AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 1: Frontier LLMs

- GPT1, GPT2, GPT3, Codex, InstructGPT, GPT4 papers. Self explanatory. GPT3.5, 4o, o1, and o3 tended to have launch events and system cards instead.
- Claude 3 and Gemini 1 papers to understand the competition. Latest iterations are Claude 3.5 Sonnet and Gemini 2.0 Flash/Flash Thinking. Also Gemma 2.
- LLaMA 1, Llama 2, Llama 3 papers to understand the leading open models. You can also view Mistral 7B, Mixtral and Pixtral as a branch on the Llama family tree.
- DeepSeek V1, Coder, MoE, V2, V3 papers. Leading (relatively) open model lab.
- Apple Intelligence paper. It's on every Mac and iPhone.

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 2: Benchmarks and Evals

- MMLU paper - the main knowledge benchmark, next to GPQA and BIG-Bench. In 2025 frontier labs use MMLU Pro, GPQA Diamond, and BIG-Bench Hard.
- MuSR paper - evaluating long context, next to LongBench, BABILong, and RULER. Solving Lost in The Middle and other issues with Needle in a Haystack.
- MATH paper - a compilation of math competition problems. Frontier labs focus on FrontierMath and hard subsets of MATH: MATH level 5, AIME, AMC10/AMC12.
- IFEval paper - the leading instruction following eval and only external benchmark adopted by Apple. You could also view MT-Bench as a form of IF.
- ARC AGI challenge - a famous abstract reasoning "IQ test" benchmark that has lasted far longer than many quickly saturated benchmarks.

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 3: Prompting, ICL & Chain of Thought

- The Prompt Report paper - a survey of prompting papers
- Chain-of-Thought paper - one of multiple claimants to popularizing Chain of Thought, along with Scratchpads and Let's Think Step By Step
- Tree of Thought paper - introducing lookaheads and backtracking (podcast)
- Prompt Tuning paper - you may not need prompts - if you can do Prefix-Tuning, adjust decoding (say via entropy), or representation engineering
- Automatic Prompt Engineering paper - it is increasingly obvious that humans are terrible zero-shot prompters and prompting itself can be enhanced by LLMs. The most notable implementation of this is in the DSPy paper/framework.

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 4: Retrieval Augmented Generation

- Introduction to Information Retrieval
- 2020 Meta RAG paper - which coined the term. The original authors have started Contextual and have coined RAG 2.0. Modern "table stakes" for RAG — HyDE, chunking, rerankers, multimodal data are better presented elsewhere.
- MTEB paper - known overfitting that its author considers it dead, but still de-facto benchmark. Many embeddings have papers - pick your poison - SentenceTransformers, OpenAI, Nomic Embed, Jina v3, cde-small-v1, ModernBERT Embed - with Matryoshka embeddings increasingly standard.
- GraphRAG paper - Microsoft's take on adding knowledge graphs to RAG, now open sourced. One of the most popular trends in RAG in 2024, alongside of ColBERT/ColPali/ColQwen (more in the Vision section).
- RAGAS paper - the simple RAG eval recommended by OpenAI.

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 5: Agents

- SWE-Bench paper. After adoption by Anthropic, Devin and OpenAI, probably the highest profile agent benchmark today (vs WebArena or SWE-Gym). Technically a coding benchmark, but more a test of agents than raw LLMs. See also SWE-Agent, SWE-Bench Multimodal and the Konwinski Prize.
- ReAct paper- ReAct started a long line of research on tool using and function calling LLMs, including Gorilla and the BFCL Leaderboard. Of historical interest - Toolformer and HuggingGPT.
- MemGPT paper - one of many notable approaches to emulating long running agent memory, adopted by ChatGPT and LangGraph. Versions of these are reinvented in every agent system from MetaGPT to AutoGen to Smallville.
- Voyager paper - Nvidia's take on 3 cognitive architecture components (curriculum, skill library, sandbox) to improve performance. More abstractly, skill library/curriculum can be abstracted as a form of Agent Workflow Memory.
- Anthropic on Building Effective Agents

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 6: Code Generation

- The Stack paper - the original open dataset twin of The Pile focused on code, starting a great lineage of open codegen work from The Stack v2 to StarCoder.
- Open Code Model papers - choose from DeepSeek-Coder, Qwen2.5-Coder, or CodeLlama. Many regard 3.5 Sonnet as the best code model but it has no paper.
- HumanEval/Codex paper - This is a saturated benchmark, but is required knowledge for the code domain. SWE-Bench is more famous for coding now, but is expensive/evals agents rather than models. Modern replacements include Aider, Codeforces, BigCodeBench, LiveCodeBench and SciCode.
- AlphaCodeium paper - Google published AlphaCode and AlphaCode2 which did very well on programming problems, but here is one way Flow Engineering can add a lot more performance to any given base model.
- CriticGPT paper - LLMs are known to generate code that can have security issues. OpenAI trained CriticGPT to spot them, and Anthropic uses SAEs to identify LLM features that cause this, but it is a problem you should be aware of.

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 7: Vision

- Non-LLM Vision work is still important: e.g. the YOLO paper (now up to v11, but mind the lineage), but increasingly transformers like DETRs Beat YOLOs too.
- CLIP paper - the first successful ViT from Alec Radford. These days, superceded by BLIP/BLIP2 or SigLIP/PaliGemma, but still required to know.
- MMVP benchmark (LS Live)- quantifies important issues with CLIP. Multimodal versions of MMLU (MMMU) and SWE-Bench do exist.
- Segment Anything Model and SAM 2 paper (our pod) - the very successful image and video segmentation foundation model. Pair with GroundingDINO.
- Early fusion research: Contra the cheap “late fusion” work like LLaVA (our pod), early fusion covers Meta’s Flamingo, Chameleon,

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 8: Voice

- Whisper paper - the successful ASR model from Alec Radford. Whisper v2, v3 and distil-whisper and v3 Turbo are open weights but have no paper.
- AudioPaLM paper - our last look at Google's voice thoughts before PaLM became Gemini. See also: Meta's Llama 3 explorations into speech.
- NaturalSpeech paper - one of a few leading TTS approaches. Recently v3.
- Kyutai Moshi paper - an impressive full-duplex speech-text open weights model with high profile demo. See also Hume OCTAVE.
- OpenAI Realtime API: The Missing Manual -

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 9: Image/Video Diffusion

- Latent Diffusion paper - effectively the Stable Diffusion paper.
- DALL-E / DALL-E-2 / DALL-E-3 paper - OpenAI's image generation.
- Imagen / Imagen 2 / Imagen 3 paper - Google's image gen. See also Ideogram.
- Consistency Models paper - this distillation work with LCMs spawned the quick draw viral moment of Dec 2023. These days, updated with sCMs.
- Sora blogpost - text to video - no paper of course beyond the DiT paper

Credit: Latent Space

AI Engineer Reading List

Papers/models/blogs across 10 fields in AI

Section 10: Finetuning

- LoRA/QLoRA paper - the de facto way to finetune models cheaply, whether on local models or with 4o (confirmed on pod). FSDP+QLoRA is educational.
- DPO paper - the popular, if slightly inferior, alternative to PPO, now supported by OpenAI as Preference Finetuning.
- ReFT paper - instead of finetuning a few layers, focus on features instead.
- Orca 3/AgentInstruct paper - see the Synthetic Data picks at NeurIPS but this is a great way to get finetue data.
- RL/Reasoning Tuning papers - RL Finetuning for o1 is debated, but Let's Verify Step By Step and Noam Brown's many public talks give hints for how it works.

Credit: Latent Space