Note: This reading list is tentative and subject to change, especially for later lectures. For paper discussions (even lecture numbers), you will read only one paper assigned to your group! Dates without **"Paper discussions"** (odd lecturer numbers) have papers that will be referenced in the lecture slides and you are not obliged to read them. So, **although there are many papers listed here, you will be evaluated for reading only 11-12 of them throughout the entire semester.**

Introduction

1. 8/23: **Lecture references.** Motivation, course overview, and requirements; examples of projects.
    - Project ideas: link.
    - EMNLP 2020 tutorial: Interpreting Predictions of NLP Models. Link.
    - NAACL 2022 Tutorial on Human-centered Evaluations of Explanations. Link.
    - Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018. Link.
    - Ribeiro et al. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016. Link. [LIME]
    - Li et al. Understanding Neural Networks through Representation Erasure. Arxiv 2017. Link. [leave-one-out]
    - Zellers et al. From Recognition to Cognition: Visual Commonsense Reasoning. CVPR 2019. Link.
    - Aggarwal et al. Explanations for CommonsenseQA: New Dataset and Models. ACL 2021. Link.
    - Wiegreffe and Marasović. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. NeurIPS 2021. Link.
    - Jay Alammar. The Illustrated Transformer. Link.
    - Koh and Liang. Understanding Black-box Predictions via Influence Functions. ICML 2017. Link.
    - Yeh et al. Representer Point Selection for Explaining Deep Neural Networks. NeurIPS 2018. Link.
    - Miller. Explanation in artificial intelligence: Insights from the social sciences. AIJ 2019. Link.
    - Yang et al. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. COLING 2020. Link.
    - Jacovi and Goldberg. Aligning Faithful Interpretations with their Social Attribution. TACL 2021. Link.
    - Chen et al. KACE: Generating Knowledge-Aware Contrastive Explanations for NLI. ACL 2021. Link.
    - Ross et al. Explaining NLP Models via Minimal Contrastive Editing (MiCE). Findings of ACL 2021. Link.
    - Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. Findings of ACL 2021. Link.

- Wu et al. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. ACL 2021. Link.
- Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021. Link.
- Ribera and Lapedriza. Can we do better explanations? A proposal of user-centered explainable AI. IUI Workshops 2019. Link.

Background

2. 8/25: **Lecture references.** Background: Transformer. Pretraining-finetuning. Data artifacts.
    - Vaswani et al. Attention Is All You Need. NeurIPS 2017. Link.
    - The Illustrated Transformer. Link.
    - Richter and Wattenhofer. Normalized Attention Without Probability Cage. Arxiv 2020. Link.
    - Kovaleva et al. Revealing the Dark Secrets of BERT. EMNLP 2019. Link.
    - Marasović. BERT-base forward pass. Link.
    - Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL 2018. Link.
    - Sequence Models by Noah A. Smith. Link.
    - Bowman. The Dangers of Underclaiming: Reasons for Caution When Reporting How NLP Systems Fail. ACL 2022. Link.
    - Liu et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Arxiv 2019. Link.
    - Radford et al. Language Models are Unsupervised Multitask Learners. 2018. Link.
    - Brown et al. Language Models are Few-Shot Learners. NeurIPS 2020. Link.
    - Raffel et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR 2020. Link.
    - Jia and Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP 2017. Link.
    - Gururangan*, Swayamdipta*, et al. Annotation Artifacts in Natural Language Inference Data. NAACL 2018. Link.
    - Agrawal et al. Analyzing the Behavior of Visual Question Answering Models. EMNLP 2016. Link.
    - Pezeshkpour et al. Combining Feature and Instance Attribution to Detect Artifacts. ACL Findings 2022. Link.

Which part of the input led to a prediction?

3. 8/30: **Lecture references.** Gradient-based post-hoc explanations.
    - Wang et al. Non-local Neural Networks. CVPR 2018. Link.
    - Dosvitskoy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021. Link.
    - Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018. Link.

- Ribeiro et al. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016. Link. [LIME]
- Li et al. Understanding Neural Networks through Representation Erasure. Arxiv 2017. Link. [leave-one-out]
- Simonyan et al. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ICLR Workshop, 2014. Link.
- Han et al. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. ACL 2020. Link.
- Shrikumar et al. Learning Important Features Through Propagating Activation Differences. ICML 2017. Link. [DeepLIFT]
- Bach et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. Plos One 2015. Link. [LRP]
- Selvaraju et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. IJCV 2019. Link. [Grad-CAM]
- Smilkov et al. SmoothGrad: removing noise by adding noise. ICML Workshop on Visualization for Deep Learning, 2017. Link.
- Sundararajan et al. Axiomatic Attribution for Deep Networks. ICML 2017. Link.

4. 9/1: Gradient-based post-hoc explanations. **Paper discussions.**
    - Group 1**:** Hooker et al. A Benchmark for Interpretability Methods in Deep Neural Networks. NeurIPS 2019. Link.
    - Group 2: Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS 2018. Link.
    - Group 3: Wang et al. Gradient-based Analysis of NLP Models is Manipulable. EMNLP Findings 2020. Link.
    - Group 4: Pruthi et al. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? TACL 2022. Link.

5. 9/6: Attention-based post-hoc explanations.
    - Olah. Understanding LSTM Networks. Link.
    - Bahdanau et al. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015. Link.
    - Yang et al. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. Arxiv 2019. Link.
    - Bibal et al. Is Attention Explanation? An Introduction to the Debate. ACL 2022. Link.
    - Jain and Wallace. Attention is not Explanation. NAACL 2019. Link.
    - Wiegreffe and Pinter. Attention is not not Explanation. EMNLP 2019. Link.
    - Brunner et al. On Identifiability in Transformers. ICLR 2020. Link.
    - Sun and Marasović. Effective Attention Sheds Light On Interpretability. Findings of EACL 2021. Link.

6. 9/8: Attention-based post-hoc explanations. **Paper discussions.**
    - Group 1: Kobayashi et al. Attention is not only a weight: Analyzing transformers with vector norms. EMNLP 2020. Link.

- Group 2: Tutek and Šnajder. Staying True to Your Word: (How) Can Attention Become Explanation? Workshop on Representation Learning for NLP 2020. Link.
- Group 3: Bastings and Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? BlackboxNLP 2020. Link.

7. 9/13: Select-then-predict; faithfulness.
    - Jacovi and Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? ACL 2020. Link.
    - Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence (2019). Link.
    - Lei et al. Rationalizing Neural Predictions. EMNLP 2016. Link.
    - Bastings et al. Interpretable Neural Predictions with Differentiable Binary Variables. ACL 2019. Link.
    - Wiegreffe and Marasović. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. NeurIPS 2021. Link.
    - Yu et al. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. EMNLP 2019. Link.
    - DeYoung et al. ERASER: A Benchmark to Evaluate Rationalized NLP Models. ACL 2020. Link.
    - Carton et al. Evaluating and Characterizing Human Rationales. EMNLP 2020. Link.
    - Kingma and Welling. Auto-Encoding Variational Bayes. ICLR 2014. Link.
    - Jain et al. Learning to Faithfully Rationalize by Construction. ACL 2020. Link.
    - Doshi-Velez and Kim. Towards A Rigorous Science of Interpretable Machine Learning. Arxiv 2017. Link.
    - Chen et al. FRAME: Evaluating Simulatability Metrics for Free-Text Rationales. Arxiv 2022. Link.
    - Feng et al. Pathologies of Neural Models Make Interpretations Difficult. EMNLP 2018. Link.
    - Alvarez-Melis and Jaakkola. On the Robustness of Interpretability Methods. 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018). Link.

8. 9/15. Concept-based and hierarchical explanations.
    - Kim et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML 2018. Link.
    - Nejadgholi et al. Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors. ACL 2022. Link.
    - Antognini and Faltings. Rationalization through Concepts. Findings of ACl 2021. Link.
    - Rigotti et al. Attention-based Interpretability with Concept Transformers. ICLR 2022. Link. (Check out related works cited in this paper)
    - Reading papers:

    i. Fong. Reading a Computer Science Research Paper. Inroads — SIGCSE Bulletin. Link.
    ii. Keshav. How to Read a Paper. ACM SIGCOMM Computer Communication Review. Link.
    iii. Mitzenmacher. How to read a research paper. 2015. Link.
  - Reviewing papers:
    i. Rogers and Augenstein. How to review for ACL Rolling Review. Link.
    ii. Rogers and Augenstein. What Can We Do to Improve Peer Review in NLP? EMNLP Findings 2020.
    iii. NLP Highlights Podcast: On Writing Quality Peer Reviews, with Noah A. Smith. Link.
    iv. Two example good reviews from NAACL 2018 presented in their reviewing form
    v. Discursive advice in ACL 2017 from leading lights in the field: Mirella Lapata, Marco Baroni, Yoav Artzi, Emily Bender, Joel Tetreault, Ani Nenkova, and Tim Baldwin
    vi. Advice on Reviewing for EMNLP 2020
    vii. Allman. Thoughts on Reviewing. ACM Computer Communication Review Editorial Zone 2008. Link.
    viii. Cormode. How NOT to review a paper: The tools and techniques of the adversarial reviewer. SIGMOD Record 2008. Link.
    ix. Smith. The Task of the Referee. IEEE Computer 1990. Link.

9. 9/20: **Paper discussions.**
   - Group 1: Jacovi and Goldberg. Aligning Faithful Interpretations with their Social Attribution. TACL 2021. Link. ***("Only" until Section 10!)***
   - Group 2: Carton et al. Evaluating and Characterizing Human Rationales. EMNLP 2020. Link.
   - Group 3: Rigotti et al. Attention-based Interpretability with Concept Transformers. ICLR 2022. Link.

In plain English, why is this input assigned this label?

10. 9/22: Free-text explanations.
    - Marasović et al. Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs. EMNLP Findings 2020. Link.
    - Kayser et al. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. ICCV 2021. Link.
    - Wiegreffe et al. Measuring Association Between Labels and Free-Text Rationales. EMNLP 2021. Link.
    - Marasović et al. Few-Shot Self-Rationalization with Natural Language Prompts. NAACL Findings 2022. Link.

- Lampinen et al. Can language models learn from explanations in context? Arxiv 2022. Link.
- Wei et al. Chain of Thought Prompting Elicits Reasoning in Large Language Models. Arxiv 2022. Link.
- Zelikman et al. STaR: Bootstrapping Reasoning With Reasoning. Arxiv 2022. Link.

11. 9/27: Free-text explanations. **Paper discussions.**
    - Group 1: Majumder et al. Knowledge-Grounded Self-Rationalization via Extractive and Natural Language Explanations. ICML 2022. Link.
    - Group 2: Chen et al. FRAME: Evaluating Simulatability Metrics for Free-Text Rationales. Arxiv 2022. Link.
    - Group 3: Ye and Durrett. The Unreliability of Explanations in Few-Shot In-Context Learning. Arxiv 2022. Link.

    Extra:
    - Camburu et al. Make Up Your Mind! Adversarial Generation of Inconsistent Natural Language Explanations. ACL 2020. Link.

Which training examples caused the prediction?

12. 9/29: Influence functions.
    - Koh and Liang. Understanding Black-box Predictions via Influence Functions. ICML 2017. Link.
    - Guo et al. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. EMNLP 2021. Link.
    - Pruthi et al. Estimating Training Data Influence by Tracing Gradient Descent. NeurIPS 2020. Link.
    - Yeh et al. Representer Point Selection for Explaining Deep Neural Networks. NeurIPS 2018. Link.
    - Han et al. Explaining Black Box Predictions and Unveiling Data Artifacts through Influence Functions. ACL 2020. Link.
    - Basu et al. Influence Functions in Deep Learning Are Fragile. ICLR 2021. Link.
13. 104: Alternatives for computing influence. **Paper discussions.**
    - Group 1: Guo et al. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. EMNLP 2021. Link.
    - Group 3: Han et al. ORCA: Interpreting Prompted Language Models via Locating Supporting Data Evidence in the Ocean of Pretraining Data. Arxiv 2022. Link.

Which part of the input should be changed to change the prediction to a given label?

14. 10/6: Contrastive editing; contrastive vector representation.
    - Miller. Explanation in artificial intelligence: Insights from the social sciences. In AIJ 2019. Chapter 2. Link.
    - Yang et al. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. COLING 2020. Link.
    - Jacovi and Goldberg. Aligning Faithful Interpretations with their Social Attribution. TACL 2021. Link.

- Chen et al. KACE: Generating Knowledge-Aware Contrastive Explanations for NLI. ACL 2021. Link.
- Ross et al. Explaining NLP Models via Minimal Contrastive Editing (MiCE). ACL Findings 2021. Link.
- Paranjape et al. Prompting Contrastive Explanations for Commonsense Reasoning Tasks. ACL Findings 2021. Link.
- Wu et al. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. ACL 2021. Link.
- Jacovi et al. Contrastive Explanations for Model Interpretability. EMNLP 2021. Link.

Extra:

- Goyal et al. Counterfactual Visual Explanations. ICML 2019. Link.
- ALICE: Active Learning with Contrastive Natural Language Explanations. EMNLP 2020. Link.
- Kanehira et al. Multimodal Explanations by Predicting Counterfactuality in Videos. CVPR 2019. Link.

15. 10/11: Fall break.
16. 10/13: Fall break.

Psychology of human explanations

17. 10/18: Foundations.
    - Lombrozo. Explanation and abductive inference. Oxford handbook of thinking and reasoning 2012. Link.
    - Aronowitz and Lombrozo. Experiential Explanation. Topics in Cognitive Science 2020. Link.
    - Kuhn. How do people know? Psychological science 2001. Link.
    - Wilson and Keil. The shadows and shallows of explanation. Minds and machines 1998. Link.
    - Keil. Folkscience: Coarse interpretations of a complex reality. Trends in cognitive sciences 2003. Link.
    - Giffin, Wilkenfeld, and Lombrozo. The explanatory effect of a label: Explanations with named categories are more satisfying. Cognition 2017. Link.
    - Lombrozo. Explanatory preferences shape learning and inference. Trends in Cognitive Sciences 2016. Link.
    - Blanchard, Vasilyeva, and Lombrozo. Stability, breadth and guidance. Philosophical Studies 2018. Link.
    - Scharrer et al. The seduction of easiness: How science depictions influence laypeople's reliance on their own evaluation of scientific information. Learning and Instruction 2012. Link.

    Extra: Integrating psychological frameworks.
    - Yang et al. A Psychological Theory of Explainability. ICML 2022. Link.

- Jacovi et al. Diagnosing AI Explanation Methods with Folk Concepts of Behavior. Arxiv 2022. Link.
- González et al. On the Interaction of Belief Bias and Explanations. ACL Findings 2021. Link.

Application-grounded, human-subject evaluations of explanations

18. 10/20: Experimental design, use cases, and challenges.
    - Buçinca et al. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. IUI 2020. Link.
    - Laio and Varshney. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. Book chapter.
    - Suresh et al. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. CHI 2021. Link.
    - Lai and Tan. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. FAccT 2019. Link.
    - Zhang et al. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. FAccT 2020. Link.
    - Wang and Yin. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. IUI 2021. Link.
    - McKnight et al. Developing and validating trust measures for e-commerce: An integrative typology. Information systems research 2002. Link.
    - Cheng et al. Explaining Decision-Making Algorithms through UI:
    - Strategies to Help Non-Expert Stakeholders. CHI 2019. Link.
    - Lai et al. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv 2021. Link.
    - Kaur et al. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. CHI 2020. Link.
    - Dodge et al. Explaining models: an empirical study of how explanations impact fairness judgment. IUI 2019. Link.
    - Kaur et al. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. FAccT 2022. Link.
19. 10/25: Human-subject evaluations of explanations in NLP and CV. **Paper discussions.**
    - Group 1: González et al. Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations. ACL 2021. Link.
    - Group 3: Colin et al. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. arXiv 2021. Link.

Extra:

- Parrish et al. Single-Turn Debate Does Not Help Humans Answer Hard Reading-Comprehension Questions. ACL 2022 Workshop on Learning with Natural Language Supervision. Link.
- Mozannar et al. Teaching Humans When To Defer to a Classifier via Examplars. AAAI 2022.
- Feng and Boyd-Graber. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. IUI 2019. Link.
- Chu et al. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. arXiv 2020. Link.
- Bansal et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021. Link.
- Arora et al. Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations. AAAI 2022. Link.

Explainability as a dialog

20. 10/27: Principles, roadmap, risks, and research opportunities.
    - Miller et al. Explanation in Artificial Intelligence: Insights from the Social Sciences. Section 5. Link.
    - Lakkaraju et al. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. arXiv 2022. Link.
    - Slack et al. TalkToModel: Understanding Machine Learning Models With Open Ended Dialogues. Arxiv 2022. Link.
21. 11/1: Implementations of explainability as a dialog. **Paper discussions.**
    - Group 1: Jung et al. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. Link.
    - Group 3: Xie et al. Calibrating Trust of Multi-Hop Question Answering Systems with Decompositional Probes. Link

    Extra:
    - Feldhus et al. Mediators: Conversational Agents Explaining NLP Model Behavior. IJCAI-ECAI 2022 Workshop on Explainable Artificial Intelligence (XAI) 2022. Link.
    - Wu et al. INSCIT: Information-Seeking Conversations with Mixed-Initiative Interactions. Link.

Trust in AI

22. 11/3: Prerequisites, causes, and goals of human trust in AI.
    - Jacovi et al. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. FAccT 2021. Link.
23. 11/8: Delegability; ML for positive impact; concrete measures for evaluating trust. **Paper discussions.**
    - Group 1: Lubars et al. Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability. NeurIPS 2019. Link.
    - Group 3: Jin et al. How Good Is NLP? A Sober Look at NLP Tasks through the Lens of Social Impact. ACL 2021. Link.

Extra:

- Deng et al. Trust, but Verify: Using Self-Supervised Probing to Improve Trustworthiness. ECCV 2022.
- Fortuna et al. Cartography of Natural Language Processing for Social Good (NLP4SG): Searching for Definitions, Statistics and White Spots. ACL Workshop on NLP for Positive Impact. Link.
- Rong et al. User Trust on an Explainable AI-based Medical Diagnosis Support System. Workshop on Trust and Reliance in AI-Human Teams at CHI 2022. Link.
- Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research? Workshop on Trust and Reliance in AI-Human Teams at CHI 2022. Link.

Advanced Topics

24. 11/10: Few-shot learning
    - A Visual Guide to Using BERT for the First Time by Jay Alammar
    - Ravichander et al. CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation. EMNLP 2022. Link.
    - Ouyang et al. Training language models to follow instructions with human feedback. Arxiv 2022. Link.
    - Wei et al. Chain of Thought Prompting Elicits Reasoning in Large Language Models. Arxiv 2022. Link.
    - Wang et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. Arxiv 2022. Link.
    - Chung et al. Scaling Instruction-Finetuned Language Models. Arxiv 2022. Link.
    - Jung et al. Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations. Link.
    - Press et al. Measuring and Narrowing the Compositionality Gap in Language Models. Arxiv 2022. Link.
    - Arora et al. Ask Me Anything: A simple strategy for prompting language models. Arxiv 2022. Link.
    - Zhao et al. Calibrate Before Use: Improving Few-Shot Performance of Language Models. ICML 2021. Link.
    - Min et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? EMNLP 2022. Link.
    - Webson et al. Do Prompt-Based Models Really Understand the Meaning of their Prompts? NAACL 2022. Link.
    - Yoo et al. Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations. EMNLP 2022. Link.
    - How does in-context learning work? A framework for understanding the differences from traditional supervised learning
    - Zhang et al. Robustness of Demonstration-based Learning Under Limited Data Scenario. EMNLP 2022. Link.
    - Bragg et al. FLEX: Unifying Evaluation for Few-Shot NLP. NeurIPS 2021. Link.

- Perez et al. True Few-Shot Learning with Language Models. NeurIPS 2021. Link.
- Liu et al. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. Arxiv 2021. Link.
- Yang et al. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. AAAI 2022. Link.
- Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. Arxiv 2022. Link.

25. 11/15: Self-attention vs. convolution. **Paper discussions.**
    - Group 1: Cordonnier et al. On the Relationship between Self-Attention and Convolutional Layers. ICLR 2020. Link.
    - Group 3: Raghu et al. Do Vision Transformers See Like Convolutional Neural Networks? NeurIPS 2021. Link.

No more readings after this 🎉🥳

26. 11/22: Project meetings and/or peer user studies
27. 11/24: Break
28. 11/29: Project Presentations
29. 12/1: Project Presentations
30. 12/6: Looking Back