Advanced Prompt Engineering and Multi-modal AI

Advanced Prompt Engineering

Prompt Engineering Guide | Prompt Engineering Guide Master Prompt Engineering and building AI Agents in our NEW courses! Use PROMPTING20 for 20% off Enroll nowPrompt EngineeringIntroductionLLM SettingsBasics of PromptingPrompt ElementsGeneral Tips for Designing PromptsExamples of PromptsPrompting TechniquesZero-shot PromptingFew-shot PromptingChain-of-Thought PromptingMeta PromptingSelf-ConsistencyGenerate Knowledge PromptingPrompt ChainingTree of ThoughtsRetrieval Augmented GenerationAutomatic Reasoning and Tool-useAutomatic Prompt EngineerActive-PromptDirectional Stimulus PromptingProgram-Aided Language ModelsReActReflexionMultimodal CoTGraph PromptingAgentsIntroduction to AgentsAgent ComponentsGuidesOptimizing PromptsOpenAI Deep ResearchReasoning LLMs4o Image GenerationContext Engineering GuideApplicationsFine-tuning GPT-4oFunction CallingContext Caching with LLMsGenerating DataGenerating Synthetic Dataset for RAGTackling Generated Datasets DiversityGenerating CodeGraduate Job Classification Case StudyPrompt FunctionPrompt HubClassificationSentiment ClassificationFew-Shot Sentiment ClassificationCodingGenerate Code SnippetGenerate MySQL QueryDraw TiKZ DiagramCreativityRhymesInfinite PrimesInterdisciplinaryInventing New WordsEvaluationEvaluate Plato's DialogueInformation ExtractionExtract Model NamesImage GenerationDraw a Person Using AlphabetMathematicsEvaluating Composite FunctionsAdding Odd NumbersQuestion AnsweringClosed Domain Question AnsweringOpen Domain Question AnsweringScience Question AnsweringReasoningIndirect ReasoningPhysical ReasoningText SummarizationExplain A ConceptTruthfulnessHallucination IdentificationAdversarial PromptingPrompt InjectionPrompt LeakingJailbreakingModelsChatGPTClaude 3Code LlamaFlanGeminiGemini AdvancedGemini 1.5 ProGemmaGPT-4Grok-1LLaMALlama 3Mistral 7BMistral LargeMixtralMixtral 8x22BOLMoPhi-2SoraLLM CollectionRisks & MisusesAdversarial PromptingFactualityBiasesLLM Research FindingsLLM AgentsRAG for LLMsLLM ReasoningRAG FaithfulnessLLM In-Context RecallRAG Reduces HallucinationSynthetic DataThoughtSculptInfini-AttentionLM-Guided CoTTrustworthiness in LLMsLLM TokenizationWhat is Groq?PapersToolsNotebooksDatasetsAdditional ReadingsEnglishLightPrompt EngineeringPrompt Engineering Guide Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics. Prompt engineering skills help to better understand the capabilities and limitations of large language models (LLMs). Researchers use prompt engineering to improve the capacity of LLMs on a wide range of common and complex tasks such as question answering and arithmetic reasoning. Developers use prompt engineering to design robust and effective prompting techniques that interface with LLMs and other tools. Prompt engineering is not just about designing and developing prompts. It encompasses a wide range of skills and techniques that are useful for interacting and developing with LLMs. It's an important skill to interface, build with, and understand capabilities of LLMs. You can use prompt engineering to improve safety of LLMs and build new capabilities like augmenting LLMs with domain knowledge and external tools. Motivated

by the high interest in developing with LLMs, we have created this new prompt engineering guide that contains all the latest papers, advanced prompting techniques, learning guides, model-specific prompting guides, lectures, references, new LLM capabilities, and tools related to prompt engineering. Want to learn more? Learn more about advanced prompt engineering techniques and best practices in our new AI courses. Join now! (opens in a new tab) Use code PROMPTING20 to get an extra 20% off.Introduction Advanced Prompting Techniques: - Chain-of-Thought: Step-by-step reasoning - Tree of Thoughts: Exploring multiple reasoning paths - Self-consistency: Multiple reasoning paths with voting - Program-aided Language Models: Code generation for reasoning Prompt Design Patterns: - Few-shot learning with diverse examples - Role-playing and persona prompts - Template-based prompt construction - Dynamic prompt generation and optimization Evaluation and Optimization: - A/B testing for prompt effectiveness - Automated prompt optimization techniques - Performance metrics and benchmarking - Cost optimization strategies

Multi-modal AI Systems

Multimodal learning - Wikipedia Jump to content From Wikipedia, the free encyclopedia Machine learning methods using multiple input modalities Part of a series onMachine learningand data mining Paradigms Supervised learning Unsupervised learning Semi-supervised learning Self-supervised learning Reinforcement learning Meta-learning Online learning Batch learning Curriculum learning Rule-based learning Neuro-symbolic AI Neuromorphic engineering Quantum machine learning Problems Classification Generative modeling Regression Clustering Dimensionality reduction Density estimation Anomaly detection Data cleaning AutoML Association rules Semantic analysis Structured prediction Feature engineering Feature learning Learning to rank Grammar induction Ontology learning Multimodal learning Supervised learning(classification * regression) Apprenticeship learning Decision trees Ensembles Bagging Boosting Random forest k-NN Linear regression Naive Bayes Artificial neural networks Logistic regression Perceptron Relevance vector machine (RVM) Support vector machine (SVM) Clustering BIRCH CURE Hierarchical k-means Fuzzy Expectation-maximization (EM) DBSCAN OPTICS Mean shift Dimensionality reduction Factor analysis CCA ICA LDA NMF PCA PGD t-SNE SDL Structured prediction Graphical models Bayes net Conditional random field Hidden Markov Anomaly detection RANSAC k-NN Local outlier factor Isolation forest Neural networks Autoencoder Deep learning Feedforward neural network Recurrent neural network LSTM GRU ESN reservoir computing Boltzmann machine Restricted GAN Diffusion model SOM Convolutional neural network U-Net LeNet AlexNet DeepDream Neural field Neural radiance field Physics-informed neural networks Transformer Vision Mamba Spiking neural network Memtransistor Electrochemical RAM (ECRAM) Reinforcement learning Q-learning Policy gradient SARSA Temporal difference (TD) Multi-agent Self-play Learning with humans Active learning Crowdsourcing Human-in-the-loop Mechanistic interpretability RLHF Model diagnostics Coefficient of determination Confusion matrix Learning curve ROC curve Mathematical foundations Kernel machines Bias-variance tradeoff Computational learning theory Empirical risk minimization Occam learning PAC learning Statistical learning VC theory Topological deep learning

Journals and conferences AAAI ECML PKDD NeurIPS ICML ICLR IJCAI ML JMLR Related articles Glossary of artificial intelligence List of datasets for machine-learning research List of datasets in computer vision and image processing Outline of machine learning vte

Multimodal learning is a type of deep learning that integrates and processes multiple types of data, referred to as modalities, such as text, audio, images, or video. This integration allows for a more holistic understanding of complex data, improving model performance in tasks like visual question answering, cross-modal retrieval,[1] text-to-image generation,[2] aesthetic ranking,[3] and image captioning.[4] Large multimodal models, such as Google Gemini and GPT-4o, have become increasingly popular since 2023, enabling increased versatility and a broader understanding of real-world phenomena.[5] Motivation[edit] Data usually comes with different modalities which carry different information. For example, it is very common to caption an image to convey the information not presented in the image itself. Similarly, sometimes it is more straightforward to use an image to describe information which may not be obvious from text. As a result, if different words appear in similar images, then these words likely describe the same thing. Conversely, if a word is used to describe seemingly dissimilar images, then these images may represent the same object. Thus, in cases dealing with multi-modal data, it is important to use a model which is able to jointly represent the information such that the model can capture the combined information from different modalities. Multimodal transformers[edit] This section is an excerpt from Transformer (deep learning architecture) Multimodality.[edit] Transformers can also be used/adapted for modalities (input or output) beyond just text, usually by finding a way to "tokenize" the modality. Multimodal models can either be trained from scratch, or by finetuning. A 2022 study found that Transformers pretrained only on natural language can be finetuned on only 0.03% of parameters and become competitive with LSTMs on a variety of logical and visual tasks, demonstrating transfer learning.[6] The LLaVA was a vision-language model composed of a language model (Vicuna-13B)[7] and a vision model (ViT-L/14), connected by a linear layer. Only the linear layer is finetuned.[8] Vision transformers[9] adapt the transformer to computer vision by breaking down input images as a series of patches, turning them into vectors, and treating them like tokens in a standard transformer. Conformer[10] and later Whisper[11] follow the same pattern for speech recognition, first turning the speech signal into a spectrogram, which is then treated like an image, i.e. broken down into a series of patches, turned into vectors and treated like tokens in a standard transformer. Perceivers[12][13] are a variant of Transformers designed for multimodality. For image generation, notable architectures are DALL-E 1 (2021), Parti (2022),[14] Phenaki (2023),[15] and Muse (2023).[16] Unlike later models, DALL-E is not a diffusion model. Instead, it uses a decoder-only Transformer that autoregressively generates a text, followed by the token representation of an image, which is then converted by a variational autoencoder to an image.[17] Parti is an encoder-decoder Transformer, where the encoder processes a text prompt, and the decoder generates a token representation of an image.[18] Muse is an encoder-only Transformer that is trained to predict masked image tokens from unmasked image tokens. During generation, all input tokens are masked, and the highest-confidence predictions are included for the next iteration, until all tokens are predicted.[16] Phenaki is a text-to-video model. It is a

bidirectional masked transformer conditioned on pre-computed text tokens. The generated tokens are then decoded to a video.[15] Multimodal large language models[edit] This section is an excerpt from Large language model Multimodality.[edit] Multimodality means having multiple modalities, where a "modality" refers to a type of input or output, such as video, image, audio, text, proprioception, etc.[19] For example, Google PaLM model was fine-tuned into a multimodal model and applied to robotic control.[20] LLaMA models have also been turned multimodal using the tokenization method, to allow image inputs,[21] and video inputs.[22] GPT-4o can process and generate text, audio and images.[23] Such models are sometimes called large multimodal models (LMMs).[24] A common method to create multimodal models out of an LLM is to "tokenize" the output of a trained encoder. Concretely, one can construct an LLM that can understand images as follows: take a trained LLM, and take a trained image encoder E {\displaystyle E} . Make a small multilayered perceptron f {\displaystyle f} , so that for any image y {\displaystyle y} , the post-processed vector f (E (y)) {\displaystyle f(E(y))} has the same dimensions as an encoded token. That is an "image token". Then, one can interleave text tokens and image tokens. The compound model is then fine-tuned on an image-text dataset. This basic construction can be applied with more sophistication to improve the model. The image encoder may be frozen to improve stability.[25] The model Flamingo demonstrated in 2022 the effectiveness of the tokenization method, fine-tuning a pair of pretrained language model and image encoder to perform better on visual question answering than models trained from scratch.[26] Multimodal deep Boltzmann machines[edit] A Boltzmann machine is a type of stochastic neural network invented by Geoffrey Hinton and Terry Sejnowski in 1985. Boltzmann machines can be seen as the stochastic, generative counterpart of Hopfield nets. They are named after the Boltzmann distribution in statistical mechanics. The units in Boltzmann machines are divided into two groups: visible units and hidden units. Each unit is like a neuron with a binary output that represents whether it is activated or not.[27] General Boltzmann machines allow connection between any units. However, learning is impractical using general Boltzmann Machines because the computational time is exponential to the size of the machine[citation needed]. A more efficient architecture is called restricted Boltzmann machine where connection is only allowed between hidden unit and visible unit, which is described in the next section. Multimodal deep Boltzmann machines can process and learn from different types of information, such as images and text, simultaneously. This can notably be done by having a separate deep Boltzmann machine for each modality, for example one for images and one for text, joined at an additional top hidden layer.[28] Applications[edit] Multimodal machine learning has numerous applications across various domains: Cross-modal retrieval: cross-modal retrieval allows users to search for data across different modalities (e.g., retrieving images based on text descriptions), improving multimedia search engines and content recommendation systems. Models like CLIP facilitate efficient, accurate retrieval by embedding data in a shared space, demonstrating strong performance even in zero-shot settings.[29] Classification and missing data retrieval: multimodal Deep Boltzmann Machines outperform traditional models like support vector machines and latent Dirichlet allocation in classification tasks and can predict missing data in multimodal datasets, such as images and text. Healthcare diagnostics: multimodal models integrate medical

imaging, genomic data, and patient records to improve diagnostic accuracy and early disease detection, especially in cancer screening.[30][31][32] Content generation: models like DALLE generate images from textual descriptions, benefiting creative industries, while cross-modal retrieval enables dynamic multimedia searches.[33] Robotics and human-computer interaction: multimodal learning improves interaction in robotics and AI by integrating sensory inputs like speech, vision, and touch, aiding autonomous systems and human-computer interaction. Emotion recognition: combining visual, audio, and text data, multimodal systems enhance sentiment analysis and emotion recognition, applied in customer service, social media, and marketing. See also[edit] Hopfield network Markov random field Markov chain Monte Carlo References[edit] ^ Hendriksen, Mariya; Bleeker, Maurits; Vakulenko, Svitlana; van Noord, Nanne; Kuiper, Ernst; de Rijke, Maarten (2021). "Extending CLIP for Category-to-image Retrieval in E-commerce". arXiv:2112.11294 [cs.CV]. ^ "Stable Diffusion Repository on GitHub". CompVis - Machine Vision and Learning Research Group, LMU Munich. 17 September 2022. Archived from the original on January 18, 2023. Retrieved 17 September 2022. ^ LAION-AI/aesthetic-predictor, LAION AI, 2024-09-06, retrieved 2024-09-08 ^ Mokady, Ron; Hertz, Amir; Bermano, Amit H. (2021). "ClipCap: CLIP Prefix for Image Captioning". arXiv:2111.09734 [cs.CV]. ^ Zia, Tehseen (January 8, 2024). "Unveiling of Large Multimodal Models: Shaping the Landscape of Language Models in 2024". Unite.ai. Retrieved 2024-06-01. ^ Lu, Kevin; Grover, Aditya; Abbeel, Pieter; Mordatch, Igor (2022-06-28). "Frozen Pretrained Transformers as Universal Computation Engines". Proceedings of the AAAI Conference on Artificial Intelligence. 36 (7): 7628-7636. doi:10.1609/aaai.v36i7.20729. ISSN 2374-3468. ^ "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality | LMSYS Org". lmsys.org. Retrieved 2024-08-11. ^ Liu, Haotian; Li, Chunyuan; Wu, Qingyang; Lee, Yong Jae (2023-12-15). "Visual Instruction Tuning". Advances in Neural Information Processing Systems. 36: 34892-34916. ^ Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob (2021-06-03). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". arXiv:2010.11929 [cs.CV]. ^ Gulati, Anmol; Qin, James; Chiu, Chung-Cheng; Parmar, Niki; Zhang, Yu; Yu, Jiahui; Han, Wei; Wang, Shibo; Zhang, Zhengdong; Wu, Yonghui; Pang, Ruoming (2020). "Conformer: Convolution-augmented Transformer for Speech Recognition". arXiv:2005.08100 [eess.AS]. ^ Radford, Alec; Kim, Jong Wook; Xu, Tao; Brockman, Greg; McLeavey, Christine; Sutskever, Ilya (2022). "Robust Speech Recognition via Large-Scale Weak Supervision". arXiv:2212.04356 [eess.AS]. ^ Jaegle, Andrew; Gimeno, Felix; Brock, Andrew; Zisserman, Andrew; Vinyals, Oriol; Carreira, Joao (2021-06-22). "Perceiver: General Perception with Iterative Attention". arXiv:2103.03206 [cs.CV]. ^ Jaegle, Andrew; Borgeaud, Sebastian; Alayrac, Jean-Baptiste; Doersch, Carl; Ionescu, Catalin; Ding, David; Koppula, Skanda; Zoran, Daniel; Brock, Andrew; Shelhamer, Evan; Henaff, Olivier (2021-08-02). "Perceiver IO: A General Architecture for Structured Inputs & Outputs". arXiv:2107.14795 [cs.LG]. ^ "Parti: Pathways Autoregressive Text-to-Image Model". sites.research.google. Retrieved 2024-08-09. ^ a b Villegas, Ruben; Babaeizadeh, Mohammad; Kindermans, Pieter-Jan; Moraldo, Hernan; Zhang, Han; Saffar, Mohammad Taghi; Castro, Santiago; Kunze, Julius; Erhan, Dumitru (2022-09-29). "Phenaki: Variable Length Video Generation from Open Domain

Textual Descriptions". {{cite journal}}: Cite journal requires |journal= (help) ^ a b Chang, Huiwen; Zhang, Han; Barber, Jarred; Maschinot, A. J.; Lezama, Jose; Jiang, Lu; Yang, Ming-Hsuan; Murphy, Kevin; Freeman, William T. (2023-01-02). "Muse: Text-To-Image Generation via Masked Generative Transformers". arXiv:2301.00704 [cs.CV]. ^ Ramesh, Aditya; Pavlov, Mikhail; Goh, Gabriel; Gray, Scott; Voss, Chelsea; Radford, Alec; Chen, Mark; Sutskever, Ilya (2021-02-26), Zero-Shot Text-to-Image Generation, arXiv:2102.12092 ^ Yu, Jiahui; Xu, Yuanzhong; Koh, Jing Yu; Luong, Thang; Baid, Gunjan; Wang, Zirui; Vasudevan, Vijay; Ku, Alexander; Yang, Yinfei (2022-06-21), Scaling Autoregressive Models for Content-Rich Text-to-Image Generation, arXiv:2206.10789 ^ Kiros, Ryan; Salakhutdinov, Ruslan; Zemel, Rich (2014-06-18). "Multimodal Neural Language Models". Proceedings of the 31st International Conference on Machine Learning. PMLR: 595-603. Archived from the original on 2023-07-02. Retrieved 2023-07-02. ^ Driess, Danny; Xia, Fei; Sajjadi, Mehdi S. M.; Lynch, Corey; Chowdhery, Aakanksha; Ichter, Brian; Wahid, Ayzaan; Tompson, Jonathan; Vuong, Quan; Yu, Tianhe; Huang, Wenlong; Chebotar, Yevgen; Sermanet, Pierre; Duckworth, Daniel; Levine, Sergey (2023-03-01). "PaLM-E: An Embodied Multimodal Language Model". arXiv:2303.03378 [cs.LG]. ^ Liu, Haotian; Li, Chunyuan; Wu, Qingyang; Lee, Yong Jae (2023-04-01). "Visual Instruction Tuning". arXiv:2304.08485 [cs.CV]. ^ Zhang, Hang; Li, Xin; Bing, Lidong (2023-06-01). "Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding". arXiv:2306.02858 [cs.CL]. ^ "OpenAI says natively multimodal GPT-4o eats text, visuals, sound - and emits the same". The Register. 2024-05-13. ^ Zia, Dr Tehseen (2024-01-08). "Unveiling of Large Multimodal Models: Shaping the Landscape of Language Models in 2024". Unite.AI. Retrieved 2025-05-30. ^ Li, Junnan; Li, Dongxu; Savarese, Silvio; Hoi, Steven (2023-01-01). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models". arXiv:2301.12597 [cs.CV]. ^ Alayrac, Jean-Baptiste; Donahue, Jeff; Luc, Pauline; Miech, Antoine; Barr, Iain; Hasson, Yana; Lenc, Karel; Mensch, Arthur; Millican, Katherine; Reynolds, Malcolm; Ring, Roman; Rutherford, Eliza; Cabi, Serkan; Han, Tengda; Gong, Zhitao (2022-12-06). "Flamingo: a Visual Language Model for Few-Shot Learning". Advances in Neural Information Processing Systems. 35: 23716-23736. arXiv:2204.14198. Archived from the original on 2023-07-02. Retrieved 2023-07-02. ^ Dey, Victor (2021-09-03). "Beginners Guide to Boltzmann Machine". Analytics India Magazine. Retrieved 2024-03-02. ^ "Multimodal Learning with Deep Boltzmann Machine" (PDF). 2014. Archived (PDF) from the original on 2015-06-21. Retrieved 2015-06-14. ^ Hendriksen, Mariya; Vakulenko, Svitlana; Kuiper, Ernst; de Rijke, Maarten (2023). "Scene-centric vs. Object-centric Image-Text Cross-modal Retrieval: A Reproducibility Study". arXiv:2301.05174 [cs.CV]. ^ Quach, Katyanna. "Harvard boffins build multimodal AI system to predict cancer". The Register. Archived from the original on 20 September 2022. Retrieved 16 September 2022. ^ Chen, Richard J.; Lu, Ming Y.; Williamson, Drew F. K.; Chen, Tiffany Y.; Lipkova, Jana; Noor, Zahra; Shaban, Muhammad; Shady, Maha; Williams, Mane; Joo, Bumjin; Mahmood, Faisal (8 August 2022). "Pan-cancer integrative histology-genomic analysis via multimodal deep learning". Cancer Cell. 40 (8): 865-878.e6. doi:10.1016/j.ccell.2022.07.004. ISSN 1535-6108. PMC 10397370. PMID 35944502. S2CID 251456162. Teaching hospital press release: "New AI technology integrates multiple data types to predict cancer outcomes". Brigham and Women's Hospital via medicalxpress.com. Archived from the original on 20 September

2022. Retrieved 18 September 2022. ^ Shi, Yuge; Siddharth, N.; Paige, Brooks; Torr, Philip HS (2019). "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". arXiv:1911.03393 [cs.LG]. ^ Shi, Yuge; Siddharth, N.; Paige, Brooks; Torr, Philip HS (2019). "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". arXiv:1911.03393 [cs.LG]. Retrieved from "https://en.wikipedia.org/w/index.php?title=Multimodal_learning&oldid=1293473618" Categories: Artificial neural networksMultimodal interactionHidden categories: CS1 errors: missing periodicalArticles with short descriptionShort description is different from WikidataArticles with excerptsAll articles with unsourced statementsArticles with unsourced statements from November 2022 Search Search Multimodal learning 8 languages Add topic Multi-modal AI Systems: - Vision-Language Models: GPT-4V, CLIP, DALL-E - Audio-Language Integration: Whisper, speech synthesis - Video Understanding: Temporal reasoning and analysis - Cross-modal retrieval and generation Applications: - Image captioning and visual question answering - Document analysis and OCR enhancement - Video content analysis and summarization - Multi-modal search and recommendation systems Technical Challenges: - Modal alignment and fusion strategies - Computational efficiency and optimization - Data quality and annotation requirements - Evaluation across multiple modalities