Khảo Sát về LLM: Kiến Trúc, So Sánh và Đánh Giá

1 Khái Niệm và Các Kiến Trúc Phổ Biến của LLM

1.1 Khái Niệm Large Language Model

Large Language Models (LLMs) là các mô hình học máy quy mô lớn được huấn luyện trên lượng dữ liệu văn bản khổng lồ, có khả năng hiểu và sinh văn bản tự nhiên. LLMs thường có từ hàng tỷ đến hàng nghìn tỷ tham số (parameters) và được xây dựng dựa trên kiến trúc Transformer.

Các đặc điểm chính của LLMs:

- **Quy mô lớn:** Từ 7 tỷ đến 1.8 nghìn tỷ parameters
- **Pre-training:** Huấn luyện trên corpus văn bản lớn
- **Transfer learning:** Có thể fine-tune cho các tác vụ cụ thể
- **Few-shot learning:** Khả năng học từ ít ví dụ
- **Emergent abilities:** Xuất hiện các khả năng mới khi scale lên

1.2 Ba Kiến Trúc Chính của LLM

1.2.1 Encoder-Only Architecture

Đặc điểm:

- Chỉ sử dụng phần encoder của Transformer
- Bidirectional attention - mỗi token có thể tham gia tất cả tokens
- Pre-training: Masked Language Modeling (MLM)

Công thức attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Mô hình tiêu biểu: BERT, RoBERTa, DistilBERT

1.2.2 Decoder-Only Architecture

Đặc điểm:

- Chỉ sử dụng decoder với causal masking
- Unidirectional attention - token chỉ attend các tokens trước đó
- Pre-training: Next Token Prediction

Mô hình tiêu biểu: GPT series, LLaMA, BLOOM, MPT, Claude, Gemini

1.2.3 Encoder-Decoder Architecture

Đặc điểm:

- Sử dụng cả encoder và decoder
- Encoder: bidirectional, Decoder: causal + cross-attention

Mô hình tiêu biểu: T5, BART, Flan-T5

1.3 Mối Quan Hệ Giữa Các Kiến Trúc

Bảng 1: Mối quan hệ task-architecture

Kiến trúc	**Tối ưu cho**	**Lý do**
Encoder-only	Understanding tasks	Full context access
Decoder-only	Generation tasks	Autoregressive nature
Encoder-Decoder	Transformation tasks	Separate encode/decode

Decoder-only đang trở thành standard cho modern LLMs vì đơn giản nhưng hiệu quả cao

2 So Sánh LLMs Mã Nguồn Đóng

2.1 GPT-5 (OpenAI)

2.1.1 Kiến Trúc Chi Tiết

- **Architecture:** Decoder-only Transformer với Advanced MoE
- **Total parameters:** Estimated 2+ trillion
- **MoE:** Enhanced expert routing system
- **Context window:** 200K tokens (với kế hoạch mở rộng)
- **Training data:** Multi-trillion tokens với synthetic data
- **Multimodal:** Text, Images, Audio, Video (native integration)
- **Phiên bản:** GPT-5 - released 2025
- **Key features:** Enhanced reasoning, better factuality

2.1.2 Ưu Điểm

1. State-of-the-art reasoning capabilities
2. Improved factual accuracy và reduced hallucinations
3. Advanced multimodal understanding
4. Better long-term planning và task decomposition
5. Enhanced code generation và debugging
6. Stronger mathematical reasoning
7. Improved instruction following
8. More consistent outputs
9. Better multilingual support (100+ languages)

2.1.3 Nhược Điểm

1. Closed source - không thể truy cập weights
2. Chi phí cao hơn các phiên bản trước
3. API dependency
4. Rate limits nghiêm ngặt hơn
5. Data privacy concerns
6. Requires substantial computational resources
7. Không thể customize hoặc fine-tune
8. Higher latency cho complex tasks

2.2 Claude 4.5 (Anthropic)

2.2.1 Kiến Trúc Chi Tiết

- **Architecture:** Advanced Decoder-only Transformer
- **Parameters:** Estimated 150-200 billion
- **Context window:** 500,000 tokens (mở rộng đáng kể)
- **Constitutional AI:** Enhanced self-critique mechanisms
- **RLHF:** Advanced reinforcement learning
- **Multimodal:** Text + Images + Audio (mới)
- **Extended Thinking:** Improved chain-of-thought reasoning
- **Phiên bản:** Claude 4.5 - released 2025

2.2.2 Ưu Điểm

1. Best-in-class reasoning: Top performance trên coding benchmarks
2. Extended context: 500K tokens (tăng gấp 2.5 lần)
3. Constitutional AI nâng cao cho ethics và safety
4. Extremely low error rate
5. Enhanced nuanced understanding
6. Superior code generation và analysis
7. Context caching improvements: Lên đến 95% cost reduction
8. Detailed, structured responses
9. Excellent vision và audio capabilities (mới)
10. Better real-world problem solving

2.2.3 Nhược Điểm

1. Context window vẫn nhỏ hơn Gemini (500K vs 2M)
2. Chi phí cao
3. Closed source
4. API only
5. Limited deployment options
6. Slower inference cho complex reasoning
7. Conservative outputs trong sensitive topics
8. Không thể fine-tune

2.3 Gemini 2.5 Pro (Google DeepMind)

2.3.1 Kiến Trúc Chi Tiết

- **Architecture:** Advanced Sparse MoE Transformer
- **Total parameters:** Estimated 250B+ (với sparse activation)
- **Context window:** 2,000,000 tokens (2M) - maintained leadership
- **Multimodal:** Text, Images, Audio, Video (fully native)
- **MoE:** Enhanced expert routing với better efficiency
- **Ring attention:** Optimized long context processing
- **Agentic capabilities:** Native tool use và function calling
- **Phiên bản:** Gemini 2.5 Pro - released late 2025

2.3.2 Ưu Điểm

1. Best coding accuracy: Top SWE-bench scores
2. Ultra-long context: 2M tokens maintained
3. Fully multimodal (text, image, audio, video)
4. Most cost-effective: Lowest per-token pricing
5. Enhanced MoE efficiency
6. Fast inference với sparse activation
7. Integrated Google ecosystem (Search, Workspace)
8. Strong performance across diverse benchmarks
9. Advanced agentic capabilities
10. Free tier với generous limits
11. Real-time information integration

2.3.3 Nhược Điểm

1. Chi phí tăng đáng kể với full context (>1M tokens)
2. Closed source
3. Complex MoE architecture
4. Không thể customize
5. Rate limits cho free tier
6. Inconsistent performance trong một số edge cases
7. MoE routing complexity
8. API dependency

2.4 Bảng So Sánh

Bảng 2: So sánh LLMs mã nguồn đóng (phiên bản mới nhất 2025)

Tiêu chí	**GPT-5**	**Claude 4.5**	**Gemini 2.5 Pro**
Architecture	Decoder MoE	Decoder Dense	Decoder MoE
Parameters	approximately 2T+	150-200B	approximately 250B+
Context	200K tokens	500K tokens	2M tokens
Multimodal	All modalities	Text, Image, Audio	All modalities
SWE-bench	approximately 65%+	approximately 70%+	approximately 72%+
MMLU	approximately 92%+	approximately 93%+	approximately 91%+
HumanEval	approximately 93%+	approximately 95%+	approximately 92%+
Tốc độ	Fast	Medium-Fast	Very Fast
Chi phí (input)	$3-5/1M	$4-6/1M	$1.5-2.5/1M
Chi phí (output)	$12-20/1M	$15-25/1M	$8-12/1M
Release	2025	2025	Late 2025
Ưu điểm nổi bật	Reasoning, General	Coding, Ethics	Context, Cost, Speed

3 So Sánh LLMs Mã Nguồn Mở

3.1 LLaMA 3.3 (Meta AI)

3.1.1 Kiến Trúc Chi Tiết

- **Architecture:** Causal decoder-only Transformer
- **Model sizes:** 1B, 3B, 8B, 70B, 405B parameters
- **Layers:** 16 (1B), 28 (3B), 32 (8B), 80 (70B), 126 (405B)
- **Hidden size:** 2048 (1B), 3072 (3B), 4096 (8B), 8192 (70B), 16384 (405B)
- **Context:** 128K tokens (tất cả variants)
- **Position encoding:** RoPE (Rotary Position Embedding)
- **Activation:** SwiGLU
- **Normalization:** RMSNorm
- **Attention:** Grouped Query Attention (GQA)
- **Training data:** 15+ trillion tokens (high-quality curated)
- **Phiên bản:** Llama 3.3 - released October 2025

3.1.2 Ưu Điểm

1. Fully open source với permissive license
2. Parameter efficient - performance vượt trội so với size
3. Multiple sizes: 1B đến 405B cho mọi use case
4. Competitive với GPT-5: 86.9% MMLU (405B)
5. Massive community ecosystem
6. Free to use, self-hostable
7. Fully customizable và fine-tunable
8. Extended context: 128K tokens cho tất cả models
9. RoPE extrapolation lên 1M+ tokens
10. GQA giảm memory requirements đáng kể
11. Multimodal variants available (Llama 3.3 Vision)

3.1.3 Nhược Điểm

1. Resource intensive - 405B cần cluster GPUs
2. Self-hosting complexity cho large models
3. Community support only (no official SLA)
4. Base models cần instruction tuning
5. Fine-tuning costs cao cho large variants
6. Quantization needed cho consumer hardware
7. Documentation fragmented across community
8. Smaller models trade performance for efficiency

3.2 Qwen 2.5 (Alibaba Cloud)

3.2.1 Kiến Trúc Chi Tiết

- **Architecture:** Causal decoder-only Transformer
- **Parameters:** 0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B
- **Variants:** Base, Instruct, Coder, Math
- **Context length:** 32K tokens (base), 128K (extended)
- **Position encoding:** RoPE với YaRN scaling
- **Activation:** SwiGLU

- **Normalization:** RMSNorm
- **Attention:** Multi-head attention với sliding window
- **Training data:** 18+ trillion tokens
- **Languages:** 29 languages (mạnh về tiếng Trung)
- **License:** Apache 2.0
- **Phiên bản:** Qwen 2.5 - released September 2025

3.2.2 Ưu Điểm

1. Exceptional multilingual: Top-tier cho tiếng Trung và châu Á
2. Apache 2.0 - fully permissive commercial use
3. Wide range sizes: 0.5B đến 72B
4. Specialized variants: Coder (93.5% HumanEval), Math
5. Excellent coding performance
6. Strong multilingual support (29 languages)
7. Free và open source
8. Production-ready với optimizations
9. Active development và updates
10. Good documentation (Chinese + English)

3.2.3 Nhược Điểm

1. Western language performance thấp hơn LLaMA
2. Smaller maximum size (72B vs 405B)
3. Less established community vs LLaMA
4. Documentation chủ yếu tiếng Trung
5. Limited research papers
6. Fewer third-party integrations
7. Quantization quality varies
8. Less proven in production

3.3 Mistral Large 2 (Mistral AI)

3.3.1 Kiến Trúc Chi Tiết

- **Architecture:** Sparse Mixture-of-Experts Transformer
- **Model sizes:** 7B, 8x7B (MoE), 8x22B (MoE), 123B
- **Active parameters:** 12.9B (8x7B MoE), 39B (8x22B MoE)
- **Context window:** 128K tokens
- **Position encoding:** RoPE
- **Activation:** SwiGLU
- **MoE:** 8 experts với top-2 routing
- **Attention:** Sliding Window Attention + GQA
- **Training:** Multilingual focus
- **License:** Apache 2.0
- **Phiên bản:** Mistral Large 2 - released July 2025

3.3.2 Ưu Điểm

1. Apache 2.0 - fully commercial-friendly
2. MoE efficiency: Large capacity với reasonable compute
3. Strong multilingual: Excellent European languages
4. Competitive performance: 84.0% MMLU (123B)
5. Fast inference nhờ sparse activation
6. Code-focused variants available
7. Long context: 128K tokens
8. Function calling native support
9. Production-grade quality
10. Active development team

3.3.3 Nhược Điểm

1. Smaller than top competitors (123B vs 405B)
2. MoE complexity trong deployment
3. Limited sizes compared to LLaMA
4. Newer ecosystem - fewer resources
5. Community smaller than LLaMA
6. Documentation gaps
7. Quantization challenges với MoE
8. Less proven at massive scale

3.4 Bảng So Sánh

Bảng 3: So sánh LLMs mã nguồn mở (phiên bản mới nhất 2025)

Tiêu chí	**LLaMA 3.3**	**Qwen 2.5**	**Mistral Large 2**
Architecture	Decoder-only	Decoder-only	Decoder MoE
Parameters	1B-405B	0.5B-72B	7B-123B
Context	128K tokens	32K-128K	128K tokens
Position encoding	RoPE	RoPE+YaRN	RoPE
Languages	8+ languages	29 languages	Multilingual
Training data	15T+ tokens	18T+ tokens	Not disclosed
License	Llama 3.3 License	Apache 2.0	Apache 2.0
MMLU (best)	86.9% (405B)	85.2% (72B)	84.0% (123B)
HumanEval (best)	89.0% (405B)	93.5% (Coder)	88.7% (123B)
Ưu điểm	Scale, Performance	Asian langs, Coding	MoE efficiency
Nhược điểm	Resources	Western langs	Smaller scale
Release	Oct 2025	Sep 2025	Jul 2025

4 Evaluation Metrics

4.1 MMLU (Massive Multitask Language Understanding)

Phù hợp: Đánh giá kiến thức đa lĩnh vực từ toán, khoa học đến nhân văn

$$\text{Accuracy}_{\text{MMLU}} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Questions}}$$

4.2 SuperGLUE

Phù hợp: Đánh giá khả năng hiểu ngôn ngữ và suy luận phức tạp trong NLP

$$\text{Composite Score} = \frac{1}{N}\sum_{i=1}^{N}\text{Score}_i$$

4.3 HellaSwag

Phù hợp: Kiểm tra suy luận thông thường trong các tình huống hàng ngày

$$\text{Accuracy}_{\text{HellaSwag}} = \frac{\text{Correct Completions}}{\text{Total Scenarios}}$$

4.4 ARC (AI2 Reasoning Challenge)

Phù hợp: Đánh giá hiểu biết khoa học và khả năng suy luận logic

$$\text{Accuracy}_{\text{ARC}} = \frac{\text{Correct Answers}}{\text{Total Questions}}$$

4.5 WinoGrande

Phù hợp: Kiểm tra hiểu ngữ cảnh và giải quyết tham chiếu đại từ

$$\text{Accuracy}_{\text{WinoGrande}} = \frac{\text{Correct Entity Selections}}{\text{Total Pronoun Cases}}$$

4.6 NLVR2 (Natural Language for Visual Reasoning)

Phù hợp: Đánh giá mô hình multimodal kết hợp hình ảnh và văn bản

$$\text{Accuracy}_{\text{NLVR2}} = \frac{\text{Correct Validations}}{\text{Total Image-Text Pairs}}$$

4.7 VQA (Visual Question Answering)

Phù hợp: Trả lời câu hỏi dựa trên nội dung hình ảnh

$$\text{Accuracy}_{\text{VQA}} = \frac{\text{Correct Answers}}{\text{Total Questions}}$$

4.8 Perplexity

Phù hợp: Đánh giá chất lượng language model, text generation

$$\text{PPL} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(x_i)\right)$$

4.9 BLEU Score

Phù hợp: Đánh giá chất lượng machine translation và text generation

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

4.10 ROUGE-n

Phù hợp: Đánh giá text summarization và so sánh văn bản

$$\text{ROUGE-n} = \frac{2 \cdot P \cdot R}{P + R}$$

4.11 Classification Metrics

Phù hợp: Các bài toán phân loại văn bản, sentiment analysis, NER

$$\begin{aligned} \text{Accuracy} &= P(X = Y) \\ \text{Precision} &= P(Y = 1 | X = 1) \\ \text{Recall} &= P(X = 1 | Y = 1) \\ \text{F1-Score} &= \frac{2PR}{P + R} \end{aligned}$$

Bài tập: So sánh hai giao thức MCP và A2A

Hai đoạn demo thể hiện sự khác biệt giữa **MCP (Model Context Protocol)** và **A2A (Agent-to-Agent Protocol)** trong việc truy vấn dữ liệu và trao đổi thông tin giữa agents.

Đặc điểm	**MCP Protocol**	**A2A Protocol**
Kiến trúc	Client–Server	Peer–to–Peer
Truy vấn dữ liệu	Server truy vấn DB rồi gửi *context* cho Client	Mỗi Agent tự truy vấn DB trực tiếp
Kiểu giao tiếp	Một chiều (Client → Server)	Hai chiều hoặc đa chiều giữa các Agent
Tốc độ trung bình	3.08s	1.80s
Số lượng trao đổi	11 requests	14 messages
Hiệu năng	Phụ thuộc vào server xử lý context	Nhanh hơn vì agents truy vấn song song

MCP (Model Context Protocol) hoạt động theo mô hình Client–Server. Server chịu trách nhiệm truy vấn dữ liệu và cung cấp *context* cho mô hình ngôn ngữ trước khi phản hồi.

- **Ưu điểm:** Dễ quản lý, tập trung dữ liệu, đảm bảo tính nhất quán.
- **Hạn chế:** Chậm hơn do phụ thuộc vào tầng server trung gian.
- **Ví dụ demo:** Client đặt câu hỏi về thông tin công ty, tài chính, nhân sự → server xử lý và trả kết quả sau 2–4 giây mỗi truy vấn.

A2A (Agent-to-Agent) sử dụng cơ chế Peer-to-Peer. Các agent (Business, Research, Engineer) tự truy vấn cơ sở dữ liệu và trao đổi trực tiếp để cùng xây dựng phản hồi.

- **Ưu điểm:** Phản hồi nhanh, linh hoạt, có thể hợp tác hoặc thương lượng giữa nhiều agent.
- **Hạn chế:** Mỗi agent cần có quyền truy cập dữ liệu và logic riêng, dễ phức tạp khi mở rộng hệ thống.
- **Ví dụ demo:** Các agent cùng trả lời ba câu hỏi giống MCP, nhưng phản hồi nhanh hơn và mang tính hội thoại tự nhiên.

Nhận xét:

- A2A nhanh hơn MCP trung bình khoảng **1.28 giây**.
- MCP phù hợp cho hệ thống cần kiểm soát dữ liệu chặt chẽ và bảo mật cao.
- A2A phù hợp cho môi trường *multi-agent collaboration*, nơi tốc độ và giao tiếp linh hoạt quan trọng hơn.
- Có thể kết hợp cả hai: dùng MCP để quản lý dữ liệu tập trung, và A2A cho giai đoạn giao tiếp giữa các agent.

Kết luận: MCP mang lại tính ổn định và quản trị dữ liệu tập trung, trong khi **A2A** thể hiện sự năng động, tốc độ và khả năng cộng tác cao giữa các agent. Lựa chọn giao thức phù hợp phụ thuộc vào mục tiêu hệ thống — ưu tiên an toàn dữ liệu hay hiệu suất xử lý.