**Track 1:**

Required:

GPT-4 Blog Post: [https://openai.com/research/gpt-4](https://openai.com/research/gpt-4" \t "_blank)

Introducing LLaMA: A foundational, 65-billion-parameter large language model: [https://ai.meta.com/blog/large-language-model-llama-meta-ai/](https://ai.meta.com/blog/large-language-model-llama-meta-ai/" \t "_blank)

PaLM-E: An embodied multimodal language model: [https://blog.research.google/2023/03/palm-e-embodied-multimodal-language.html](https://blog.research.google/2023/03/palm-e-embodied-multimodal-language.html" \t "_blank)

Tree of Thoughts: Deliberate Problem Solving with Large Language Models: [https://arxiv.org/abs/2305.10601](https://arxiv.org/abs/2305.10601" \t "_blank)

Optional:

GPT-4 Technical Report: [https://arxiv.org/abs/2303.08774](https://arxiv.org/abs/2303.08774" \t "_blank)

LLaMA: Open and Efficient Foundation Language Models: [https://arxiv.org/abs/2302.13971](https://arxiv.org/abs/2302.13971" \t "_blank)

PaLM-E: An Embodied Multimodal Language Model: [https://arxiv.org/abs/2303.03378](https://arxiv.org/abs/2303.03378" \t "_blank)

**Track 2:**

Required:

A Comprehensive Survey of Compression Algorithms for Language Models: [https://arxiv.org/abs/2401.15347](https://arxiv.org/abs/2401.15347" \t "_blank)

Toolformer: Language Models Can Teach Themselves to Use Tools: [https://arxiv.org/abs/2302.04761](https://arxiv.org/abs/2302.04761" \t "_blank)

LoRA: Low-Rank Adaptation of Large Language Models: [https://arxiv.org/abs/2106.09685](https://arxiv.org/abs/2106.09685" \t "_blank)

QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models: [https://arxiv.org/abs/2309.14717](https://arxiv.org/abs/2309.14717" \t "_blank)

QLoRA: Efficient Finetuning of Quantized LLMs: [https://arxiv.org/abs/2305.14314](https://arxiv.org/abs/2305.14314" \t "_blank)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: [https://arxiv.org/abs/1810.04805](https://arxiv.org/abs/1810.04805" \t "_blank)

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter: [https://arxiv.org/abs/1910.01108](https://arxiv.org/abs/1910.01108" \t "_blank)

Optional:

Toolformer github:

<https://github.com/conceptofmind/toolformer>

Quantization:

Accurate LoRA-Finetuning Quantization of LLMs via Information Retention: [https://arxiv.org/pdf/2402.05445.pdf](https://arxiv.org/pdf/2402.05445.pdf" \t "_blank)

Pruning and Sparsity:

LLM-Pruner: On the Structural Pruning of Large Language Models: <https://arxiv.org/abs/2305.11627>

Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time: <https://proceedings.mlr.press/v202/liu23am.html>

Distillation:

Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes: <https://arxiv.org/abs/2305.02301>

Efficient Prompting:

Did You Read the Instructions? Rethinking the Effectiveness of Task Definitions in Instruction Learning: <https://arxiv.org/abs/2306.01150>

**Track 3:**

Required:

Continual Learning for Large Language Models: A Survey: <https://arxiv.org/html/2402.01364v2>

Towards Continual Knowledge Learning of Language Models: <https://arxiv.org/abs/2110.03215>

Continual Pre-Training Mitigates Forgetting in Language and Vision: <https://arxiv.org/abs/2205.09357>

Lifelong Pretraining: Continually Adapting Language Models to Emerging Corpora: <https://arxiv.org/abs/2110.08534>

Continual Training of Language Models for Few-Shot Learning: <https://arxiv.org/abs/2210.05549>

Continual Pre-training of Language Models: <https://arxiv.org/abs/2302.03241>

**Transformer:**

Attention Is All You Need: <https://arxiv.org/abs/1706.03762>

<https://www.youtube.com/watch?v=ugWDIIOHtPA&list=PLJV_el3uVTsOK_ZK5L0Iv_EQoL1JefRL4&index=61>