Minor in AI

Advances in LLMs

April 23, 2025

Introduction

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing by leveraging deep neural networks, particularly transformer-based architectures, trained on massive corpora of text. These models have gone from simple next-word predictors to multi-modal, instruction-following systems capable of reasoning and dialogue.

1 Timeline of LLM Advancements

1.1 2018 — The Foundations

The year 2018 saw the emergence of the transformer architecture and the birth of pretraining followed by fine-tuning. This paradigm shift laid the groundwork for all subsequent LLMs.

Pretrain-Finetune Paradigm: Models are first trained on general language tasks using large unlabeled datasets (pretraining), and then refined on specific tasks with labeled data (finetuning). This allows for effective transfer learning.

Masked Language Modeling (MLM): Introduced by BERT, MLM involves randomly masking some tokens in the input and training the model to predict them. This enables bidirectional context understanding.

- **GPT (OpenAI)**: Unidirectional transformer trained to predict the next word (autoregressive).
- **BERT (Google)**: Bidirectional encoder using MLM; strong in understanding tasks.

1.2 2019 — Scaling Begins

As compute and data scaled up, models with more parameters were introduced, setting the stage for emergent behaviors like few-shot learning.

GPT-2 (1.5B parameters): Trained on the WebText dataset, it exhibited surprising zero-shot and few-shot capabilities without task-specific finetuning.

XLNet: Improved over BERT by removing the independence assumption of masked tokens and using a permutation-based objective.

T5 (Text-to-Text Transfer Transformer): Framed all NLP tasks as text-to-text problems (e.g., translation: "Translate English to German: ..."), unifying them under one architecture.

- GPT-2 demonstrated contextual generation in long text.
- XLNet bridged the gap between autoregressive and autoencoding models.
- T5 simplified NLP pipelines through a unified format.

1.3 2020 — Bigger and Smarter

2020 brought exponential growth in model size, culminating in GPT-3 with 175 billion parameters. This enabled powerful few-shot and in-context learning.

In-Context Learning: Rather than fine-tuning weights, the model uses the prompt context to learn tasks on the fly, e.g., providing few examples in the input prompt.

Mixture-of-Experts (MoE): A sparse model design where only a few sub-networks (experts) are activated per input. This allows models to scale to trillions of parameters without a linear increase in compute.

- **GPT-3 (OpenAI):** Few-shot learning without parameter updates.
- **Switch Transformer (Google):** Trillion-parameter MoE model.

1.4 2021 — Specialization and Scaling

LLMs began specializing in different domains like code, and multilingual understanding became a key focus.

Codex: Built on GPT-3, Codex is fine-tuned on billions of lines of code, enabling models like GitHub Copilot to assist with real-time code generation.

Gopher and PaLM: These models focused on scaling both data and parameters (up to 540B), with strong performance on reasoning and multilingual benchmarks.

- **Codex:** Specialized for code — powering developer tools.
- **PaLM (Google):** Dense architecture with superior reasoning ability.

1.5 2022 — Human Feedback Era

This year marked a critical pivot toward alignment and safety using human preferences.

Reinforcement Learning from Human Feedback (RLHF): Instead of only maximizing language modeling likelihood, models are trained with feedback from human evaluators to rank outputs, improving helpfulness and reducing toxicity.

Open Models: Community efforts like BLOOM, OPT, and GLM emerged to ensure transparency and openness in LLM research.

- **ChatGPT (OpenAI):** Uses RLHF for safer, instruction-following chat.
- **BLOOM, OPT, GLM:** Open-source LLMs democratizing access.

1.6 2023 — Multimodal and Instruction-Tuned Models

LLMs moved beyond text to handle vision, audio, and even tool-use via APIs.

Multimodal LLMs: Models like Flamingo, GPT-4, and Qwen-VL take both text and images as inputs, enabling richer interaction and better grounding in reality.

Instruction Tuning: Models are fine-tuned on curated prompts and responses to better follow user instructions.

- **GPT-4, Claude, Gemini:** Multimodal, instruction-tuned, safer LLMs.
- **Vicuna, Alpaca, Falcon, MPT:** Open-sourced, instruction-following chatbots trained on real conversations.

From simple autoregressive predictors to multimodal and instruction-tuned agents, the journey of LLMs is a testament to the power of scaling, alignment, and democratization. As the field matures, future directions include grounding with external tools, improving faithfulness, and incorporating memory and planning.

2 Key Takeaways

1. Understanding the architectural shifts is key to understanding model capabilities.
2. LLMs are not just getting bigger — they're getting smarter and safer.
3. Human alignment (RLHF) is essential for real-world deployment.
4. Open-source initiatives will define the next phase of research.