Instruction tuning

“Remember to sign up for presentations.”

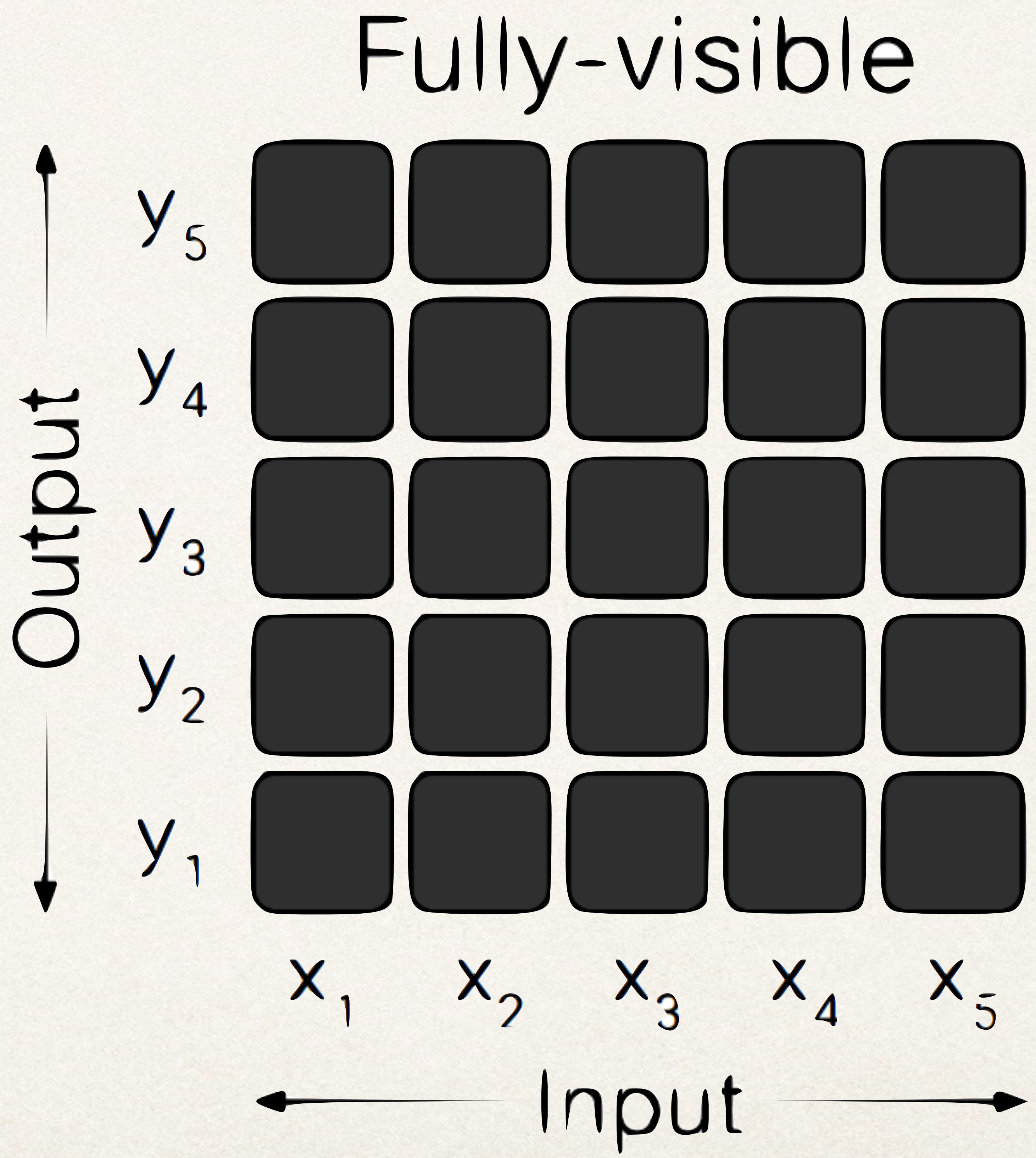
–Kenneth

Training

- Pretraining
    - Learn general, useful representations that are transferable to multiple contexts
        - Usually from large, non-specialised datasets (e.g., wikipedia)
- Fine-tuning
    - Using the general, pretrained parameters as inputs that are further adjusted to a specific purpose

Training: encoder

- Pretraining
  - Different tasks can be used - but mostly masked language modelling



Training: encoder

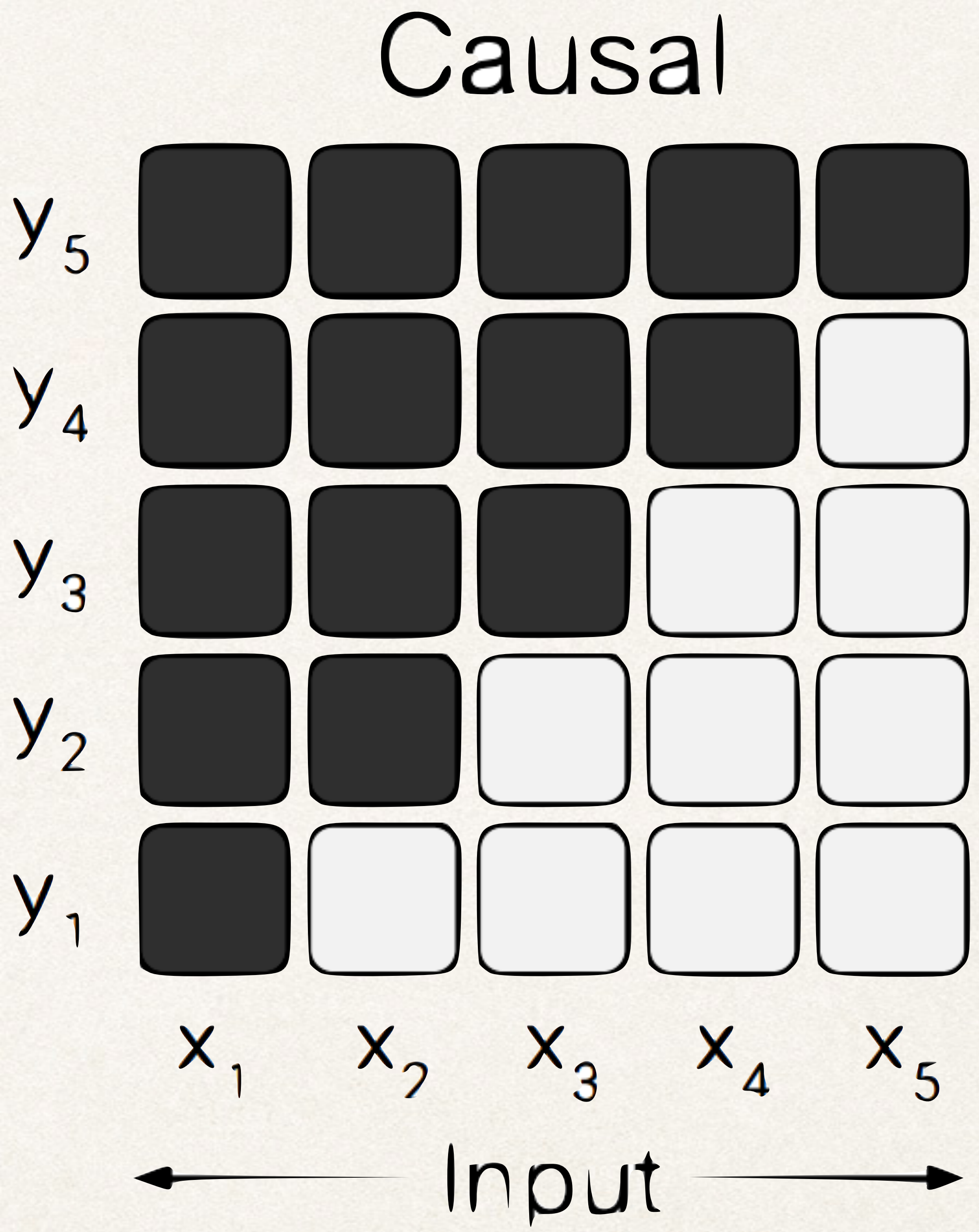
- Fine-tuning
    - The goal is to adapt the learned representations to perform well on the particular task at hand, such as text classification or named entity recognition
    - During finetuning, the encoder's weights are updated using the task-specific data, building on the general knowledge gained during pretraining

Example: medical notes

- Let’s say I have some medical notes, and I want to investigate, e.g., whether the notes can be used to investigate disease severity
- “Patient is presenting with 3-day history of sore throat, nasal congestion, and mild cough.”
- “Patient has history of COPD presenting with progressive shortness of breath, admitted Thursday.”

Training: decoder

- Pretraining
  - Language modelling task (next token prediction)



Training: decoder

- Fine-tuning
    - More language modelling

Instruction tuning

- Type of fine tuning for language models (that are meant to be interactive)

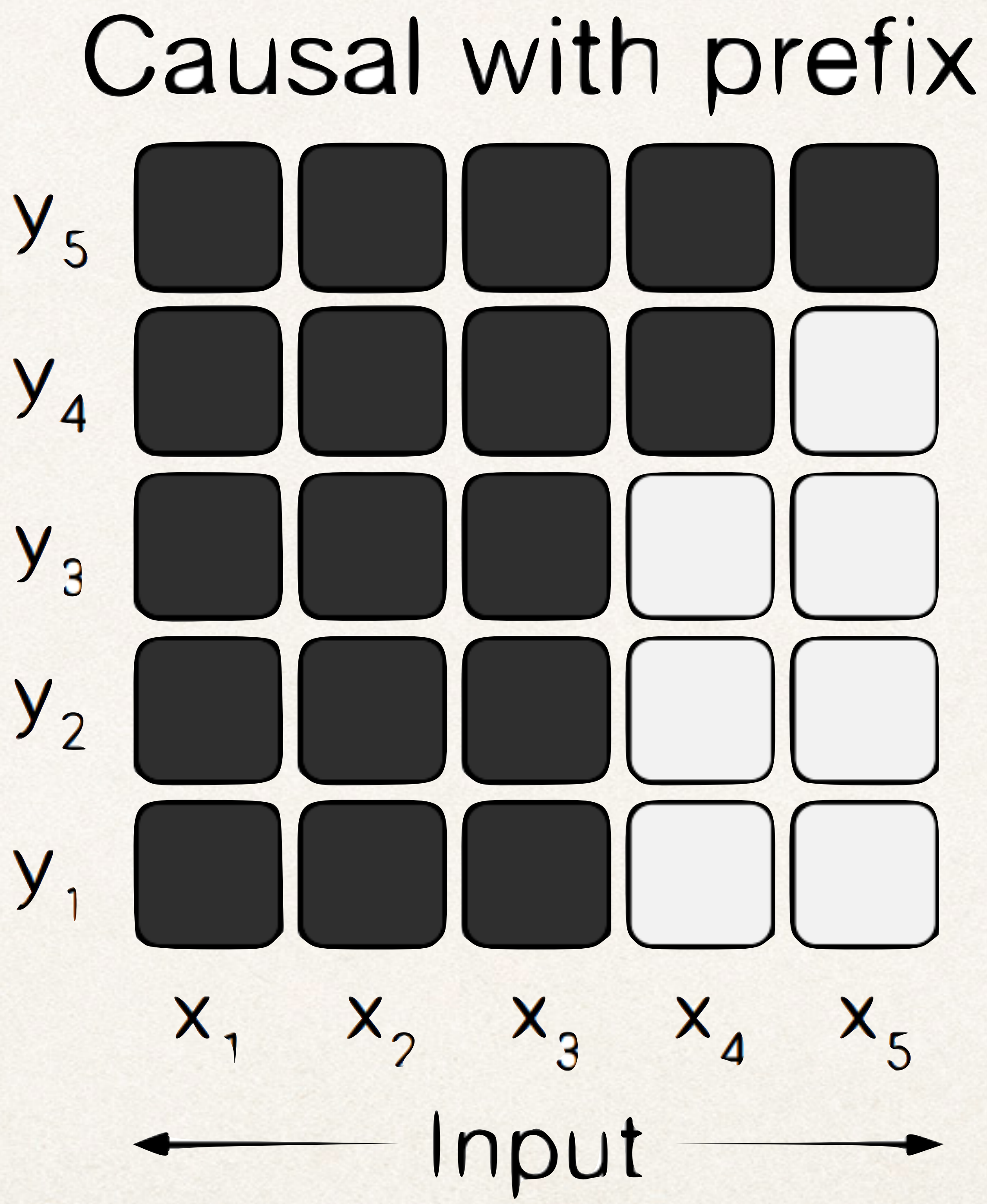
- Misalignment between the language modelling objective (next token prediction) and user objective (responding to a query) teach me how to bake bread

Absolutely, let's go through the process of baking a simple, classic loaf of bread!

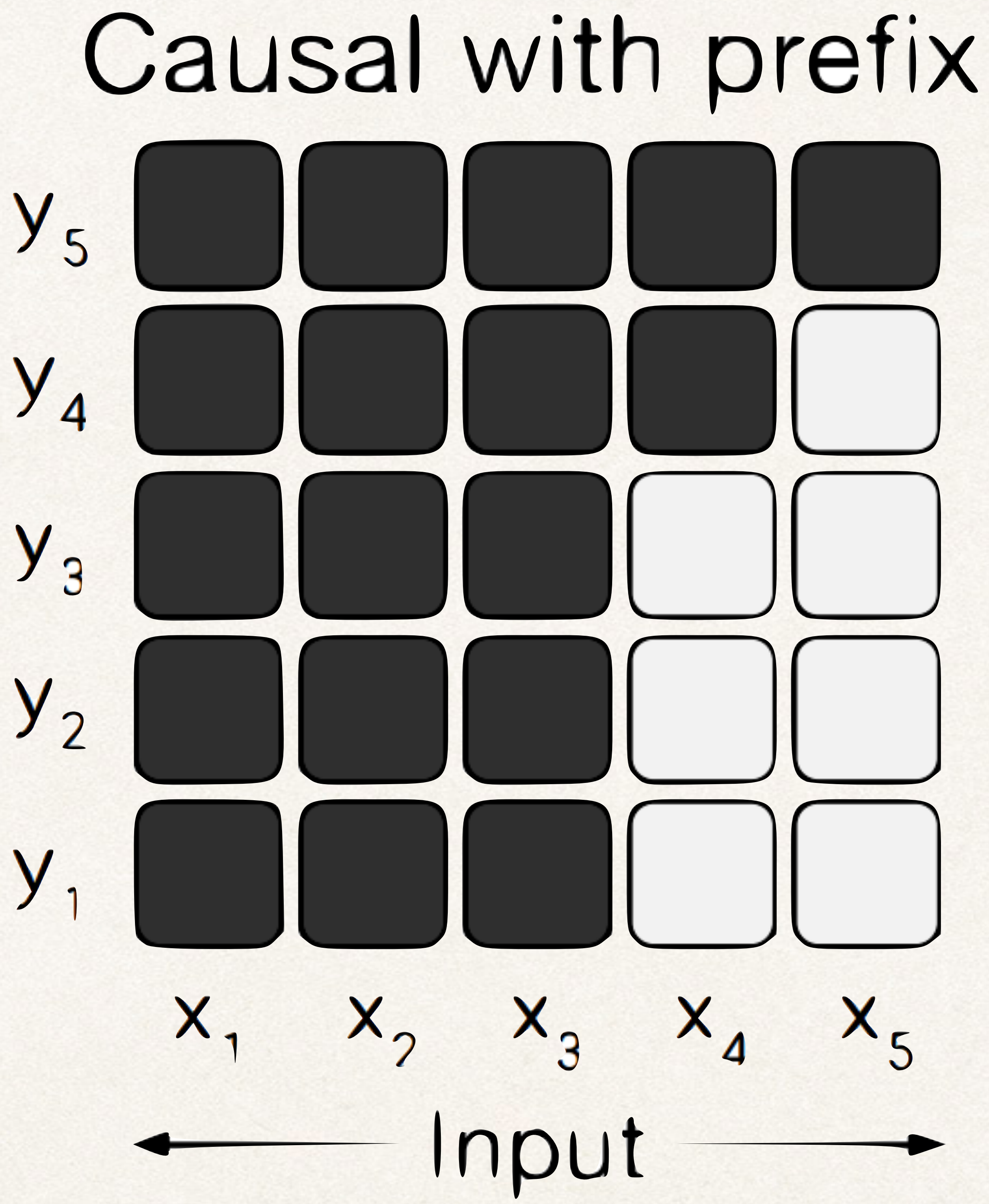
teach me how to bake bread **in a home oven**

Instruction tuning

- Language models do not answer - they append



Instruction tuning



Answer the following question: teach me how to bake bread