

Small Models, Big Impact: The BabyLM Challenge



Edoardo Vaira - Georgios Chavales

There is a **problem** with the current state of LLMs...

There is a **problem** with the current state of LLMs...
their **sizes**

Model	Parameters
GPT-4	> 1 Trillion
GPT-3.5	175 Billion
Llama 3.1	405 Billion
Chinchilla	70 Billion

Which **means...**

Which **means...**

 **Resource-Heavy:** Require lots of power and memory.

Which **means...**

- ✗ **Resource-Heavy:** Require lots of power and memory.
- ✗ **Slow:** Predictions take time.

Which **means...**

- ✗ **Resource-Heavy:** Require lots of power and memory.
- ✗ **Slow:** Predictions take time.
- ✗ **Expensive:** High training and usage costs.

Which **means...**

- ✗ **Resource-Heavy:** Require lots of power and memory.
- ✗ **Slow:** Predictions take time.
- ✗ **Expensive:** High training and usage costs.
- ✗ **Energy-Intensive:** Big carbon footprint.

Which **means...**






- ✗ **Resource-Heavy:** Require lots of power and memory.
- ✗ **Slow:** Predictions take time.
- ✗ **Expensive:** High training and usage costs.
- ✗ **Energy-Intensive:** Big carbon footprint.
- ✗ **Limited Flexibility:** Struggle with small or specific data.

So in this presentation we will give you some **ideas** on how to:

So in this presentation we will give you some **ideas** on how to:






- ✗ **Resource-Heavy**
- ✗ **Slow**
- ✗ **Expensive**
- ✗ **Energy-Intensive**
- ✗ **Limited Flexibility**

So in this presentation we will give you some **ideas** on how to:

-  **Resource-Heavy**
-  **Slow**
-  **Expensive**
-  **Energy-Intensive**
-  **Limited Flexibility**



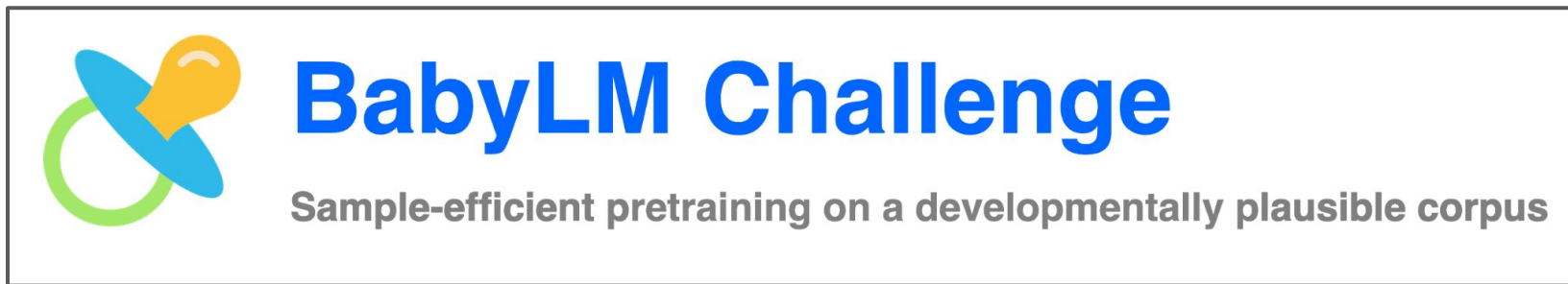
So in this presentation we will give you some **ideas** on how to:

 **Resource-Heavy**
 **Slow**
 **Expensive**
 **Energy-Intensive**
 **Limited Flexibility**



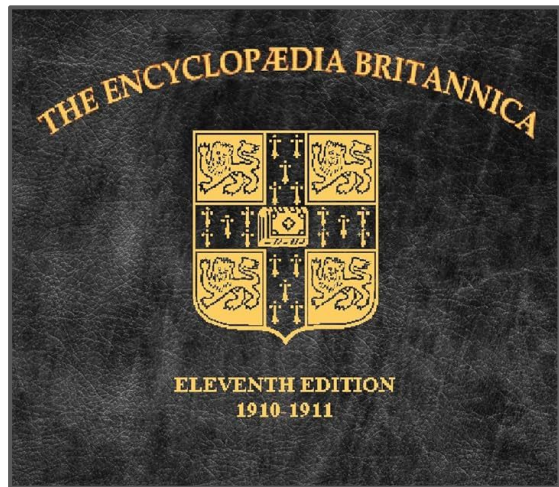
 **Lightweight**
 **Fast**
 **Affordable**
 **Green AI**
 **Versatile**

The **inspiration** for this project came from the...



Which is a **competition** where participants aim to achieve the best possible results on a **limited-size dataset** (10M or 100M words) using the most **efficient model** possible.

So we needed a **dataset**... **Encyclopaedia Britannica**



- 1911 10th edition
- 29 **volumes**
- 38 million **words** in total

https://en.wikisource.org/wiki/1911_Encyclop%C3%A6dia_Britannica

Why an **encyclopaedia**?

Textbooks Are All You Need

Suriya Gunasekar Yi Zhang Jyoti Aneja Caio César Teodoro Mendes
Allie Del Giorno Sivakanth Gopi Mojan Javaheripi Piero Kauffmann
Gustavo de Rosa Olli Saarikivi Adil Salim Shital Shah Harkirat Singh Behl
Xin Wang Sébastien Bubeck Ronen Eldan Adam Tauman Kalai Yin Tat Lee
Yuanzhi Li
Microsoft Research

High quality dataset → High quality model

How we got it:

1. Scraped from wikisource.org in **JSON** format
2. Turning the JSON files into a single **TXT** file
3. Corpus cleaning using **Regex** operations:
 - a. Removing special characters and extra white spaces
 - b. Removing references and citations to preserve main content
 - c. Standardizing punctuation and artifacts
 - d. Added article boundary tokens (<s> and </s>)

Here's how it **looks** like:

... **</s> <s>** Peipus, or Chudskoye Ozero, a lake of north-west Russia, between the governments of St Petersburg, Pskov, Livonia and Esthonia. Including its southern extension, sometimes known as Lake Pskov, it has an area of 1356 sq. m. Its shores are flat and sandy, and in part wooded; its waters deep, and they afford valuable fishing. The lake is fed by the Velikaya, which enters it at its southern extremity, and by the Embach, which flows in half way up its western shore; it drains into the Gulf of Finland by the Narova, which issues at its north-east corner. **</s> <s>** Basement, the term applied to the lowest storey of any building placed wholly or partly below the level of the ground. It is incorrectly applied to the ground storey of any building, even when, as for instance in the case of Somerset House, London, the ground floor is of plain or rusticated masonry, and the upper storey which it supports is divided up and decorated with columns or pilasters. **</s> <s>** Brenham, a city and the county-seat of Washington county, Texas, U.S.A. ...

Then we needed to **tokenize** the text...

- Byte-Pair Encoding (BPE) tokenizer
- 16,000 vocabulary size
- Sequences of 128 tokens
- Big articles split in multiple sequences
- Small articles combined in single sequence
- Padded with <pad> token

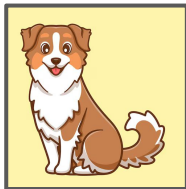
Knowledge Distillation.



Knowledge Distillation transfers knowledge from a large, accurate model (***teacher***) to a smaller, faster model (***student***). The student learns not just the correct answers but also how the teacher “**thinks**”.

The best way to understand how **knowledge distillation** works is through an example...

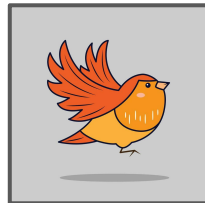
The best way to understand how **knowledge distillation** works is through an example...



“Dog”



“Cat”



“Bird”

Imagine training a machine learning model to classify **pictures of animals**.

To apply **knowledge distillation**, you follow **3 steps**:

1. Train the Teacher(s)
2. Understand how Temperature and Soft Labels work
3. Train the Student

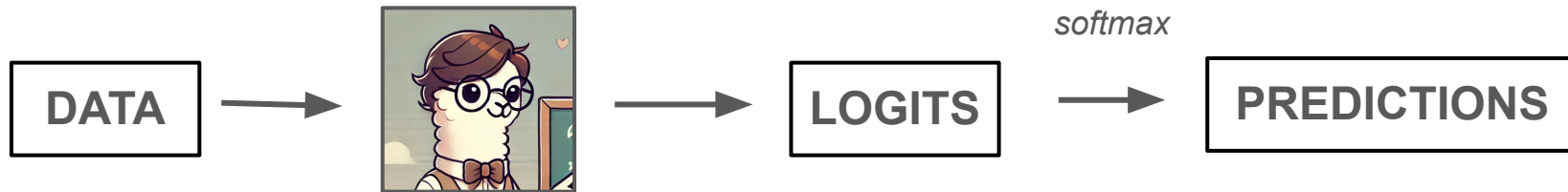
1. **Train the Teacher(s)**
2. Understand how Temperature and Soft Labels work
3. Train the Student

Training the **Teacher**

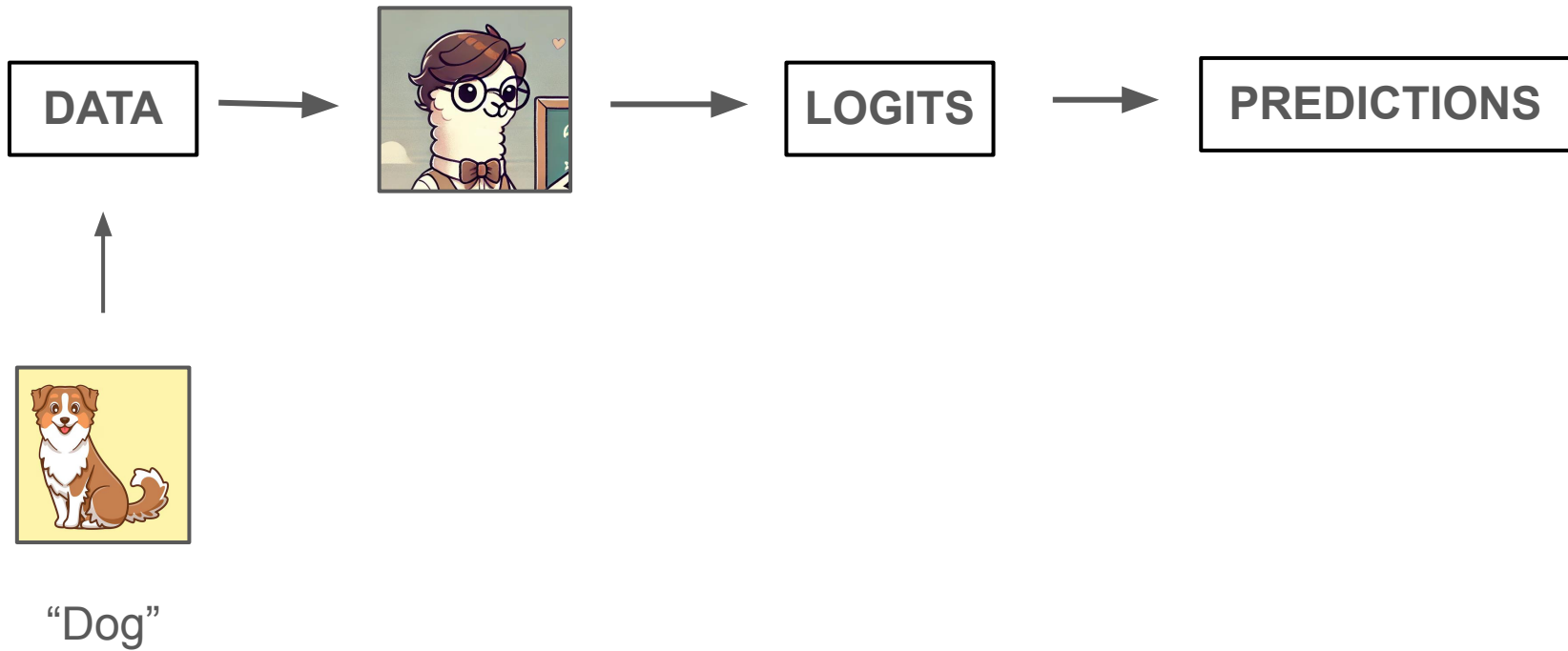


Pretty straightforward. Just train a **big** model with a **lot** of parameters and make it as **good** as possible.

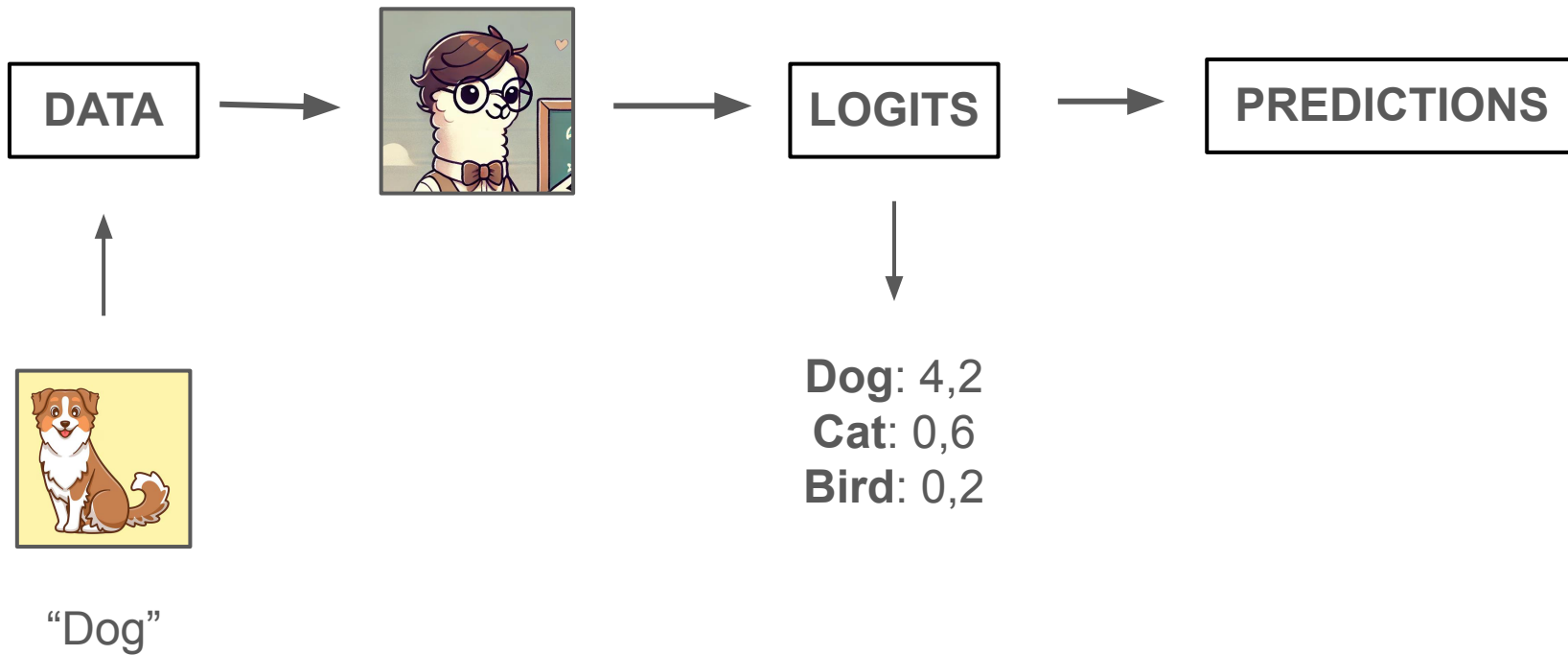
Training the **Teacher**



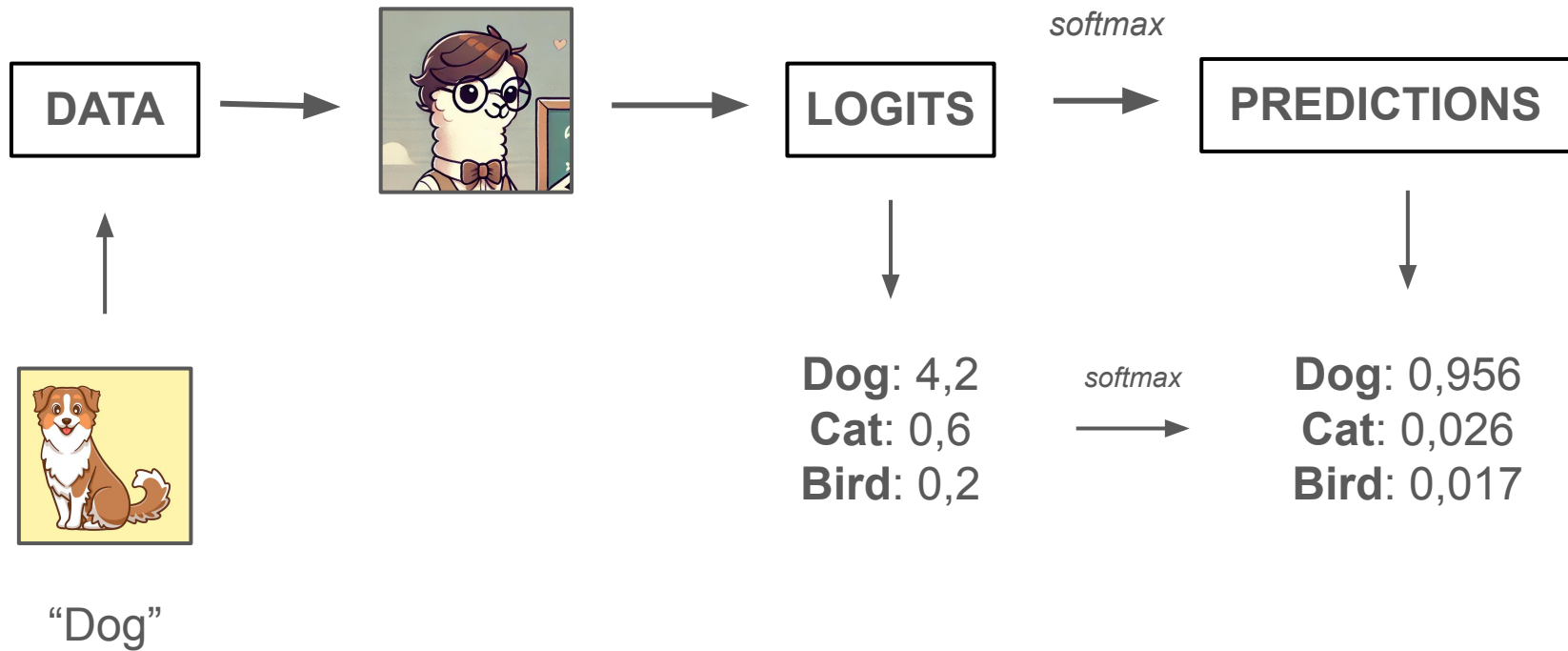
Training the **Teacher**



Training the Teacher



Training the Teacher





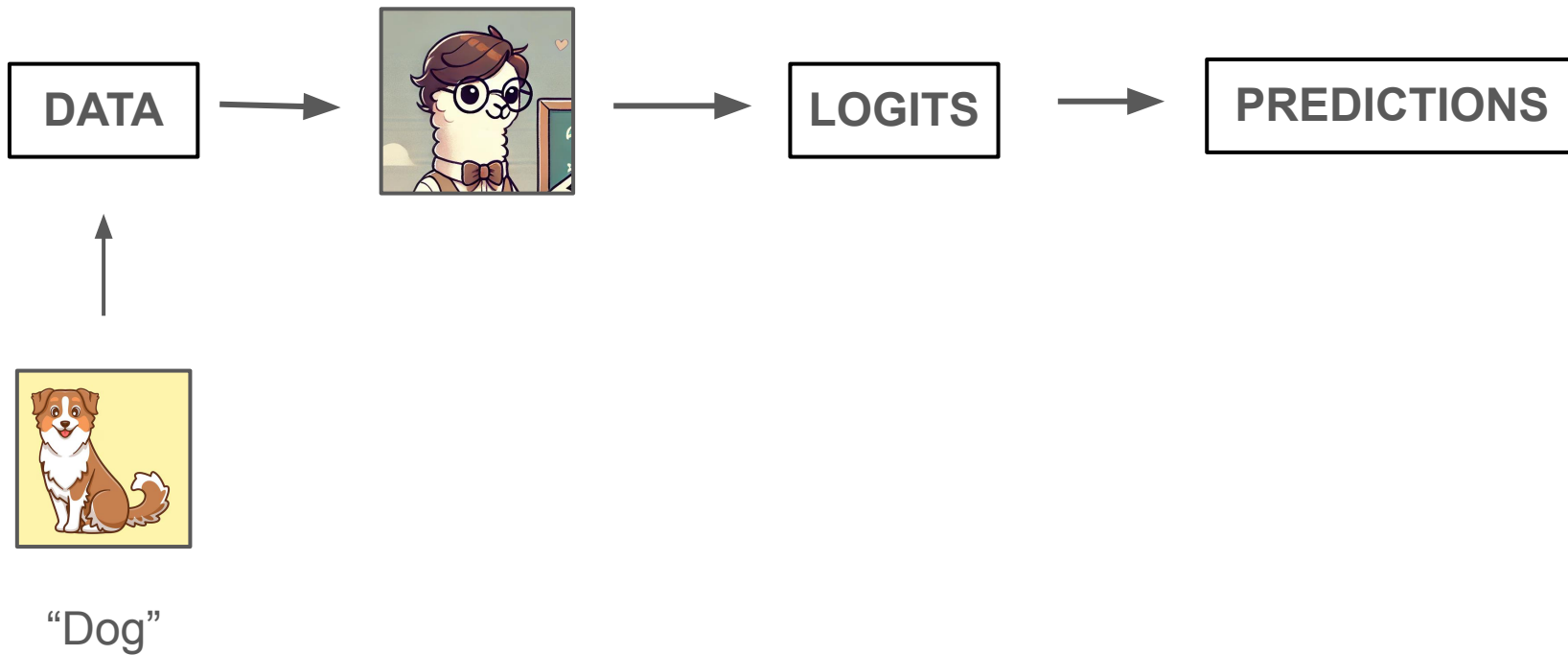
- Train the Teacher(s)
- 2. Understand how Temperature and Soft Labels work
- 3. Train the Student



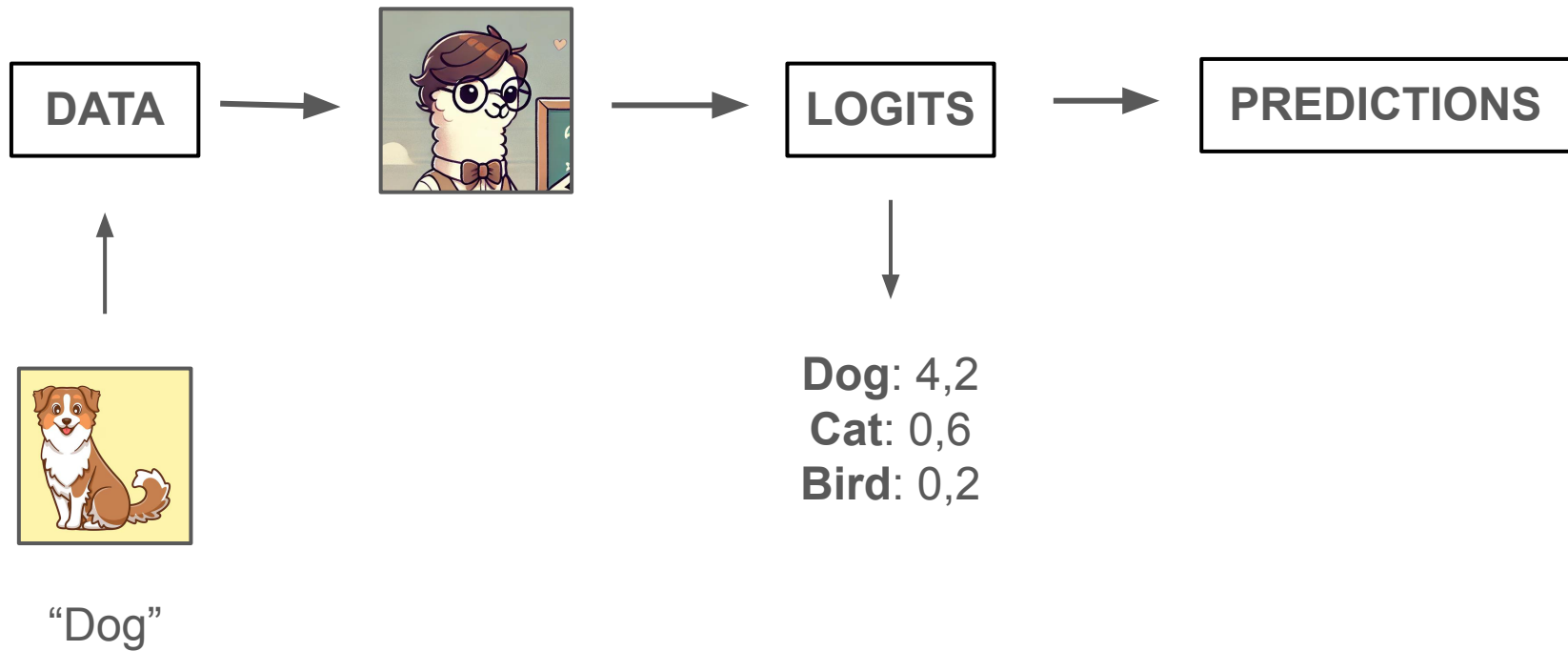
Train the Teacher(s)

2. **Understand how Temperature and Soft Labels work**
3. Train the Student

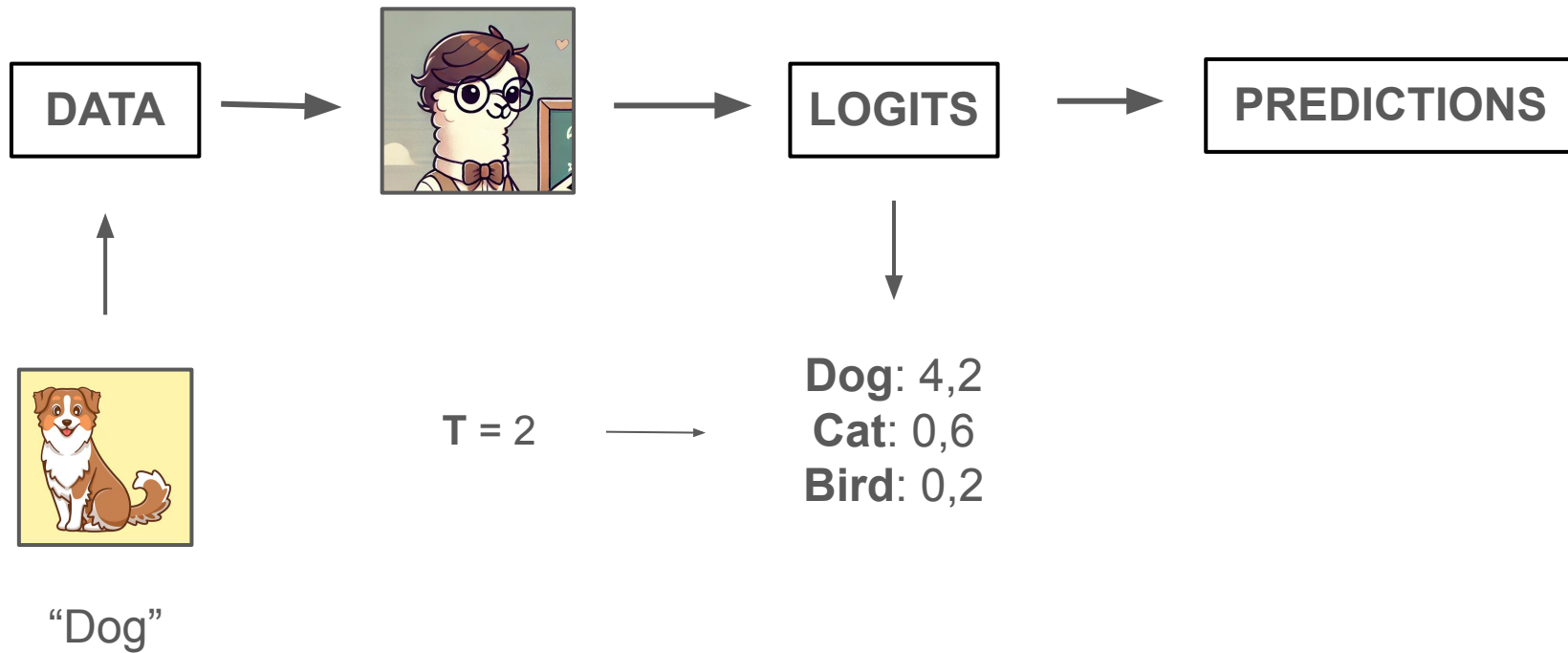
Temperature and Soft Labels



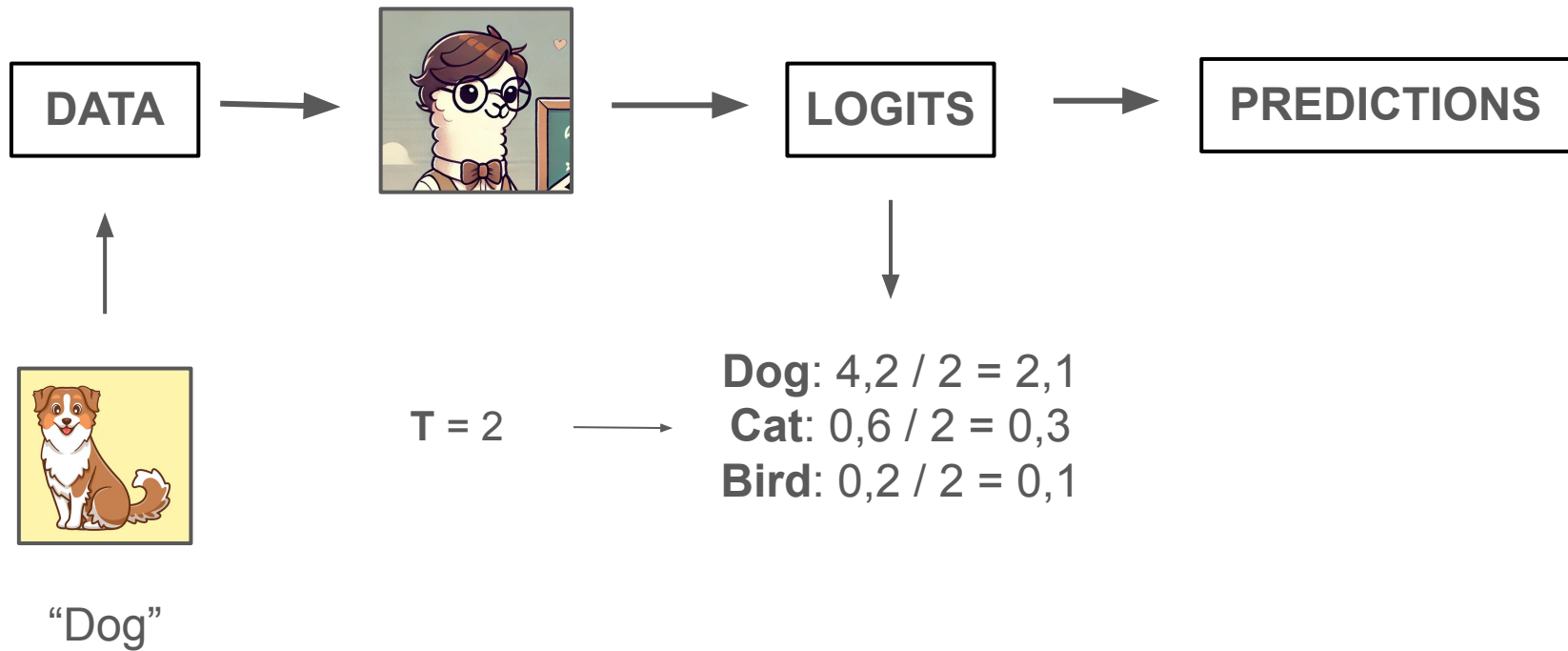
Temperature and Soft Labels



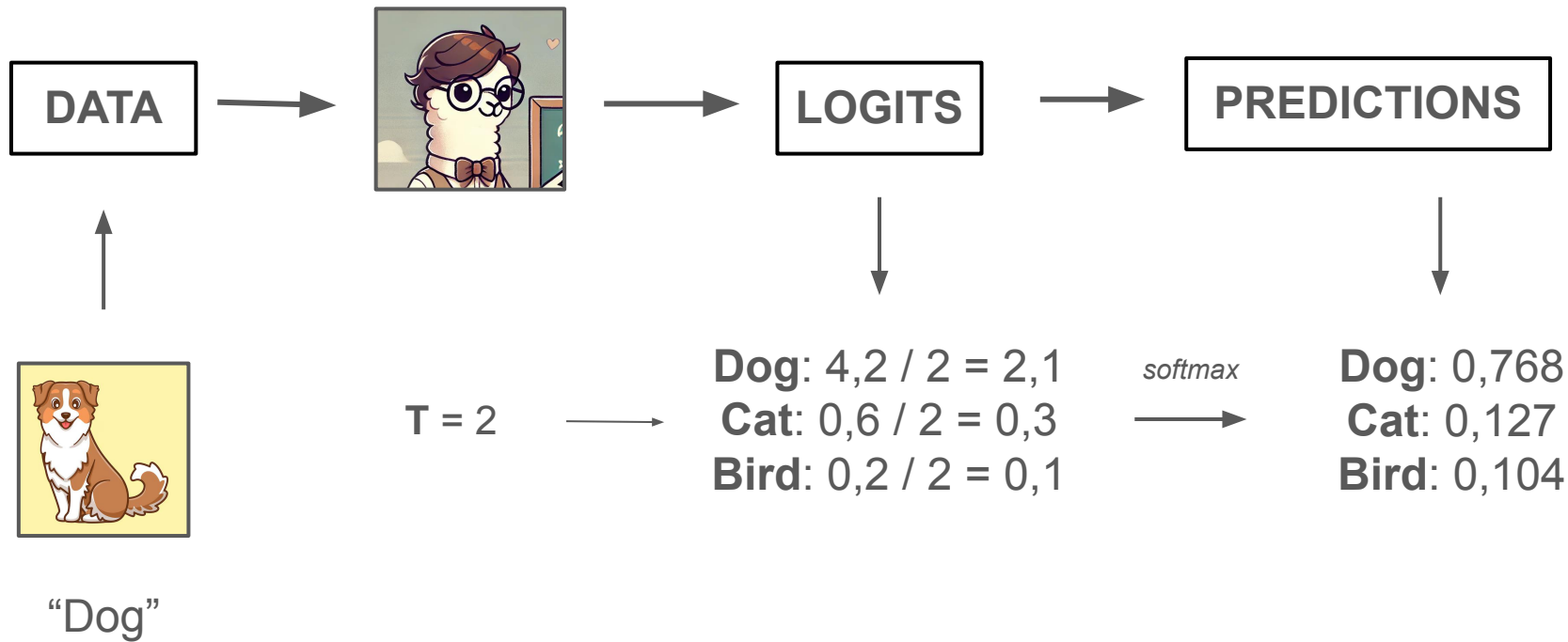
Temperature and Soft Labels



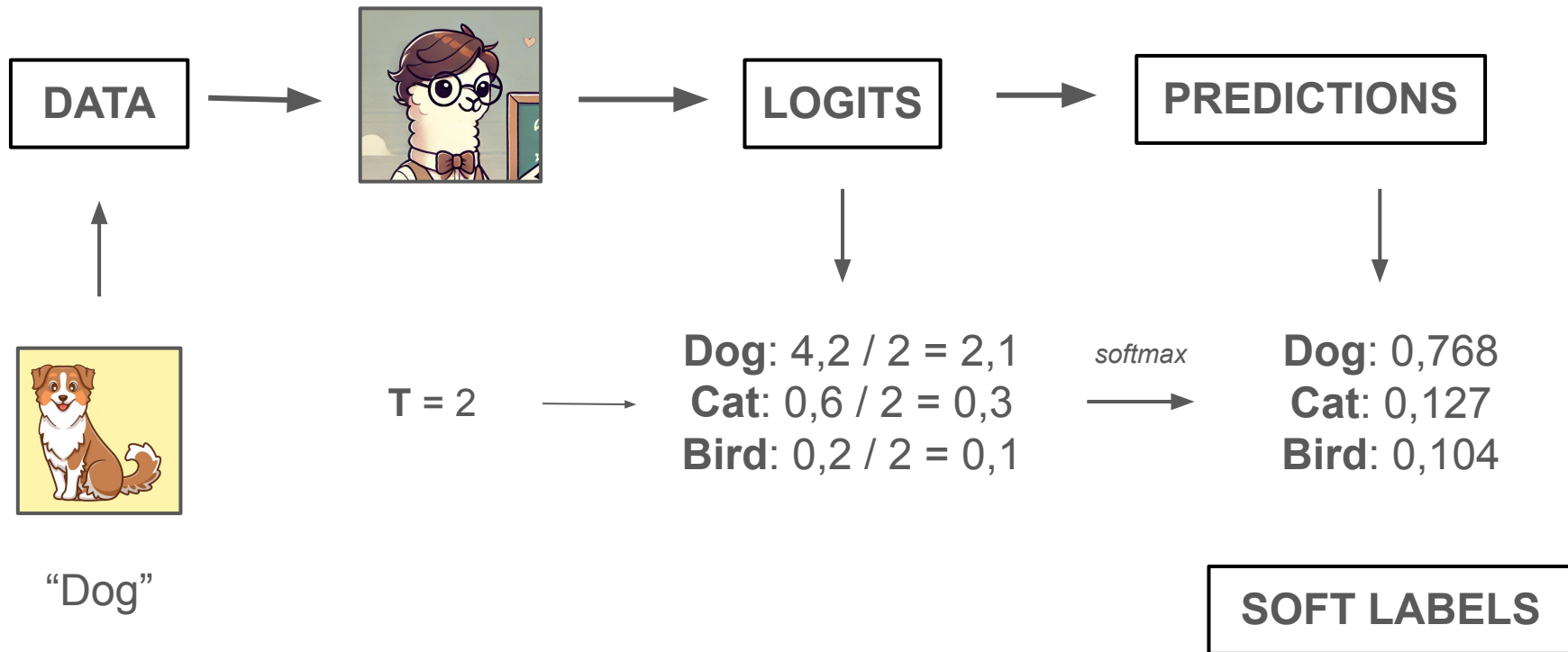
Temperature and Soft Labels



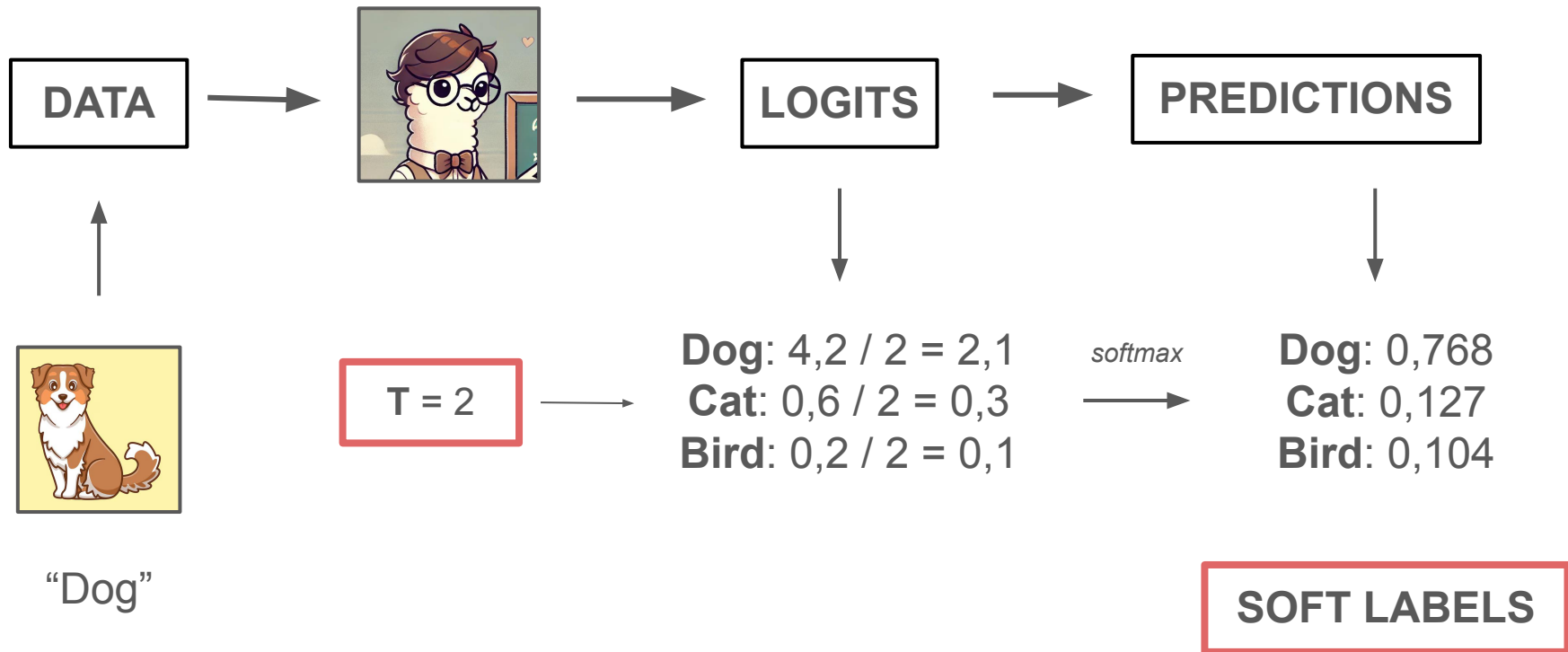
Temperature and Soft Labels



Temperature and Soft Labels



Temperature and Soft Labels





Train the Teacher(s)



Understand how Temperature and Soft Labels work

3. Train the Student



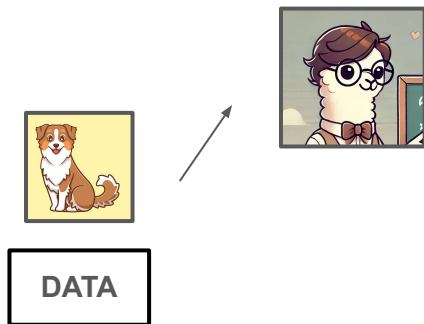
Train the Teacher(s)



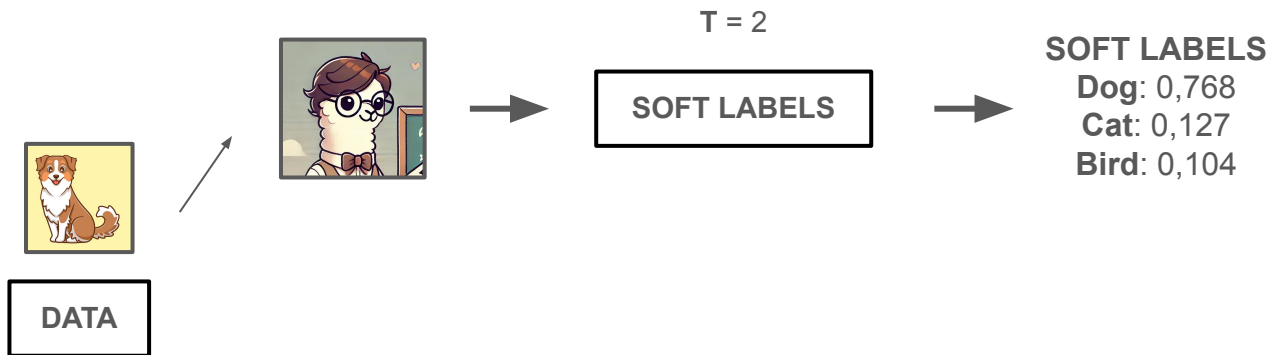
Understand how Temperature and Soft Labels work

3. Train the Student

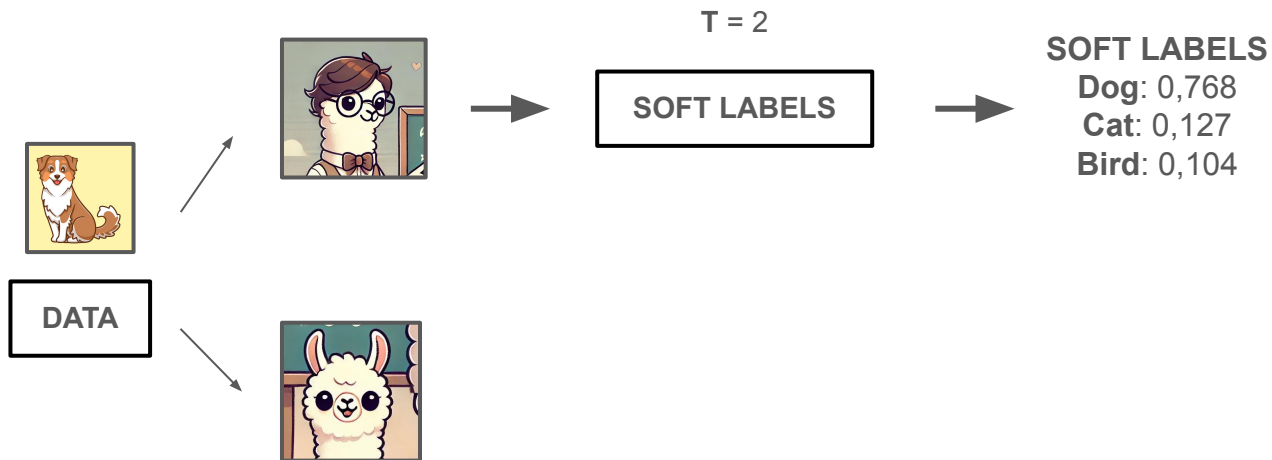
Training the **Student**



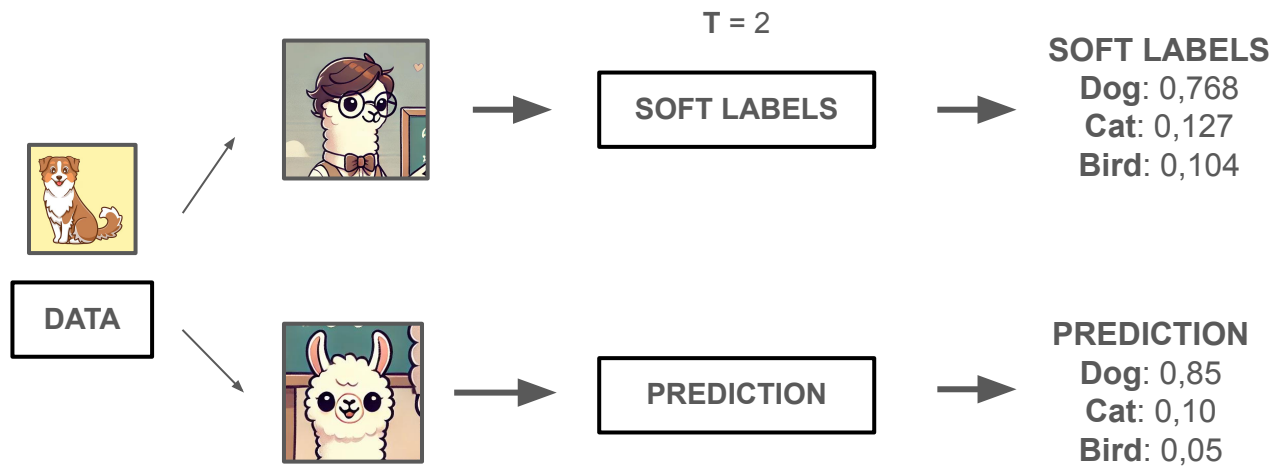
Training the Student



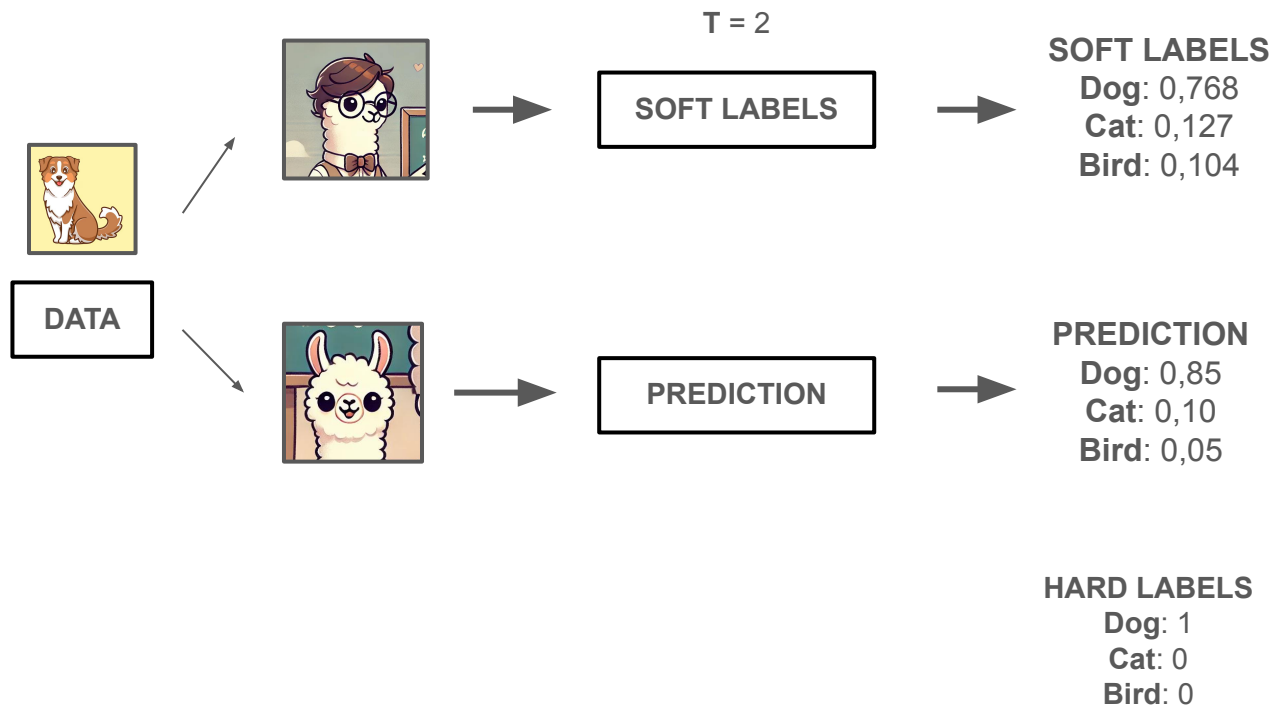
Training the Student



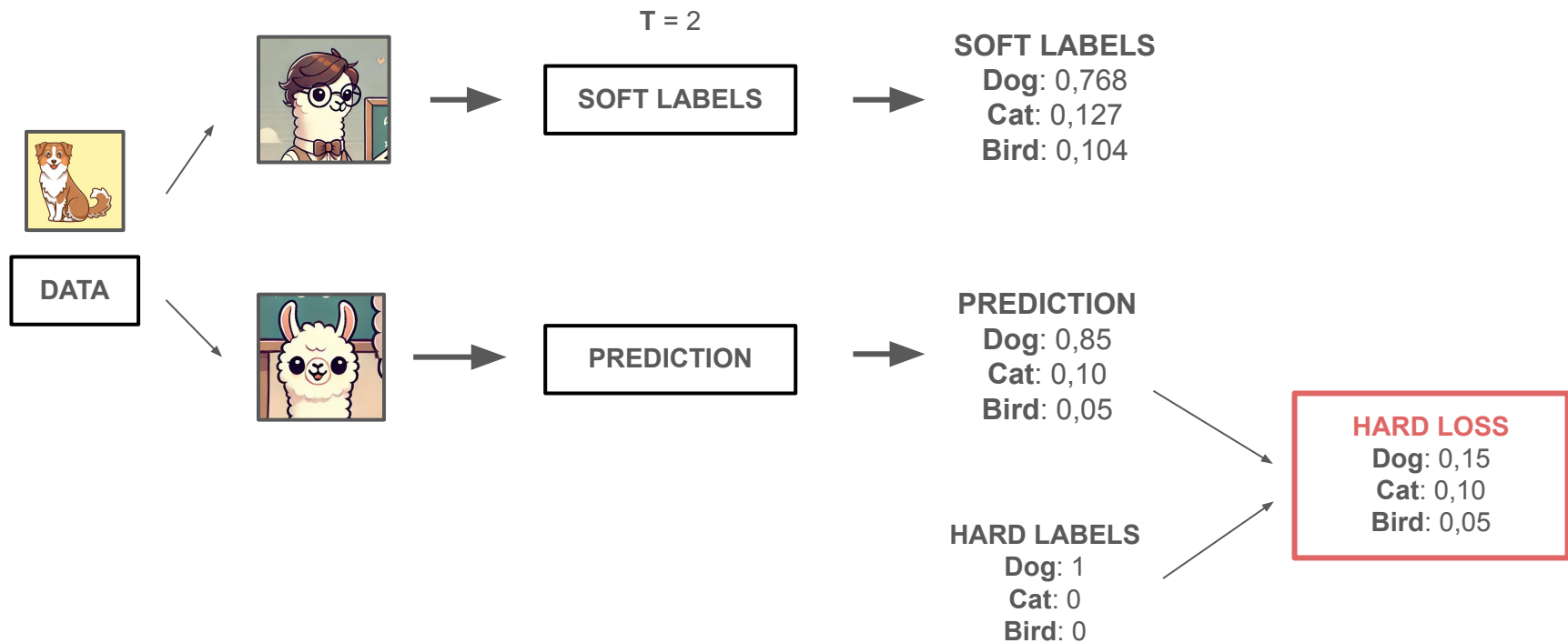
Training the Student



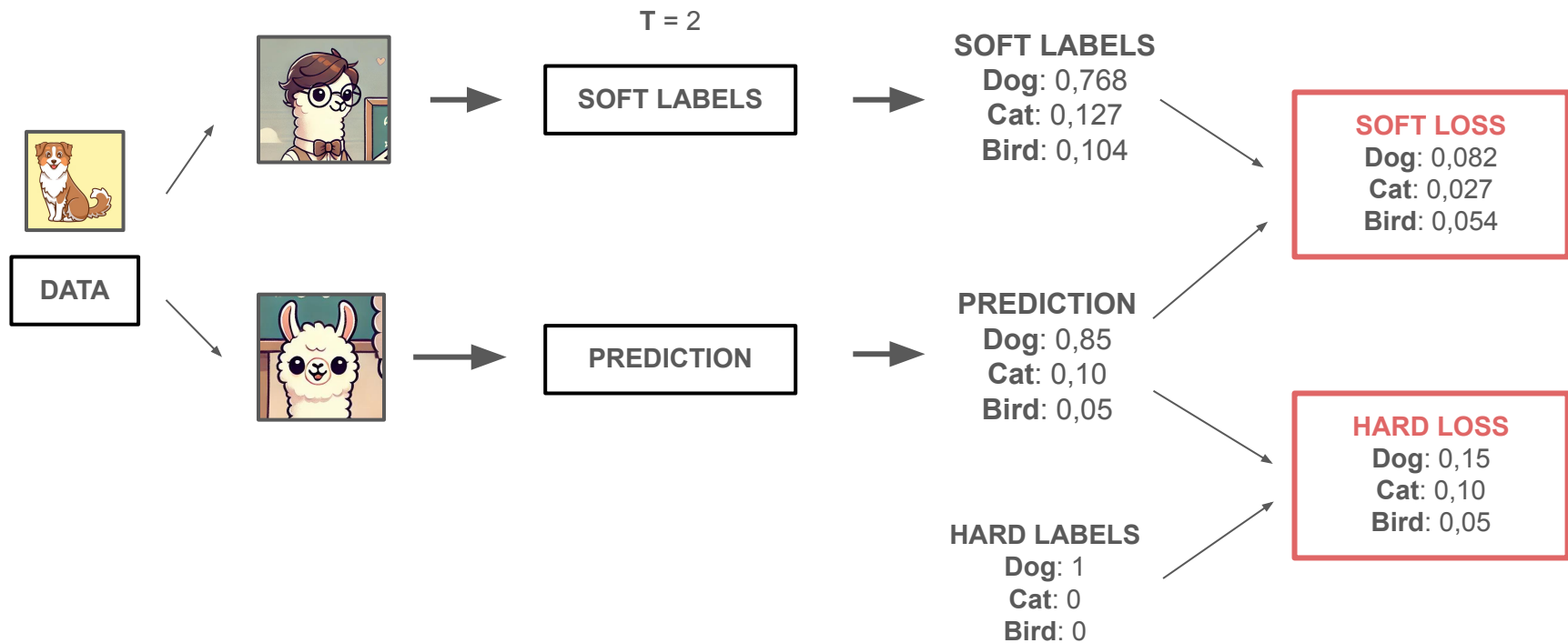
Training the Student



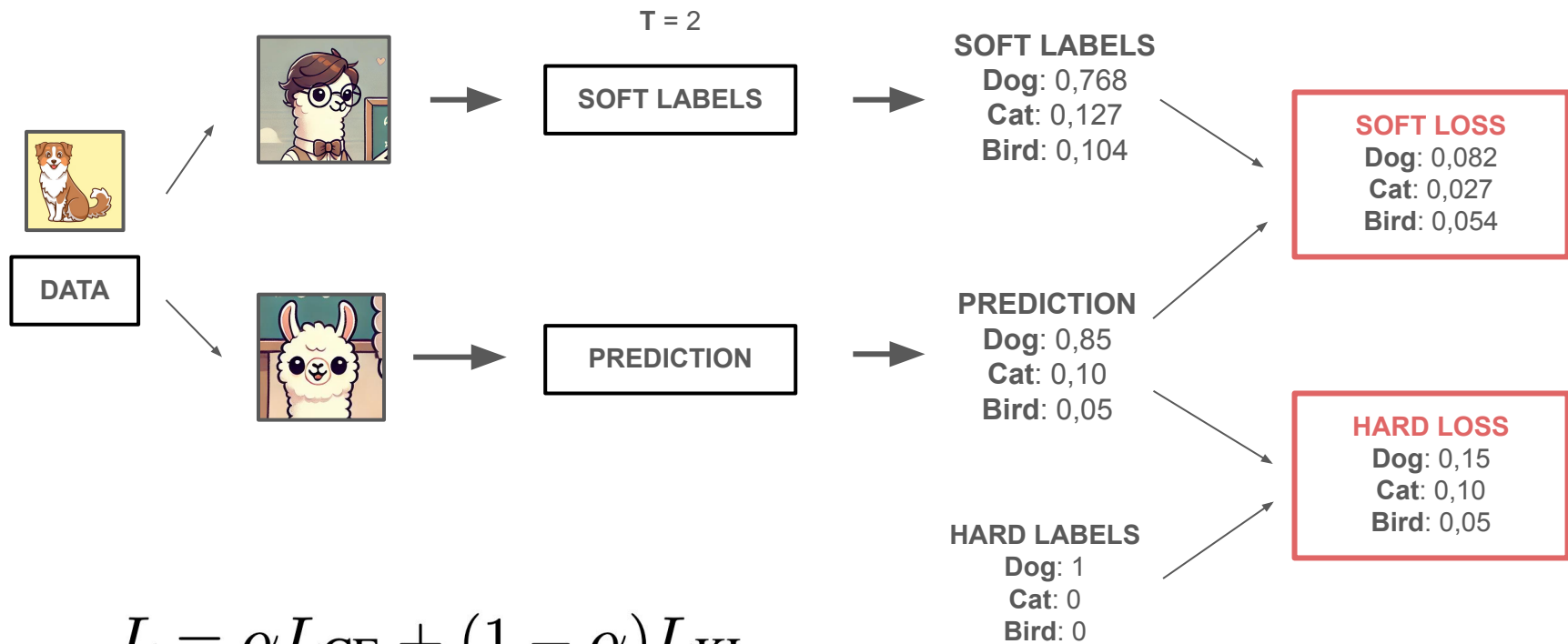
Training the Student



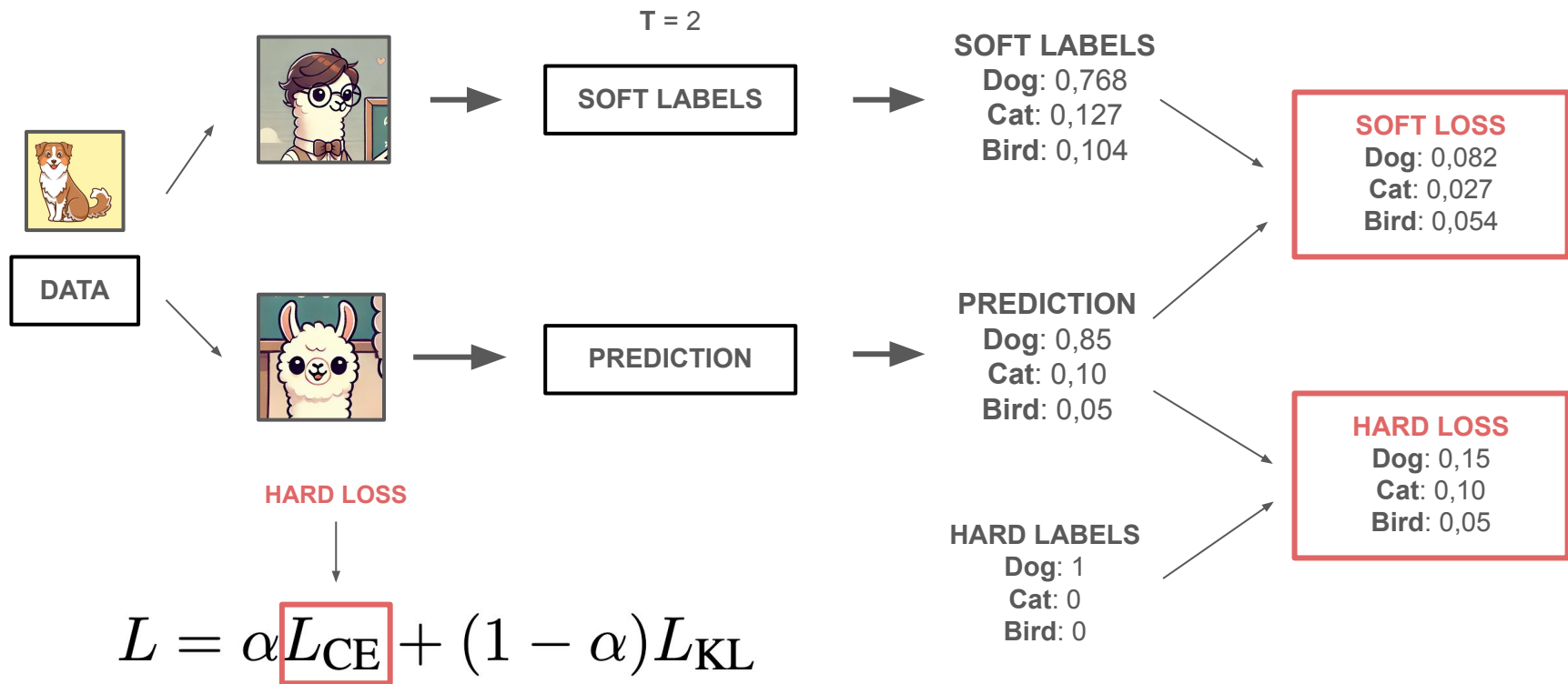
Training the Student



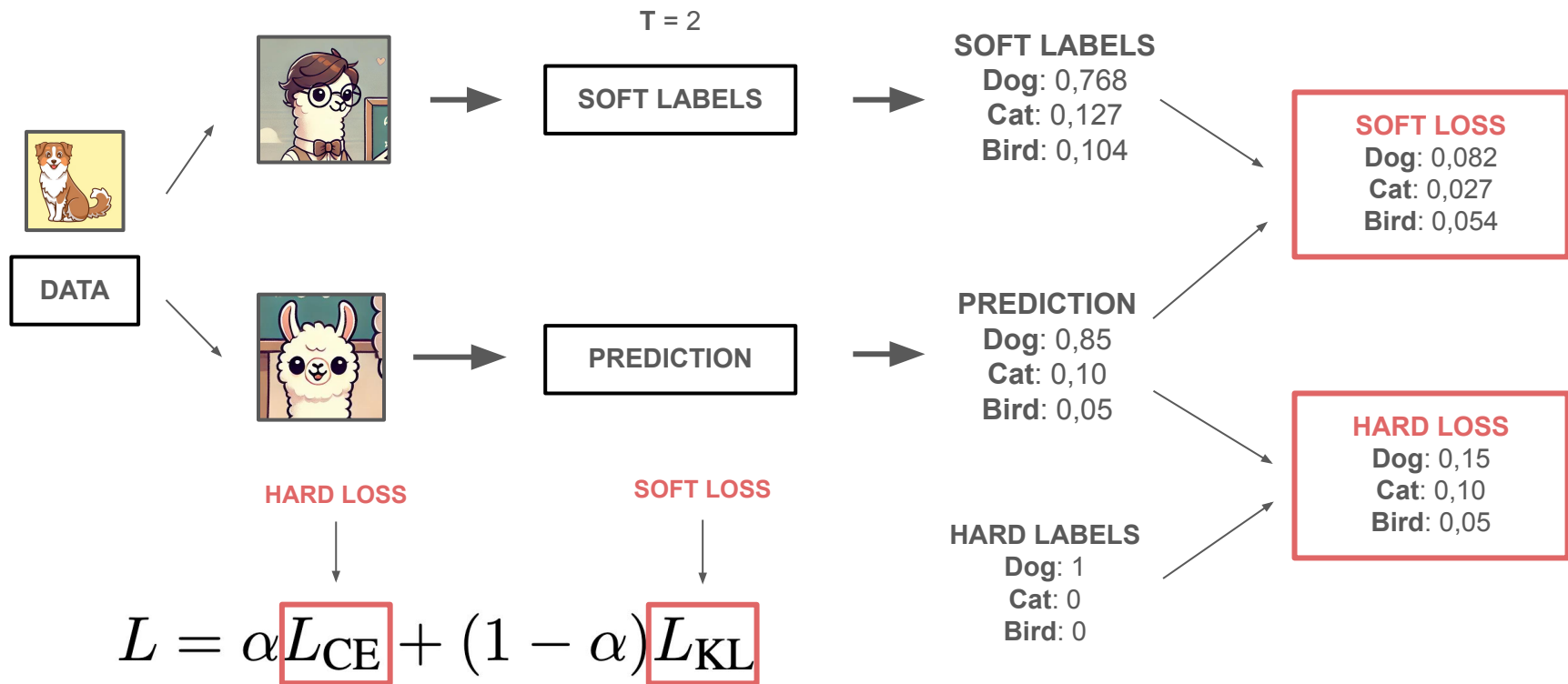
Training the Student



Training the Student



Training the Student



Our implementation

We used 2 teachers (instead of only one)

Our implementation



GPT-2

- 24 layers
- 16 attention heads
- Embedding dimension of 1536
- Intermediate dimension of 6144
- Trainable positional embeddings

=> ~705 million parameters

Our implementation



Llama

- 24 layers
- 8 attention heads
- Embedding dimension of 1024
- Intermediate dimension of 3072
- Fixed rotary positional embeddings

=> ~360 million parameters

Our implementation



BabyLlama

- 16 layers
- 8 attention heads
- Embedding dimension of 512
- Intermediate dimension of 1023

=> **just ~58 million parameters**

Our implementation



$$L = \alpha L_{\text{CE}} + (1 - \alpha) L_{\text{KL}}$$

- α : 0.5
- T: 2
- **Soft Labels** of the 2 teachers are averaged

	Learning rate	Batch size	Epochs	Warm-up steps
GPT-2	$3 \cdot 10^{-4}$	256	6	300
Llama	$2.5 \cdot 10^{-4}$	128	4	300
BabyLlama	$2.5 \cdot 10^{-4}$	32	6	200

Results

Used 2 benchmarks:

Results

Used 2 benchmarks:

- BLiMP
 - zero-shot benchmark, no further training needed
 - measures ability of model to capture linguistic phenomena (syntax, semantics, etc)
 - 67,000 sentence pairs into 67 tasks
 - model predicts probability between grammatical and ungrammatical sentences of the pair

Results

Used 2 benchmarks:




- BLiMP
 - zero-shot benchmark, no further training needed
 - measures ability of model to capture linguistic phenomena (syntax, semantics, etc)
 - 67,000 sentence pairs into 67 tasks
 - model predicts probability between grammatical and ungrammatical sentences of the pair
- BLiMP Supplementary
 - Extension of BLiMP
 - More complex and diverse tasks

Results




Comparisons with:

- 3 baseline models from BabyLM Challenge:
 - RoBERTa (125M parameters)
 - OPT (125M parameters)
 - T5 (222M parameters)
- The 2 teacher models




Results - BLiMP

					
RoBERTa	OPT	T5	GPT2	Llama	BabyLlama
125M	125M	222M	705M	360M	58M
69.5	62.6	58.8	64.5	62.8	64.7

Results - Blimp Supplementary

					
RoBERTa	OPT	T5	GPT2	Llama	BabyLlama
125M	125M	222M	705M	360M	58M
47.5	54.7	43.9	49.3	48.2	47.6

Results - Scaled BLiMP and BLiMP Suppl.

					
RoBERTa	OPT	T5	GPT2	Llama	BabyLlama
125M	125M	222M	705M	360M	58M
0.55	0.50	0.26	0.09	0.17	1.11
0.38	0.43	0.19	0.06	0.13	0.82

We also trained a model on the whole dataset (38M)
These are the results...

Results

Some qualitative results of the **Full Model**:

Prompt: “<s> Italy is”

Generated: Italy is divided into two great parts, the north and the south, and the central and south-west; while the south, including the greater part of the peninsula, and the south, the eastern and the northern parts, are separated by the great plain of Italy. The whole of the western and western part of the peninsula, except the peninsula of the Mediterranean, is a very fertile plain, and in some places has been made the capital of the kingdom of Italy. The whole of northern Italy is traversed by the main line of the Italian railway.

Results

Prompt: “<s> London was”

Generated: London was the capital of the district of the East Anglian kingdom of North Wales, the chief town of which was Beverley. It was the seat of the royal family of the earls of South Wales, who held it in the time of Edward I. It was a royal burgh of barony, and held the manor of Fordun in the time of Edward I.

Results

Prompt: “<s> January”

Generated: January, the term given to the season of the year, and the first month of the year. The term is also used for the day of the month in which the first day of the year is divided into a month and a month, the term being also used of the month of the year. The term is also applied to the month of the year in which the mean or mean sun is the same as the month of the year. The term is also used in a more general sense of the term, but it is more particularly used of the months of the year in which the season of the year is divided.

Conclusion - Future Work

- Knowledge distillation has **great** results
- **Small** but **powerful** models
- Small datasets are **enough** (if they are good...)
- Pretraining preserves knowledge in the model
- Fine Tuning or RAG can give more accurate generations

Questions?