Adapting to Context: A Case Study on In-Context Learning of Decision Tree Algorithms by Large Language Models

Motivated by: Garg, Shivam, et al. "What can transformers learn in-context? a case study of simple function classes." *Advances in Neural Information Processing Systems* 35 (2022): 30583-30598

Abdullah Azhar
INFO-259

Contents

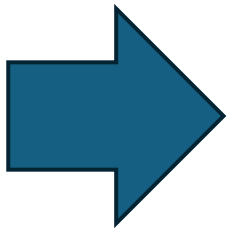
- Project Motivation - *what is in-context learning?*
- Objective - *teaching a large language model to in-context learn?*
- Roadmap and deliverables -
  - *Dataset & Model Architecture*
  - *Model Training*
  - *Model Inference (noisy and out-of-distribution prompting)*

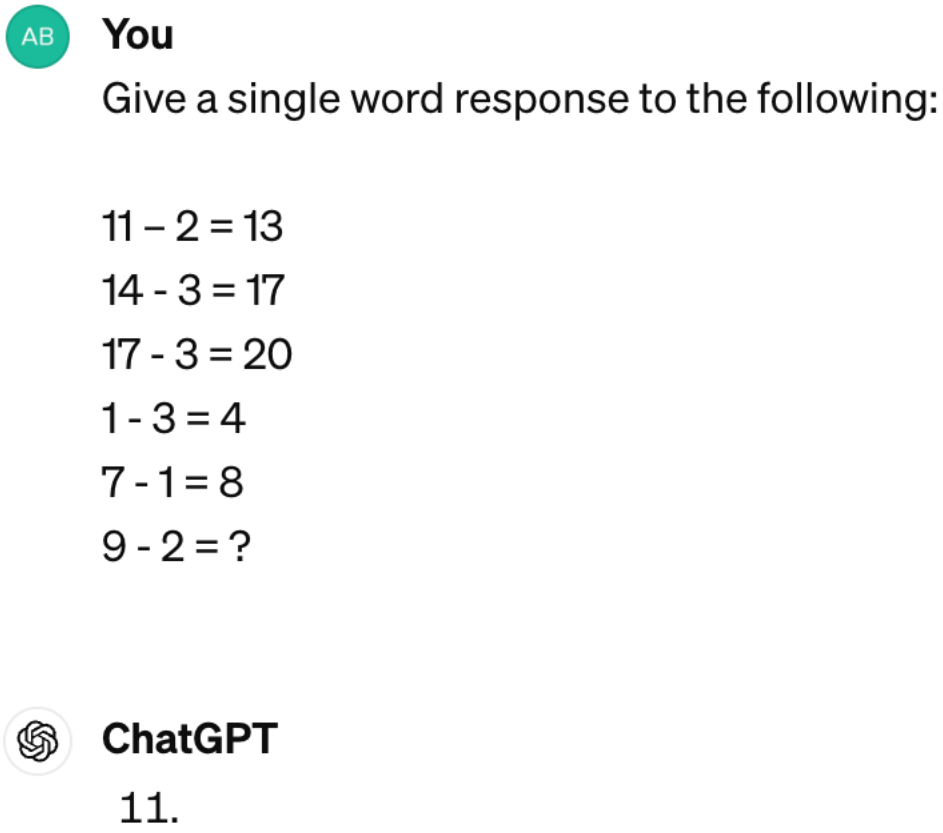
Motivation – *what is in-context learning?*

- 11 – 2 = 13
- 14 - 3 = 17
- 17 - 3 = 20
- 1 - 3 = 4
- 7 - 1 = 8
- 9 - 2 = ?

Motivation – *what is in-context learning?*

- 11 – 2 = 13
- 14 - 3 = 17
- 17 - 3 = 20
- 1 - 3 = 4
- 7 - 1 = 8
- 9 - 2 = ?





Motivation – *what is in-context learning?*

- In-context learning happens at inference time without any weight updates to the model
- However, it is unclear what the relationship is between tasks on which this succeeds and what is present in the training data?

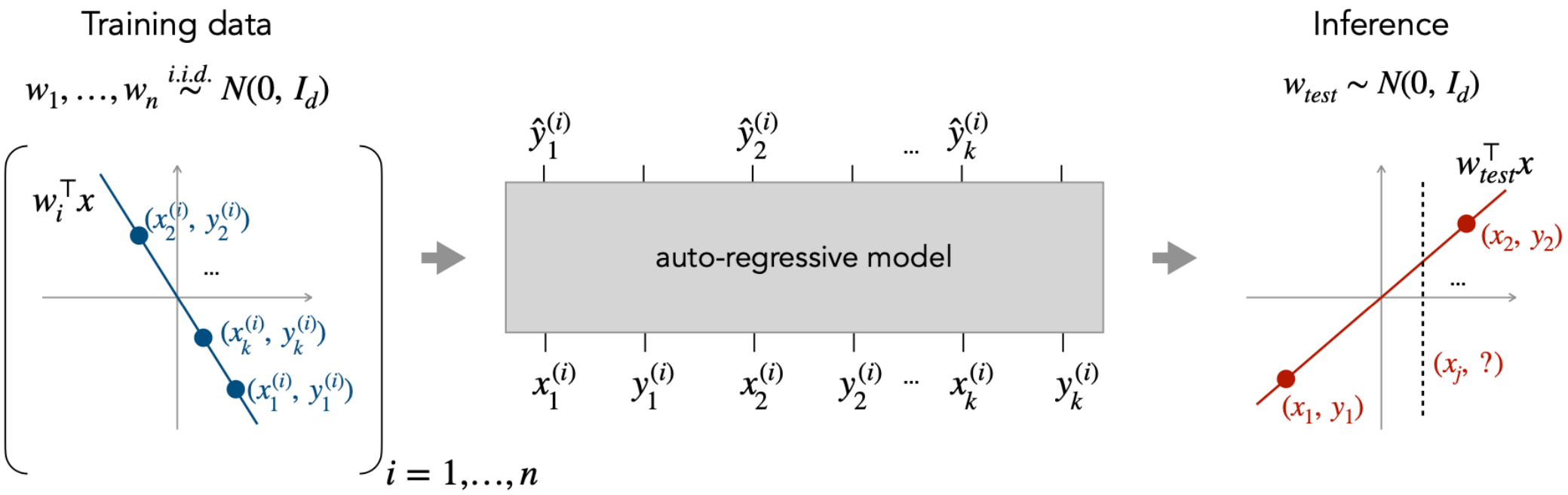
Motivation – *what is in-context learning?*

- In-context learning happens at inference time without any weight updates to the model
- However, it is unclear what the relationship is between tasks on which this succeeds and what is present in the training data?
- Is the model in the previous example simply tapping on the training data?

Objective – *teaching an LLM to incontext learn*

- Garg, Shivam, et al train a causal masking auto-regressive GPT2 Model to in-context learn
- Training done from scratch (no-fine tuning)
- Eliminating the ambiguity of training data's role in in-context learning

Objective – *teaching an LLM to incontext learn*



Garg, Shivam, et al. "What can transformers learn in-context? a case study of simple function classes." *Advances in Neural Information Processing Systems* 35 (2022): 30583-30598

Dataset and Model Architecture

- GPT2 Architecture:
  - 22 Million Parameters
  - Layers: 12
  - Heads: 8
  - Embedding Dimension: 256
- Dataset Generation:
  - Input Dimension: 20 (drawn from gaussian i.i.d. distribution)
  - Decision Tree Depth: 4
  - Split Category: signed based on uniform distribution of possible nodes

Project Roadmap - *Robustness to Noise*

- Train the transformer model on 101 prompt examples
- Varying noise levels drawn from a gaussian distribution with std: [0, 0.01, 0.1, 1, 2] to compare robustness to noise
- Perform inference on trained model by varying prompt distribution
- Prompting strategies:
  - Standard (same distribution at inference and training)
  - Random Quadrant Distribution (in-context examples belonging to different quadrants)
  - Noisy labels in the prompt examples