điệp chính

- “Đúng mô hình – đúng việc” là cách triển khai AI hiệu quả nhất.
- Hãy đánh giá bài toán trước – rồi chọn model phù hợp sau.
- Với mô hình nhỏ:
 - Dễ triển khai trên thiết bị yếu
 - Chi phí thấp
 - Đáp ứng đủ cho >80% tác vụ thông dụng

Slide 7: Gợi ý cho cộng đồng

- Với dev/engineer: test với mô hình nhỏ trước khi scale.
- Với startup: ưu tiên mô hình dễ nhúng, fine-tune được.
- Với researcher: thử nghiệm với mô hình nhỏ giúp tiết kiệm thời gian/gpu nhưng vẫn đủ insight.

1. Kết thúc (1-2 phút)

Slide 8: “Làm AI không chỉ là mạnh – mà còn phải hợp lý.”

- Gợi mở: bạn đã bao giờ *benchmark* mô hình nhỏ cho tác vụ đơn giản chưa?
- Mời kết nối: chia sẻ thêm tại sự kiện / LinkedIn / GitHub

Tổng kết kỹ thuật (2 phút)

- Model lớn hơn thường sinh văn bản sáng tạo hơn – nhưng không phải lúc nào cũng tốt hơn.
- Với tác vụ **có cấu trúc rõ, đầu ra đơn giản**, mô hình nhỏ (2B–7B) hoàn toàn phù hợp.
- Tối ưu hóa tài nguyên = tiết kiệm chi phí + tăng tốc độ hệ thống.

Thông điệp kết thúc (3 phút)

“Không phải cứ nấu mì là phải đốt lò nướng. Chọn mô hình đúng với bài toán chính là cách chúng ta **làm AI thông minh hơn – cả về kỹ thuật lẫn kinh tế.**”

- Kêu gọi cộng đồng kỹ thuật & startup:
 - Đừng default vào LLM siêu lớn cho mọi thứ.

- Xây dựng mô hình nhỏ, dễ triển khai cho bài toán nhỏ – là hướng đi đáng giá cho các nhóm nhỏ/startup.
- Nếu bạn làm open source AI: Hãy mở đầu bằng mô hình nhỏ thật tốt.

Nếu bạn muốn, mình có thể giúp bạn:

- Soạn slide outline dựa trên nội dung trên.
- Chuẩn bị ví dụ kết quả demo cho từng model.
- Viết phần script nói ngắn gọn nếu bạn cần tập trình bày.

Chúc mừng bạn đã được mời làm diễn giả tại AI HCM Meethub! Đây là một chủ đề rất thời sự và quan trọng trong bối cảnh AI phát triển như vũ bão. Dưới đây là một dàn ý chi tiết và nội dung gợi ý cho bài trình bày của bạn, nhắm đến các đối tượng tham dự đa dạng.

Bài Trình Bày: Lựa Chọn Mô Hình AI Phù Hợp: Sự Hiệu Quả Hay Model Càng Lớn Càng Tốt?

Diễn giả: [Tên của bạn]

Sự kiện: AI HCM Meethub

Mục Tiêu Bài Trình Bày:

- Giúp người nghe hiểu rõ ưu và nhược điểm của các mô hình AI lớn so với các mô hình nhỏ hơn, hiệu quả hơn.
- Cung cấp một khung sườn và các tiêu chí cụ thể để lựa chọn mô hình AI phù hợp với nhu cầu và hoàn cảnh thực tế.
- Giới thiệu các công cụ, benchmark và tài nguyên hữu ích để hỗ trợ quá trình lựa chọn.
- Thúc đẩy tư duy phản biện về xu hướng "model càng lớn càng tốt".

Dàn Ý Chi Tiết:

(1) Mở Đầu (3-5 phút)

- **Chào mừng & Giới thiệu:**
 - Giới thiệu bản thân và lý do chọn chủ đề này.
 - Chào mừng các đối tượng tham dự (nhấn mạnh sự đa dạng và tầm quan trọng của AI với từng nhóm).
- **Bối cảnh:**

- Sự bùng nổ của AI, đặc biệt là các mô hình ngôn ngữ lớn (LLMs) và mô hình nền tảng (Foundation Models).
- Cuộc đua "vũ trang" về kích thước mô hình: GPT-4, Claude 3 Opus, PaLM 2, Llama 3 (và các model mới nhất tại thời điểm đó).
- Câu hỏi day dứt: Liệu "lớn hơn" có thực sự "tốt hơn" trong mọi trường hợp?

- **Mục tiêu & Lộ trình:** Giới thiệu ngắn gọn những gì sẽ trình bày.

(2) Cơ Sốt "Model Càng Lớn Càng Tốt": Tại Sao? (5-7 phút)

- **Sức mạnh không thể phủ nhận:**
 - Khả năng học Few-shot / Zero-shot ấn tượng.
 - Hiệu suất vượt trội trên nhiều benchmark tổng quát (GLUE, SuperGLUE, MMLU...).
 - Khả năng thực hiện các tác vụ phức tạp, đòi hỏi suy luận sâu (lập luận, sáng tạo, viết mã...).
 - Tạo ra "khả năng nổi trội" (Emergent Abilities) không có ở các mô hình nhỏ hơn.
- **Ví dụ:**
 - GPT-4 vượt qua các kỳ thi chuẩn (Bar Exam, GRE...).
 - Claude 3 Opus cho thấy khả năng "tự nhận thức" (ví dụ "needle in a haystack").
 - Các mô hình lớn tạo ra hình ảnh/video/âm thanh chân thực đến kinh ngạc.
- **Tác động:** Thu hút đầu tư lớn, tạo ra các ứng dụng đột phá, thay đổi cách chúng ta tương tác với công nghệ.

(3) Mặt Trái Của "Gã Khổng Lồ": Cái Giá Phải Trả (7-10 phút)

- **Chi Phí Khổng Lồ:**
 - **Đào tạo:** Hàng triệu đến hàng trăm triệu USD (chi phí tính toán, dữ liệu, nhân lực).
 - **Inference (Suy luận):** Chi phí vận hành cao, đặc biệt khi phục vụ hàng triệu người dùng. Cần phần cứng chuyên dụng (GPU/TPU đắt đỏ).
- **Độ Trễ (Latency):**

- Các mô hình lớn thường chậm hơn, không phù hợp cho các ứng dụng thời gian thực (ví dụ: chatbot hỗ trợ khách hàng, trợ lý ảo tức thời, xe tự lái).
- **Môi Trường:**
 - Tiêu thụ năng lượng khổng lồ, dấu chân carbon đáng kể.
- **Triển Khai & Bảo Trì:**
 - Khó khăn khi triển khai trên các thiết bị biên (edge devices), di động hoặc trong môi trường tài nguyên hạn chế.
 - Phức tạp trong việc cập nhật, tinh chỉnh (fine-tuning) và giám sát.
- **Tính Chuyên Môn Hóa:**
 - Mô hình lớn có thể là "overkill" (quá mức cần thiết) cho các tác vụ cụ thể, đôi khi không hiệu quả bằng một mô hình nhỏ được đào tạo chuyên sâu.
- **Vấn đề "Black Box" & Thiên Vị:** Càng lớn càng khó diễn giải, tiềm ẩn rủi ro thiên vị và an toàn.

(4) Khi "Sự Hiệu Quả" Lên Ngôi: Lợi Ích Của Mô Hình Nhỏ/Chuyên Biệt (5-7 phút)

- **Chi phí thấp:** Dễ tiếp cận hơn cho các cá nhân, startup và doanh nghiệp vừa và nhỏ.
- **Tốc độ nhanh:** Phù hợp với các ứng dụng đòi hỏi tương tác tức thời.
- **Triển khai linh hoạt:** Có thể chạy trên di động, IoT, edge, hoặc các máy chủ thông thường.
- **Tùy chỉnh dễ dàng:** Dễ dàng fine-tune cho các nhiệm vụ hoặc bộ dữ liệu cụ thể.
- **Tiết kiệm năng lượng:** Thân thiện hơn với môi trường.
- **Khả năng kiểm soát & diễn giải:** Thường dễ hiểu và kiểm soát hơn.
- **Ví dụ:**
 - Mistral 7B / Llama 3 8B: Hiệu suất ấn tượng so với kích thước.
 - Phi-3 (Microsoft): Các mô hình nhỏ (Small Language Models - SLMs) mạnh mẽ.

- BERT/RoBERTa (và các biến thể): Vẫn là lựa chọn hàng đầu cho nhiều tác vụ NLP cụ thể.
- Các mô hình được chưng cất (Distilled models), lượng tử hóa (Quantized models).

(5) Ví Dụ So Sánh Thực Tế (5 phút)

• Tình huống 1: Chatbot Hỗ Trợ Khách Hàng cho E-commerce:

- **Nhu cầu:** Trả lời nhanh các câu hỏi thường gặp, xử lý đơn hàng, cần độ trễ thấp, chi phí hợp lý.
- **Lựa chọn A (Lớn):** GPT-4/Claude 3 Opus - Chi phí cao, độ trễ có thể là vấn đề, có thể "sáng tạo" quá mức cần thiết.
- **Lựa chọn B (Hiệu quả):** Fine-tuned Llama 3 8B / Mistral 7B + RAG (Retrieval-Augmented Generation) - Chi phí thấp, nhanh, tập trung vào kiến thức sản phẩm, dễ kiểm soát.
- **Kết luận:** Mô hình hiệu quả thường là lựa chọn tốt hơn.

• Tình huống 2: Phân Tích & Tóm Tắt Nghiên Cứu Y Khoa Phức Tạp:

- **Nhu cầu:** Hiểu sâu sắc thuật ngữ chuyên ngành, tìm ra mối liên hệ tinh vi, yêu cầu độ chính xác cực cao.
- **Lựa chọn A (Lớn):** GPT-4/Med-PaLM 2 / Claude 3 Opus - Khả năng suy luận cao, hiểu ngữ cảnh phức tạp.
- **Lựa chọn B (Hiệu quả):** Mô hình nhỏ - Có thể bỏ lỡ các chi tiết quan trọng hoặc hiểu sai các khái niệm phức tạp.
- **Kết luận:** Mô hình lớn (hoặc mô hình chuyên biệt, có thể cũng lớn) là cần thiết.

(6) Làm Thế Nào Để Lựa Chọn Mô Hình Phù Hợp? (10-12 phút)

• Bước 1: Xác Định Rõ Bài Toán & Ràng Buộc:

- **Mục tiêu là gì?** (Phân loại, sinh văn bản, tóm tắt, hỏi đáp, nhận dạng hình ảnh...).
- **Độ chính xác yêu cầu?** (Mức độ chấp nhận sai sót là bao nhiêu?).
- **Yêu cầu về độ trễ?** (Ứng dụng có cần thời gian thực không?).
- **Ngân sách?** (Cho đào tạo, fine-tuning và inference).
- **Hạ tầng hiện có/dự kiến?** (Cloud, on-premise, edge).

- **Dữ liệu?** (Có sẵn dữ liệu để fine-tuning không?).
- **Yêu cầu về diễn giải, an toàn, đạo đức?**
- **Bước 2: Khám Phá & Đánh Giá Các Lựa Chọn:**
 - **Hugging Face Hub:** Kho tàng mô hình, bộ dữ liệu và công cụ. Khám phá các mô hình theo tác vụ, kích thước, license.
 - **Bảng Xếp Hạng (Leaderboards):**
 - **Open LLM Leaderboard (Hugging Face):** So sánh các LLM mã nguồn mở.
 - **Stanford HELM (Holistic Evaluation of Language Models):** Đánh giá toàn diện trên nhiều phương diện.
 - **Papers with Code:** Theo dõi các kết quả SOTA (State-of-the-Art) trên các benchmark cụ thể.
 - **Chatbot Arena:** So sánh "mù" các chatbot dựa trên sở thích người dùng.
 - **Model Cards & Datasheets:** Đọc kỹ tài liệu đi kèm để hiểu khả năng, hạn chế, dữ liệu đào tạo và thiên vị tiềm ẩn.
 - **Nền tảng Cloud AI (Vertex AI, Azure AI, AWS SageMaker):** Cung cấp các "Model Garden" với nhiều lựa chọn và công cụ đánh giá.
- **Bước 3: Sử Dụng Benchmarks Một Cách Cân Trọng:**
 - Hiểu rằng benchmark tổng quát không phải lúc nào cũng phản ánh hiệu suất trên *tác vụ cụ thể* của bạn.
 - Ưu tiên các benchmark liên quan trực tiếp đến bài toán.
- **Bước 4: Thử Nghiệm & Tinh Chỉnh (Quan Trọng Nhất):**
 - **Không tin mù quáng vào quảng cáo!** Hãy tự mình thử nghiệm các ứng viên tiềm năng *trên dữ liệu và kịch bản sử dụng thực tế của bạn*.
 - Thực hiện A/B testing nếu có thể.
 - Đánh giá không chỉ độ chính xác mà còn cả tốc độ, chi phí và sự ổn định.
 - Cân nhắc các kỹ thuật tối ưu hóa: Quantization, Pruning, Distillation, RAG.

(7) Xu Hướng Tương Lai (3-5 phút)

- **Sự trỗi dậy của SLMs (Small Language Models).**
- **Mô hình MoE (Mixture-of-Experts):** Cố gắng cân bằng giữa kích thước và hiệu quả.
- **Tối ưu hóa phần cứng-phần mềm:** Chip AI chuyên dụng, các framework tối ưu.
- **AI đa phương thức (Multimodal AI) hiệu quả hơn.**
- **AI tự điều chỉnh (Self-optimizing AI):** Các mô hình có thể tự điều chỉnh kích thước/kiến trúc cho tác vụ.

(8) Kết Luận & Q&A (5 phút)

- **Tóm tắt thông điệp chính:**
 - Không có "viên đạn bạc" - không có mô hình nào là tốt nhất cho mọi thứ.
 - "Lớn hơn" không phải lúc nào cũng "tốt hơn".
 - Lựa chọn mô hình là một quá trình cân nhắc kỹ lưỡng giữa **Hiệu suất - Chi phí - Tốc độ - Tính Khả Thi**.
 - Tư duy thực tế và thử nghiệm là chìa khóa.
- **Lời khuyên cho từng nhóm đối tượng:**
 - **Kỹ sư:** Hãy thử nghiệm, tối ưu và đóng góp cho mã nguồn mở.
 - **Doanh nghiệp:** Tập trung vào ROI và giải pháp phù hợp với bài toán kinh doanh.
 - **Nhà đầu tư:** Nhìn vào cả xu hướng lớn và các giải pháp ngách hiệu quả.
 - **Sinh viên:** Học hỏi nền tảng, khám phá các công cụ và đừng ngại thử nghiệm.

1. Cơ bản – nhanh, rẻ, chill

GPT 4.1 mini, GPT 4.1, Gemini 2.5 Flash, o3-mini, o4-mini

Dùng để: chat thường, viết mail, dịch thuật, note ý, social post, hỗ trợ học tập cơ bản.

2. Reasoning – suy luận, coding, phân tích đa bước

o1, o4-mini-high, GPT 4o, Claude 3.7 Sonnet, Claude 4 Sonnet

Dùng để: giải toán/logic nhiều bước, viết & debug code, phân tích dữ liệu, lập kế hoạch chuyên sâu, viết luận có cấu trúc.

3. Deep Research – trường kỳ, ngữ cảnh khổng lồ, đa tác vụ:

Gemini 2.5 Pro (Deep Research), GPT o3, GPT 4.5, Llama 4, Deepseek R1, GPT o1 Pro, Claude 4 Opus

Dùng để: tổng hợp báo cáo hàng trăm trang, phân tích pháp lý, nghiên cứu khoa học, lập agent tự trị nhiều bước, dựng luận chứng đa nguồn.

- **Chi phí & tốc độ** tăng dần từ Cơ bản → Deep Research. Hãy cân đối ví tiền và deadline.
- **Đa phương thức** (ảnh, âm thanh) tốt nhất hiện tại: GPT-4o (Reasoning) và Gemini 2.0 Flash (Cơ bản).
- **Context dài:** GPT-4.1 và các mẫu Claude/Gemini mới đều > 1 M tokens – đủ để “nuốt” nguyên thư viện luận án.
- **Bảo mật dữ liệu:** Claude nổi tiếng “cẩn thận”, còn OpenAI/Gemini cho tùy chỉnh private deployment. Luôn check chính sách trước khi upload file nhạy cảm.