Endnotes

1. Model Garden on Vertex AI. Available at: https://cloud.google.com/model-garden

2. Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, William Fedus. 2022. Emergent Abilities of Large Language Models. Available at: https://arxiv.org/pdf/2206.07682.pdf

3. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. 2022. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Available at: https://arxiv.org/pdf/2005.11401.pdf

4. Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao, Department of Computer Science, Princeton University, Google Research, Brain team, REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS. Available at: https://arxiv.org/pdf/2210.03629.pdf

5. Grounding in Vertex AI. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/grounding/ground-language-models

6. Vertex Extensions. Connect models to APIs by using extensions. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/extensions/overview

7. Overview of Vertex AI Vector Search. Available at: https://cloud.google.com/vertex-ai/docs/vector-search/overview

8. What is Vertex AI Agent Builder? Available at: https://cloud.google.com/generative-ai-app-builder/docs/introduction

9. LangChain. Get your LLM application from prototype to production. Available at: https://www.langchain.com/

10. Introduction to the Vertex AI SDK for Python. Available at: https://cloud.google.com/vertex-ai/docs/python-sdk/use-vertex-ai-python-sdk

11. Introduction to Vertex AI. Available at: https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform

12. Introduction to Vertex AI Model Registry. Available at: https://cloud.google.com/vertex-ai/docs/model-registry/introduction

13. Introduction to Vertex AI Pipelines. Available at: https://cloud.google.com/vertex-ai/docs/pipelines/introduction

14. Dataplex. Available at: https://cloud.google.com/dataplex

15. BigQuery. Available at: https://cloud.google.com/bigquery?hl=en

16. PaLi-Gemma model card. Available at: https://ai.google.dev/gemma/docs/paligemma/model-card

17. Version Control. Available at: https://en.wikipedia.org/wiki/Version_control

18. Continuous integration. Available at: https://wikipedia.org/wiki/Continuous_integration

19. TFX is an end-to-end platform for deploying production ML pipelines. Available at: https://www.tensorflow.org/tfx

20. Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, Tomas Pfister. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. Available at: https://arxiv.org/pdf/2305.02301.pdf

21. Vertex Endpoints. Use private endpoints for online prediction. Available at: https://cloud.google.com/vertex-ai/docs/predictions/using-private-endpoints

22. Tuan Duong Nguyen, Marthinus Christoffel du Plessis, Takafumi Kanamori, Masashi Sugiyama, 2014. Constrained Least-Squares Density-Difference Estimation. Available at: https://www.ms.k.u-tokyo.ac.jp/sugi/2014/CLSDD.pdf

23. Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, Alexander Smola, 2012. A Kernel Two-Sample Test. Available at: https://jmlr.csail.mit.edu/papers/v13/gretton12a.html

24. Oliver Cobb, Arnaud Van Looveren, 2022. Context-Aware Drift Detection. Available at: https://arxiv.org/pdf/2203.08644.pdf

25. Google Gemma Model. Available at: https://gemini.google.com/

26. Perform metrics-based evaluation. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/models/evaluate-models

27. Gemini Team, Google, 2023. Gemini: A Family of Highly Capable Multimodal Models. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

28. Anil, Dai et al., 2023. PaLM 2 Technical Report. Available at: https://arxiv.org/abs/2305.10403

29. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, Mohammad Norouzi, 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Available at: https://arxiv.org/abs/2205.11487

30. Build the future of AI with Meta Llama 3. Available at: https://llama.meta.com/llama3

31. Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. Available at: https://arxiv.org/abs/2210.11416

32. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://arxiv.org/abs/1810.04805

33. Stable Diffusion. Available at: https://github.com/CompVis/stable-diffusion

34. Vertex AI Function Calling. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/function-calling

35. Mistral AI. Available at: https://mistral.ai/

36. Models available in Model Garden. Available at: https://cloud.google.com/vertex-ai/docs/start/explore-models#available-models

37. Vertex AI Studio. Customize and deploy generative models. Available at: https://cloud.google.com/generative-ai-studio

38. vLLM. Easy, fast, and cheap LLM serving for everyone. Available at: https://github.com/vllm-project/vllm

39. Overview of multimodal models. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/multimodal/overview

40. Text models. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text

41. Imagen on Vertex AI | AI Image Generator. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/image/overview

42. Code models overview. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/code/code-models-overview

43. Convert speech to text. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/speech/speech-to-text

44. Text-to-Speech AI. Available at: https://cloud.google.com/text-to-speech

45. Natural Language AI. Available at: https://cloud.google.com/natural-language

46. Translate docs, audio, and videos in real time with Google AI. Available at: https://cloud.google.com/translate

47. Vision AI. Available at: https://cloud.google.com/vision

48. Git. Available at: https://git-scm.com/

49. CodeGemma model card. Available at: https://ai.google.dev/gemma/docs/codegemma/model_card

50. TII's Falcon. Available at: https://falconllm.tii.ae/

51. Mistral AI. Available at: https://mistral.ai/

52. Hugging Face, 2024. Vision Transformer (ViT) Documentation. Hugging Face, [online] Available at: https://huggingface.co/docs/transformers/en/model_doc/vit

53. Mingxing Tan, Quoc V. Le, 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Available at: https://arxiv.org/abs/1905.11946

54. Anthropic Claude 3. Available at: https://www.anthropic.com/news/claude-3-haiku

55. Anthropic Claude 3 on Google Cloud Model Garden. Available at: https://cloud.google.com/blog/products/ai-machine-learning/announcing-anthropics-claude-3-models-in-google-cloud-vertex-ai

56. Vertex AI API. Available at: https://cloud.google.com/vertex-ai/docs/reference/rest

57. Vertex AI: Python SDK. Available at: https://cloud.google.com/python/docs/reference/aiplatform/latest/vertexai

58. Vertex AI: Node.js Client. Available at: https://cloud.google.com/nodejs/docs/reference/aiplatform/latest/overview

59. Vertex AI for Java. Available at: https://cloud.google.com/java/docs/reference/google-cloud-aiplatform/latest/overview

60. Customize and deploy generative models. Available at: https://cloud.google.com/generative-ai-studio

61. Design text prompts. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/text/text-prompts

62. Introduction to prompt design. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/learn/introduction-prompt-design

63. Supervised tuning. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/models/tune-models#supervised-tuning

64. RLHF model tuning. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune-text-models-rlhf

65. Vertex AI Distilation. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune-text-models

66. Create distilled text models. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/models/distill-text-models

67. Pipeline Basics. Available at: https://www.kubeflow.org/docs/components/pipelines/v2/pipelines/pipeline-basics/

68. Build a pipeline. Available at: https://cloud.google.com/vertex-ai/docs/pipelines/build-pipeline

69. Vertex AI Search extension. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/extensions/vertex-ai-search

70. What is Vertex AI Agent Builder? Available at: https://cloud.google.com/generative-ai-app-builder/docs/introduction

71. Generative AI on Vertex AI, Citation Check. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/learn/overview#citation_check

72. Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar, 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. Available at: https://arxiv.org/pdf/1908.10396.pdf

73. Get text embeddings. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/embeddings/get-text-embeddings

74. About Vertex AI Feature Store. Available at: https://cloud.google.com/vertex-ai/docs/featurestore/latest/overview

75. Google Cloud Vertex AI. Available at: https://python.langchain.com/docs/integrations/llms/google_vertex_ai_palm

76. Generative AI - Language - LangChain. Available at: https://github.com/GoogleCloudPlatform/generative-ai/tree/main/language/orchestration/langchain

77. Introduction to Vertex AI Workbench, Workbench Instances. Available at: https://cloud.google.com/vertex-ai/docs/workbench/introduction

78. Introduction to Colab Enterprise. Available at: https://cloud.google.com/colab/docs/introduction

79. Introduction to Vertex AI Experiments. Available at: https://cloud.google.com/vertex-ai/docs/experiments/intro-vertex-ai-experiments

80. Vertex AI TensorBoard Introduction to Vertex AI TensorBoard. Available at https://cloud.google.com/vertex-ai/docs/experiments/tensorboard-introduction

81. Perform metrics-based evaluation. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/models/evaluate-models

82. Perform automatic side-by-side evaluation. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/models/side-by-side-eval

83. Rapid Evaluation Vertex AI. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/rapid-evaluation

84. Citation metadata. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/learn/responsible-ai#citation_metadata

85. Responsible AI. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/learn/responsible-ai#filters-palm-api

86. Imagen on Vertex AI | AI Image Generator. Available at: https://cloud.google.com/vertex-ai/docs/generative-ai/image/overview

87. SynthID. Identifying AI-generated content with SynthID. Available at: https://deepmind.google/technologies/synthid/

88. Moderate text. Available at: https://cloud.google.com/natural-language/docs/moderating-text

89. Model bias metrics for Vertex AI. Available at: https://cloud.google.com/vertex-ai/docs/evaluation/model-bias-metrics

90. Model evaluation in Vertex AI. Available at: https://cloud.google.com/vertex-ai/docs/evaluation/introduction

91. Introduction to Vertex AI Model Monitoring. Available at: https://cloud.google.com/vertex-ai/docs/model-monitoring/overview

92. Identity and Access Management (IAM). Available at: https://cloud.google.com/iam/docs

93. Agents. Available at: https://www.kaggle.com/whitepaper-agents