Overview of in-context learning

Mehmet F. Demirel

Oct 12, 2021

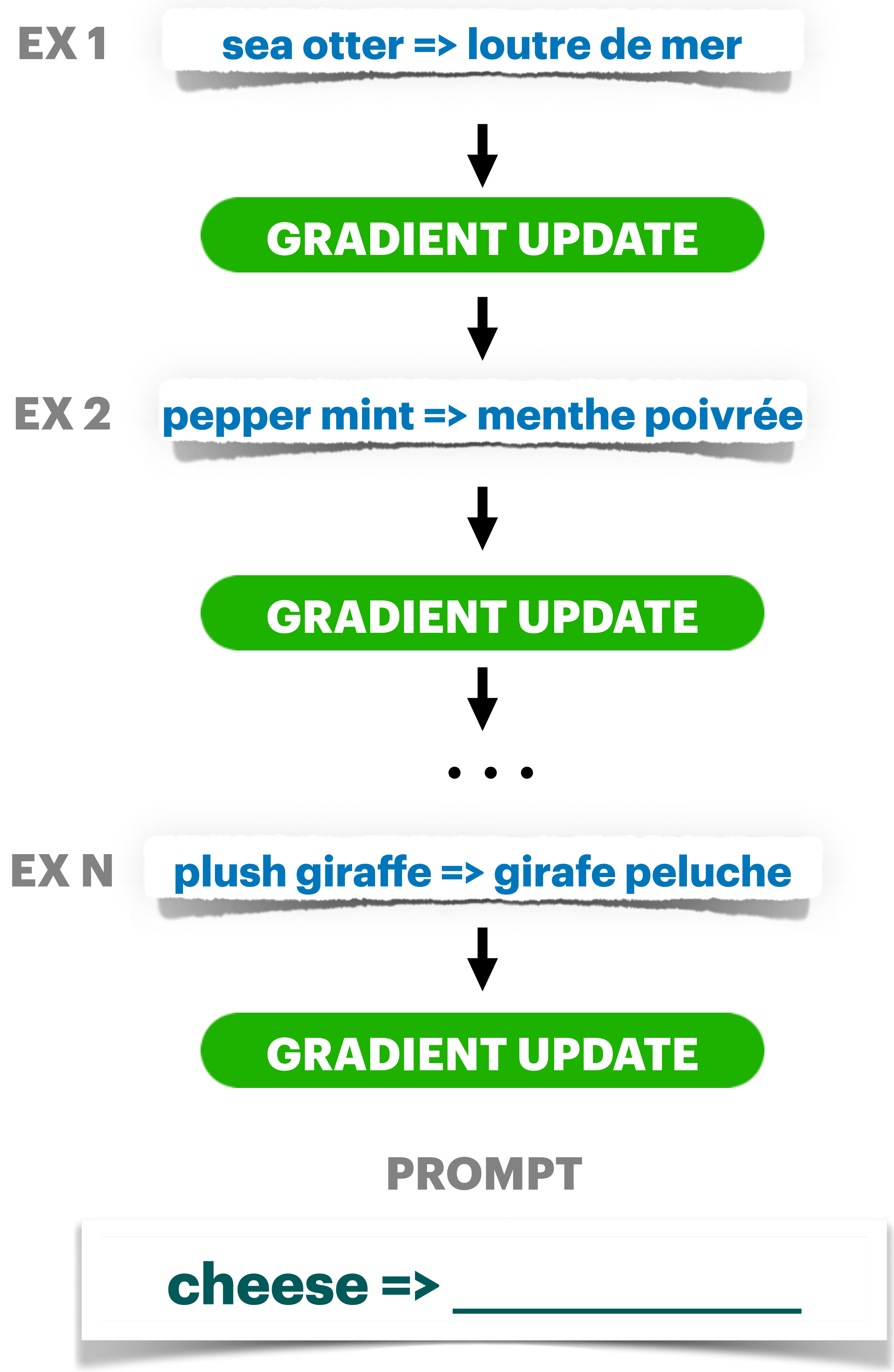
The common part: unsupervised learning

- Pretrain a language model on a large corpus of linguistic data
  - predict missing word —> **Sam took the ____ for a walk.**
  - predict next word —> **Sam took the dog for a _____.**

The traditional way: Fine-tuning

- Fine-tune the parameters of the pre-trained model for a specific downstream task using a large (thousands to hundreds of thousands) corpus of **labeled data**.
- Keep training the model via repeated gradient updates.

- **Strong performance on many benchmarks.**
- **Need a new large dataset for each task.**
- **Potential for poor out-of-distribution generalization**
- **Potential to explore spurious features of the data**



EX 1 sea otter => loutre de mer

GRADIENT UPDATE

EX 2 pepper mint => menthe poivrée

GRADIENT UPDATE

. . .

EX N plush giraffe => girafe peluche

GRADIENT UPDATE

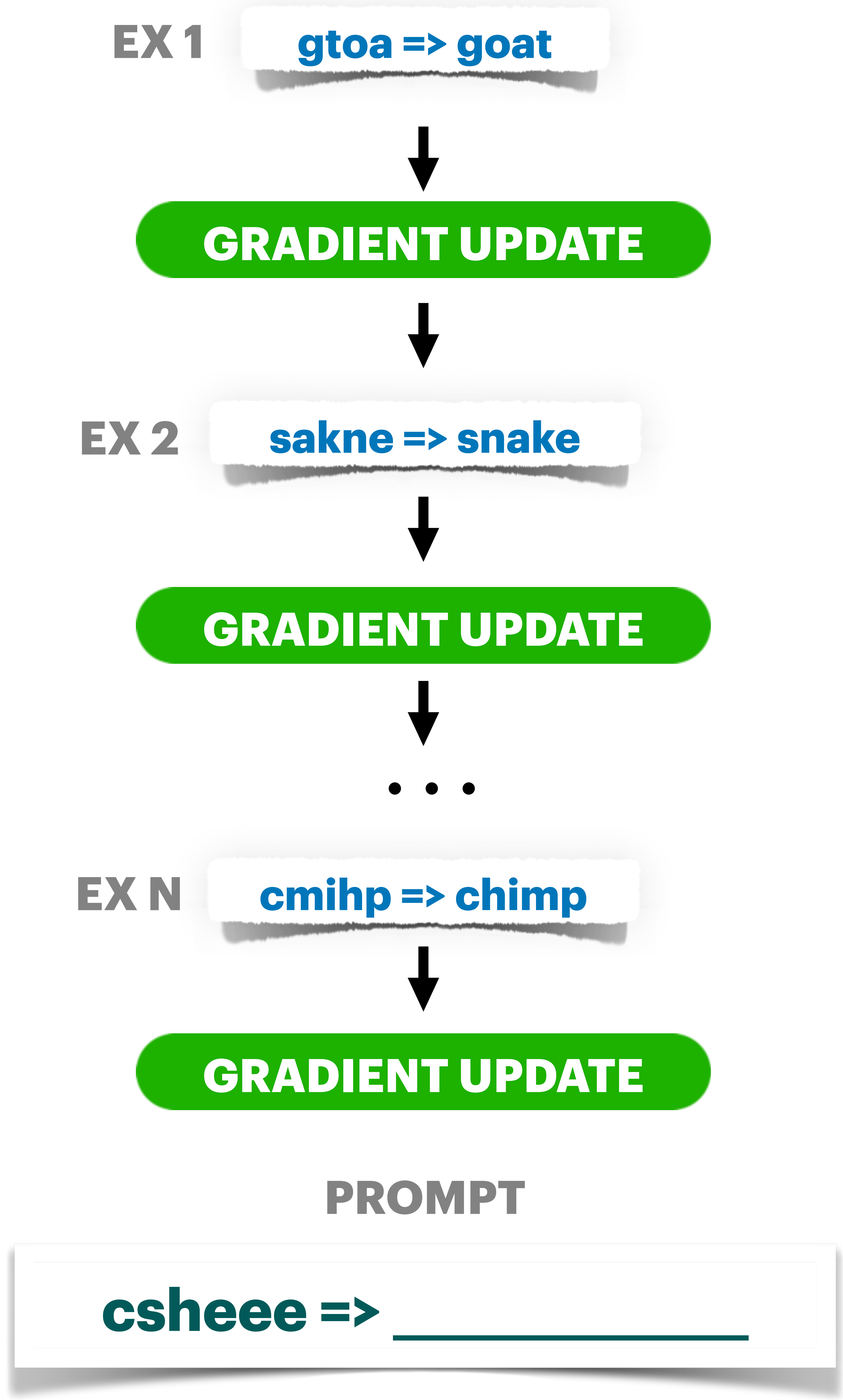
PROMPT

cheese => ______________

The traditional way: Fine-tuning

- Fine-tune the parameters of the pre-trained model for a specific downstream task using a large (thousands to hundreds of thousands) corpus of **labeled data**.
- Keep training the model via repeated gradient updates.

- **Strong performance on many benchmarks.**
- **Need a new large dataset for each task.**
- **Potential for poor out-of-distribution generalization**
- **Potential to explore spurious features of the data**



EX 1 gtoa => goat

GRADIENT UPDATE

EX 2 sakne => snake

GRADIENT UPDATE

. . .

EX N cmihp => chimp

GRADIENT UPDATE

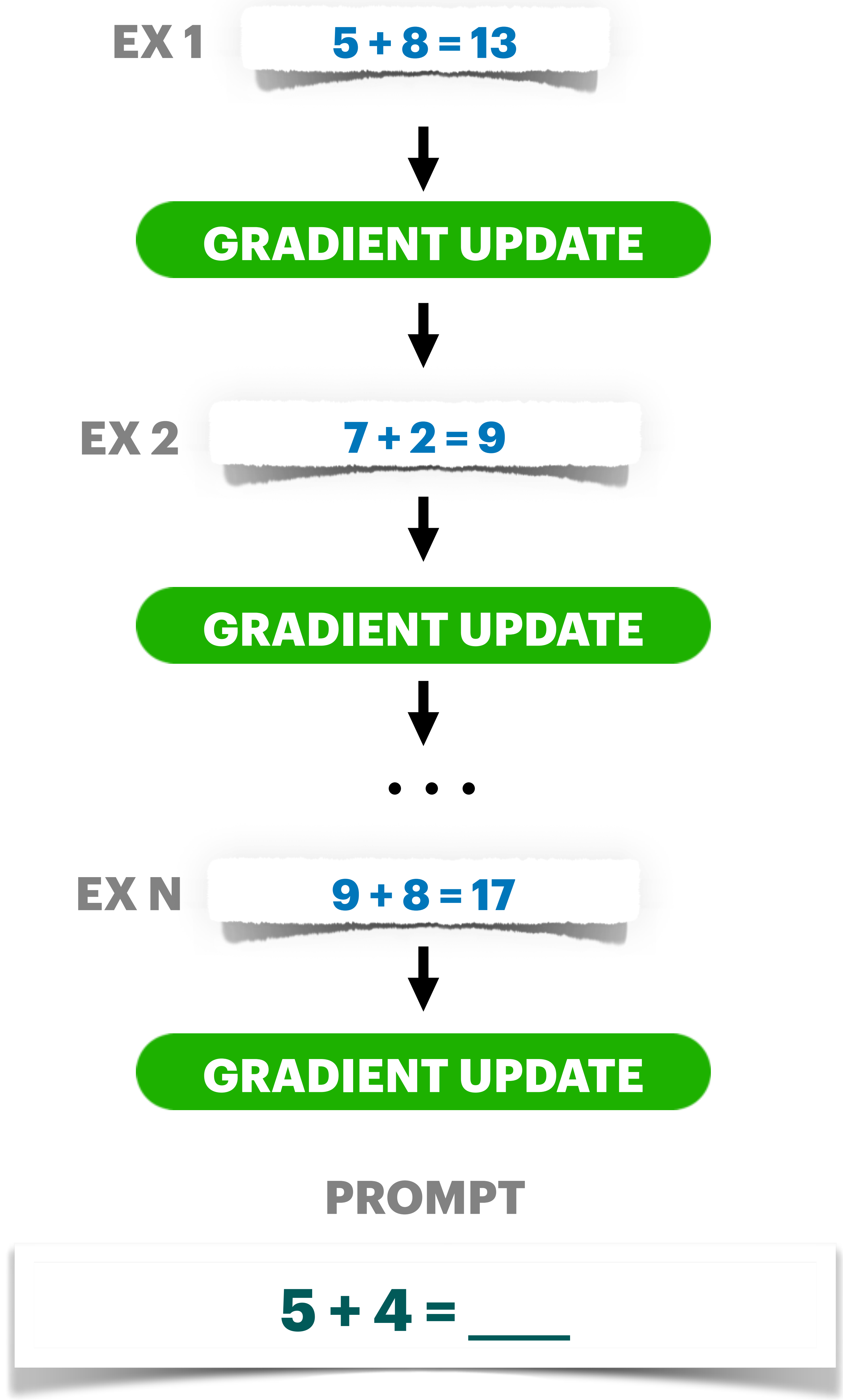
PROMPT

csheee => ____________

The traditional way: Fine-tuning

- Fine-tune the parameters of the pre-trained model for a specific downstream task using a large (thousands to hundreds of thousands) corpus of **labeled data**.
- Keep training the model via repeated gradient updates.

- **Strong performance on many benchmarks.**
- **Need a new large dataset for each task.**
- **Potential for poor out-of-distribution generalization**
- **Potential to explore spurious features of the data**



In-context learning

- No training or optimization of the model parameters in the “adaptation step”.
- Simply give the model **a task description** as well as **none/one/few examples** as the input at inference time.
    - Only the task description: **ZERO-SHOT**
    - TD + one examples : **ONE-SHOT**
    - TD + a few examples: **FEW-SHOT**
- No gradient updates are performed.

FEW-SHOT

Translate English to French

sea otter => loutre de mer

peppermint => menthe poivrée

plush giraffe => girafe peluche

cheese => ____________

ONE-SHOT

Translate English to French

sea otter => loutre de mer

cheese => ____________

ZERO-SHOT

Translate English to French

cheese => ____________

PROVIDE THE EXAMPLE(S)
IN THE CONTEXT OF
THE LANGUAGE MODEL

FEW-SHOT

Perform mathematical addition

5 + 8 = 13

7 + 2 = 9

9 + 8 = 17

5 + 4 = ____

ONE-SHOT

Perform mathematical addition

5 + 8 = 13

5 + 4 = ____

ZERO-SHOT

Perform mathematical addition

5 + 4 = ____

PROVIDE THE EXAMPLE(S)

IN THE CONTEXT OF

THE LANGUAGE MODEL

Few-shot

- **Give K examples of context and completion, and one final context whose prompt we want the model to predict.**
- **Major reduction in the need for task-specific data.**
- **Reduced potential to learn an overly narrow distribution from a large but narrow fine-tuning dataset.**
- **Still not as good as the fine-tuning SOTA, but competitive (GPT-3).**
- **Still need a few task-specific data.**

One-shot

- **Similar to few-shot, but with only one example**
- **Most closely matches the way in which some tasks are communicated to humans.**

Zero-shot

- **Provides maximum convenience (no task-specific example needed)**
- **Potential for robustness**
- **Potential for avoidance of spurious correlations**
- **Most challenging**
- **Even for humans, it is often hard to understand a task without an example.**

Foundation models

- **The survey indicates that there is an emergence of functionalities (such as in-context learning) in foundation models.**
- **Rather than task-specific data and carefully-engineered features, NLP foundation models (such as GPT-2 and GPT-3) can make inference for given tasks whose task-specific examples are provided in the context of the language model as an input at inference time with no parameter optimization required.**