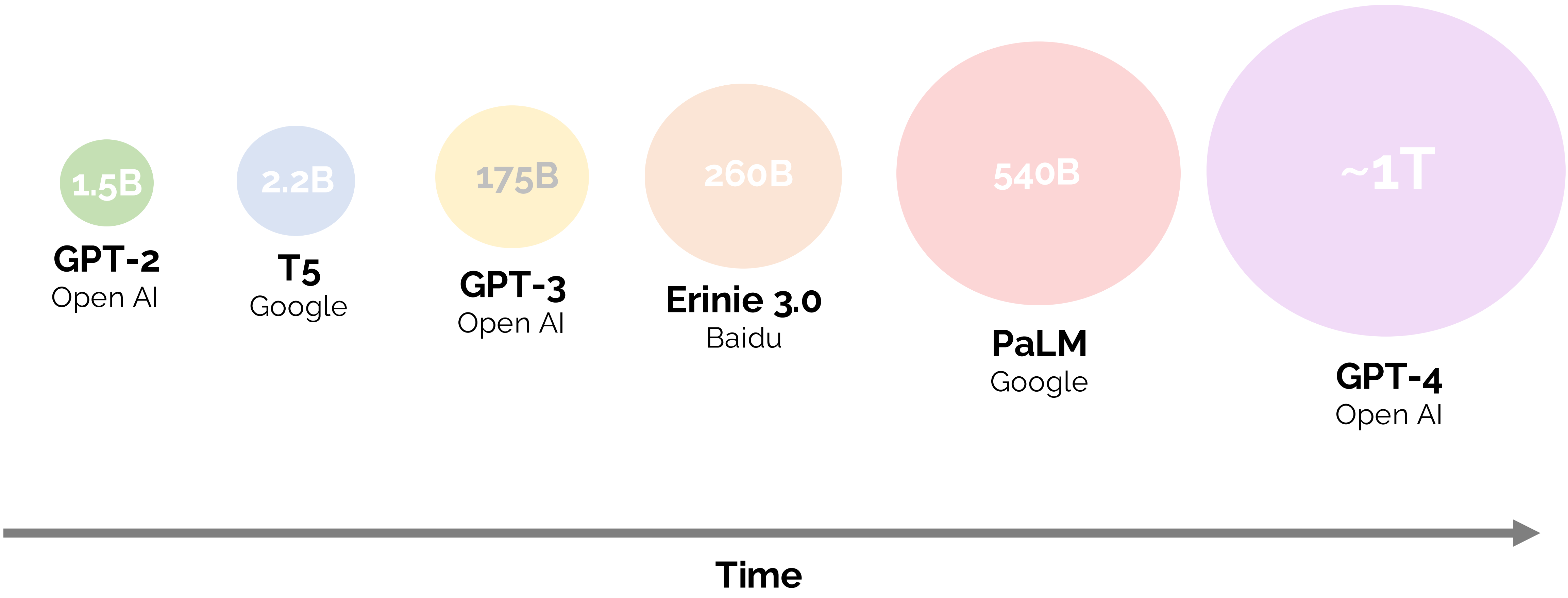
Probabilistic Token Alignment for Large Language Model Fusion

Runjia Zeng
RIT

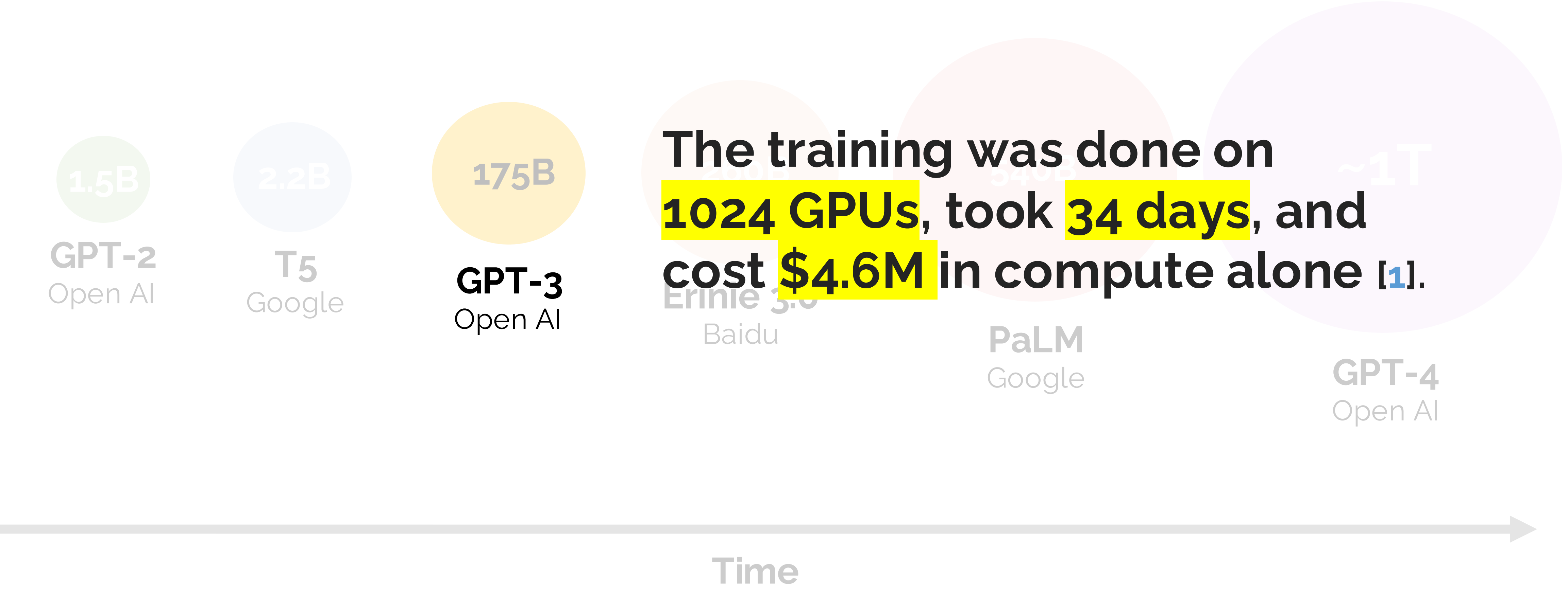
Content

Background: Curse of Scaling Law

1. Curse of Scaling Law: Parameter Explosion



1. Curse of Scaling Law: Parameter Explosion



The training was done on 1024 GPUs, took 34 days, and cost $4.6M in compute alone [1].

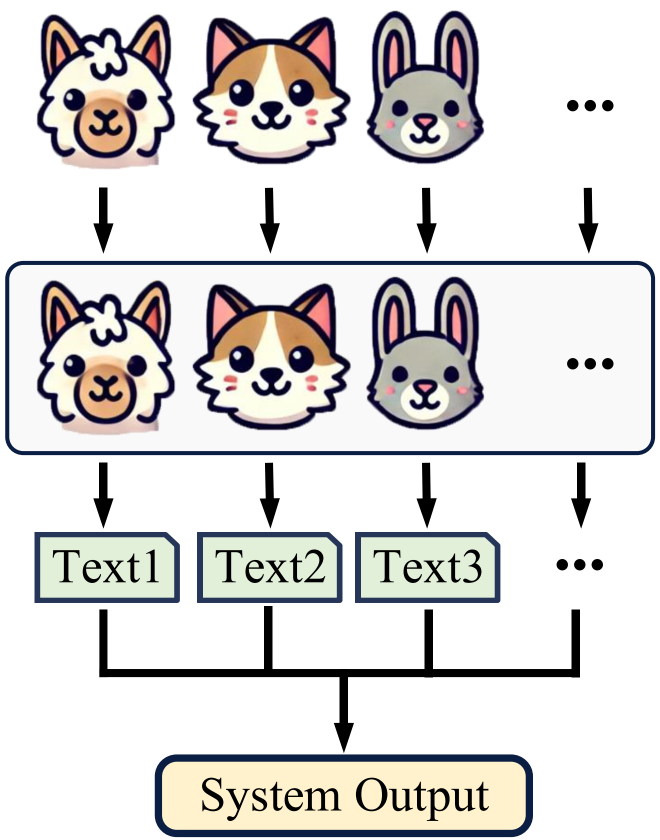
[1] Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM (2021).

Question 1:

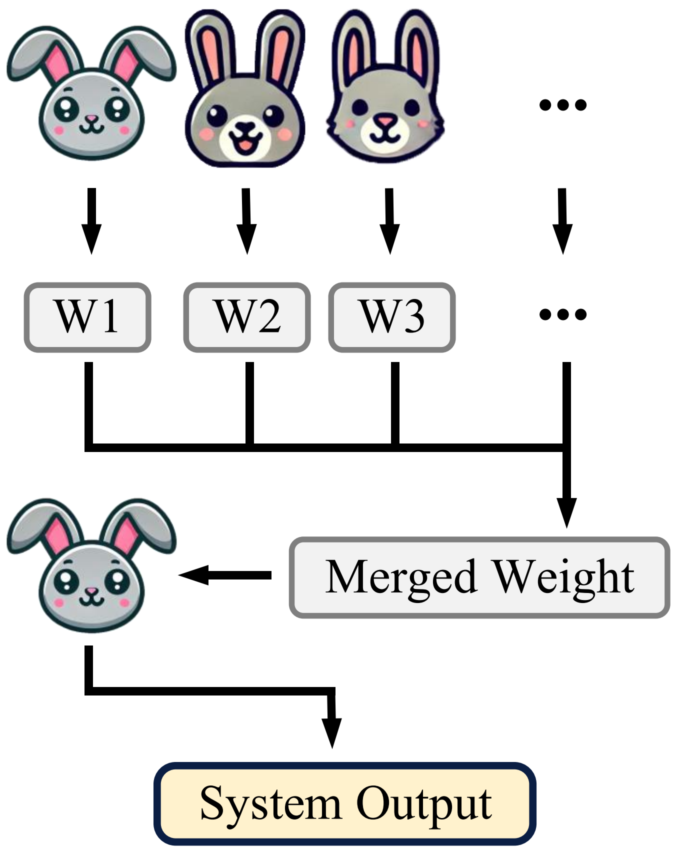
How can we construct stronger baselines without resorting to the naive application of scaling laws?

Preliminary: Model Fusion

2. Preliminary: Weight Merging

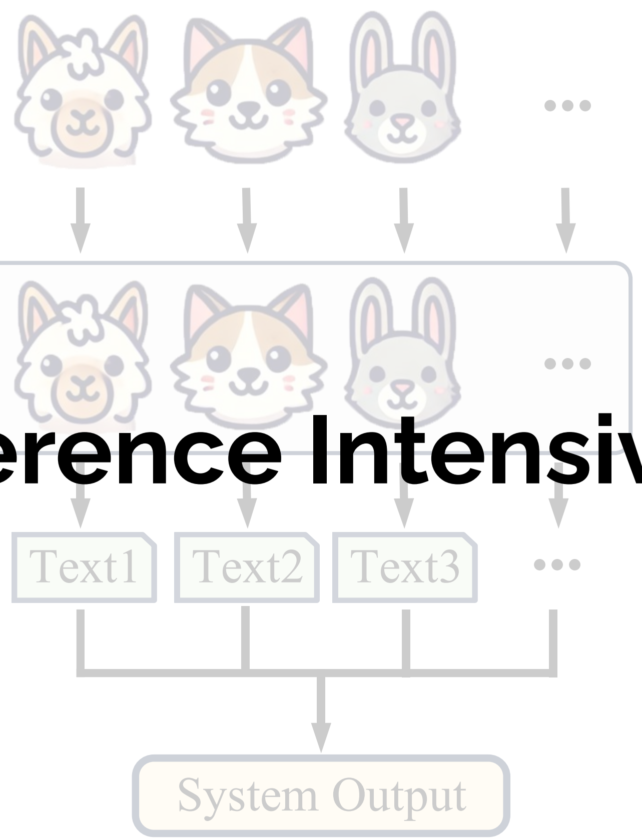


Model Emsemble



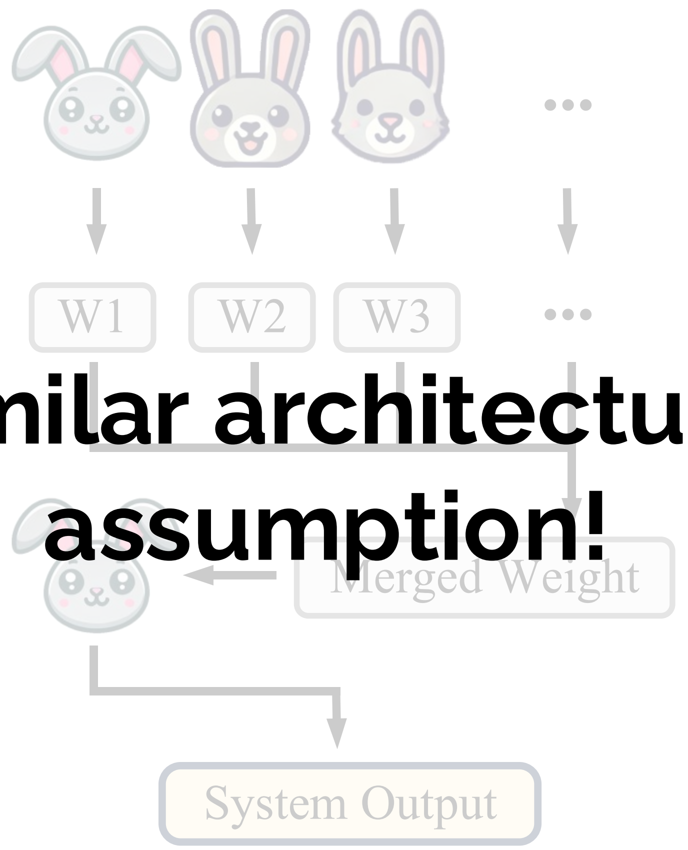
Weight Merging

2. Preliminary: Weight Merging



Model Emsemble

Inference Intensive!

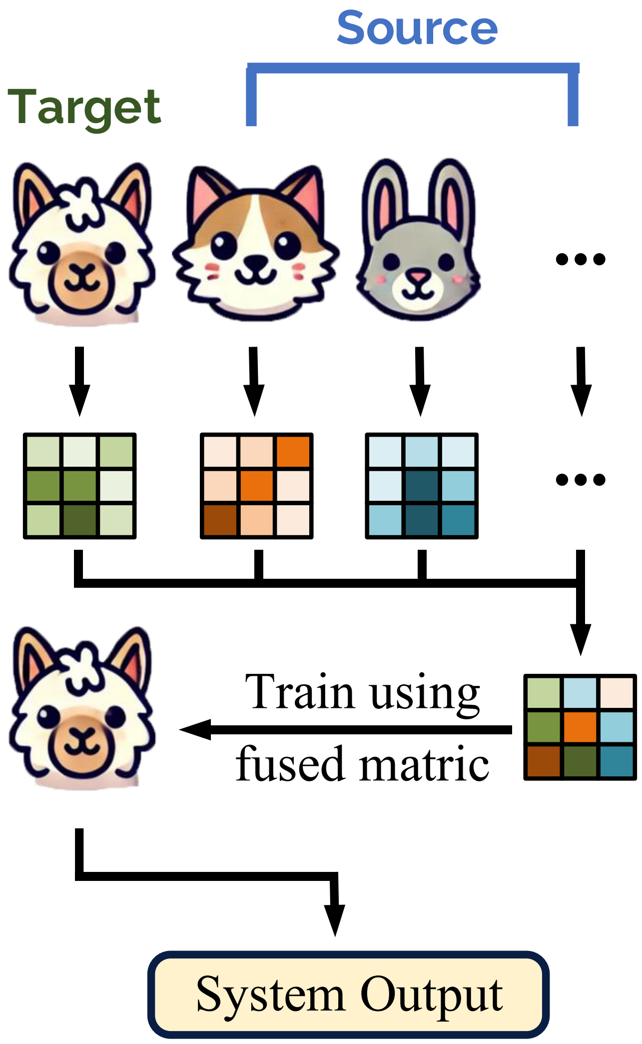


Weight Merging

Similar architecture assumption!

New Alternative: Probabilistic Token Alignment (PTA) For Knowledge Fusion

3. New Alternative: Knowledge Fusion



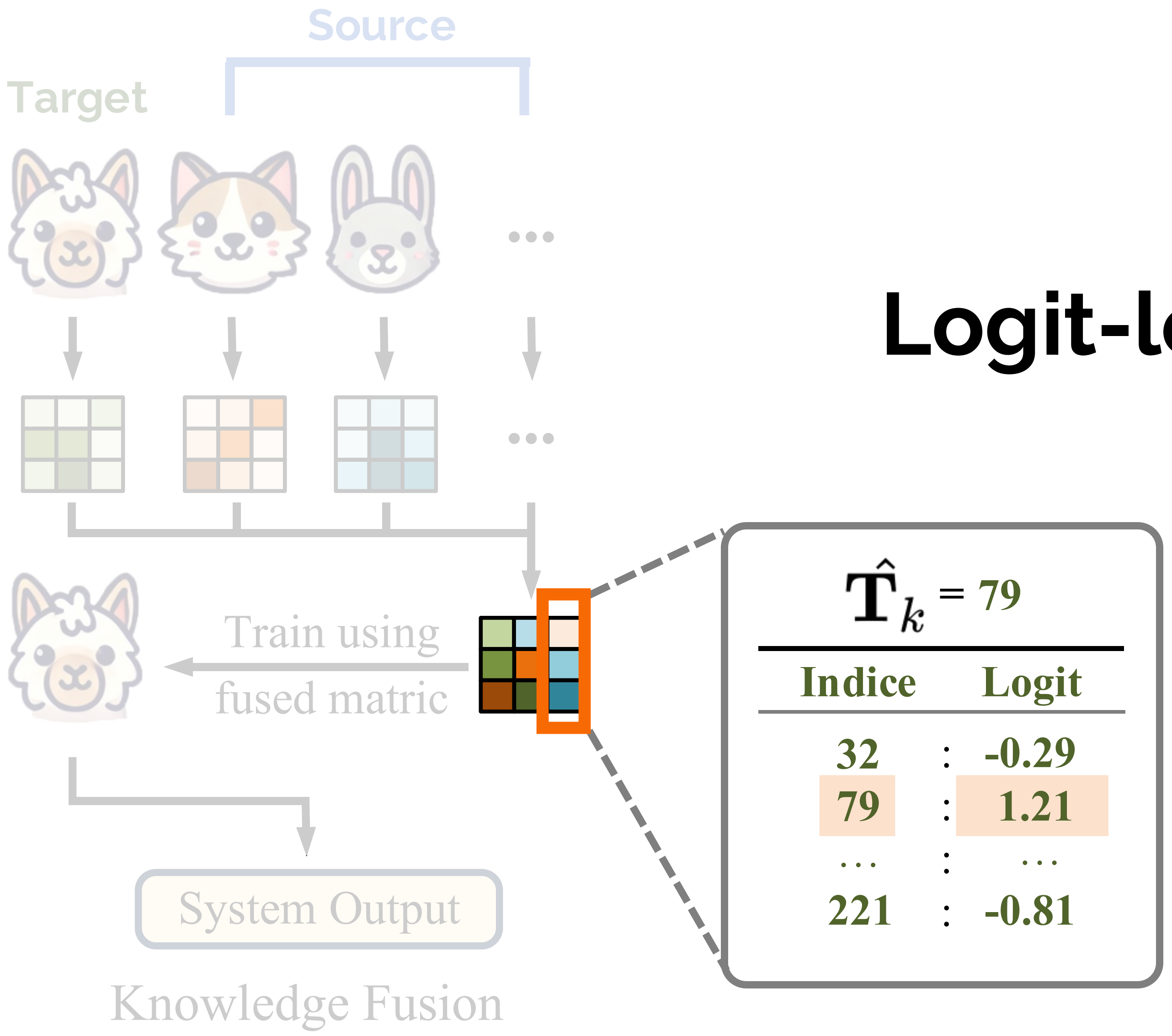
Knowledge Fusion

Logit-level Fusion

Architecture-agnostic!

Efficient Inference!

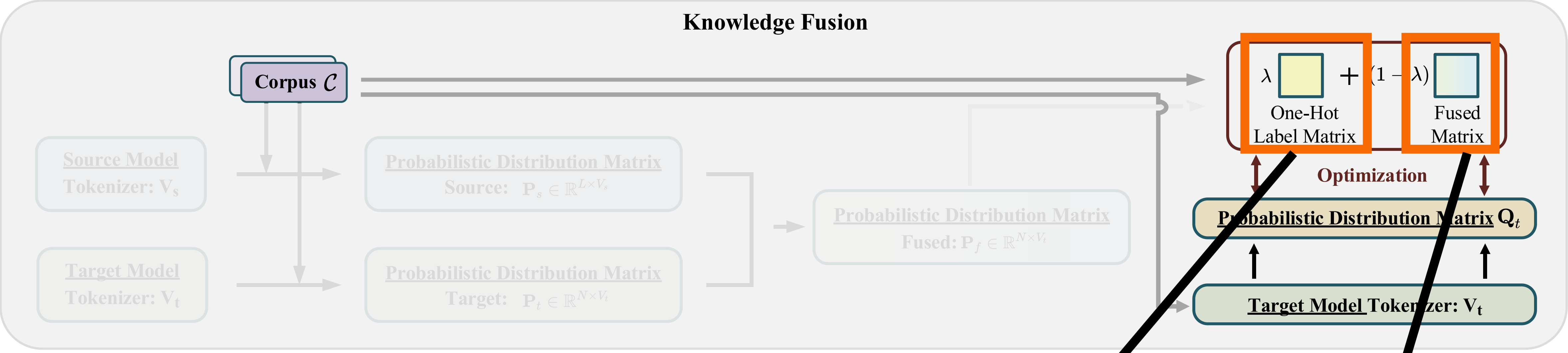
3. New Alternative: Knowledge Fusion



Logit-level Fusion

3. New Alternative: Knowledge Fusion

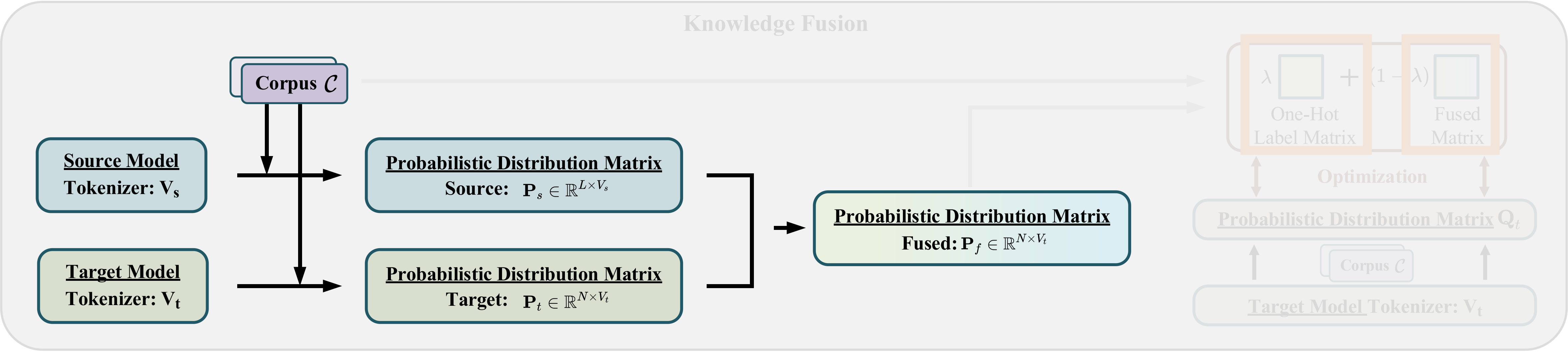
$$\mathcal{L} = \lambda \mathcal{L}_{\text{CLM}} + (1-\lambda)\mathcal{L}_{\text{Fusion}}$$



$$\mathcal{L}_{\text{CLM}} = -\mathbb{E}_{t\sim\mathcal{C}}\left[\mathbb{D}(\mathbf{Q}_t, \mathbf{O}_t)\right]$$

$$\mathcal{L}_{\text{Fusion}} = -\mathbb{E}_{t\sim\mathcal{C}}\left[\mathbb{D}(\mathbf{Q}_t, \mathcal{P}_f)\right]$$

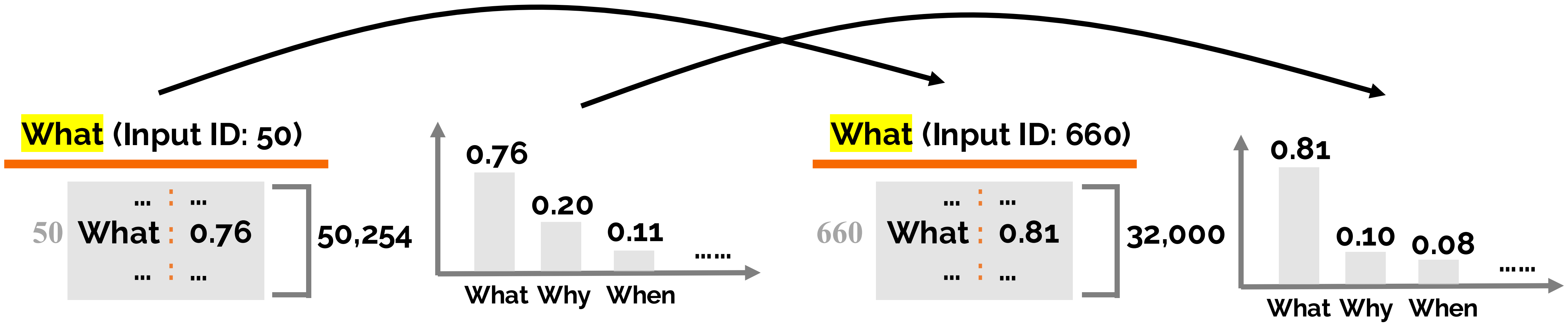
3. New Alternative: Knowledge Fusion



Question 2:

How can we align various LLMs with different tokenizers?

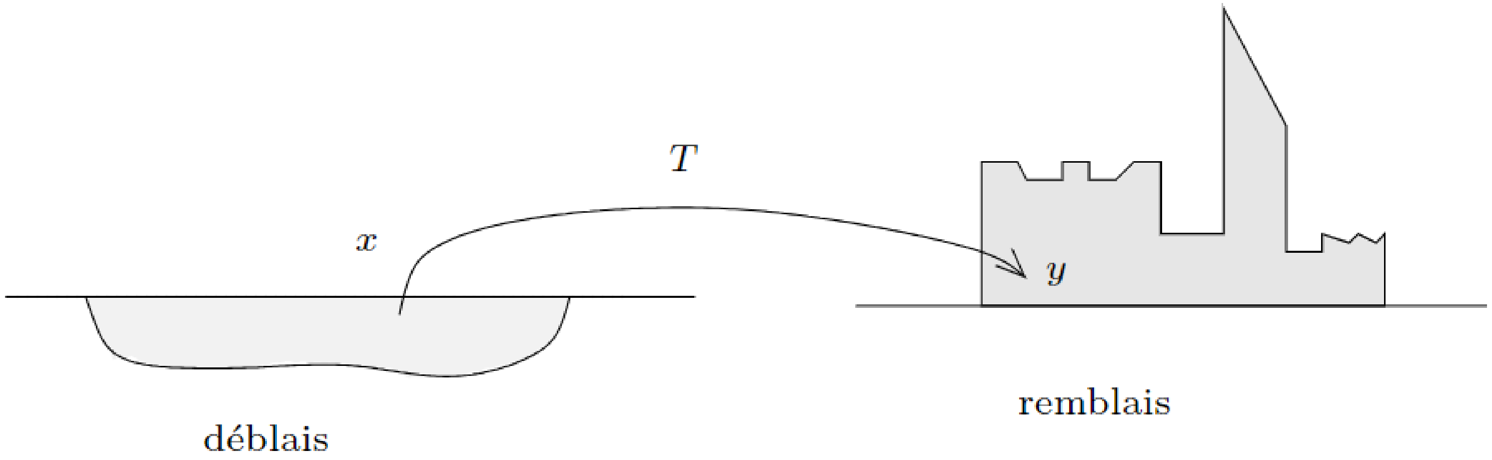
3. New Alternative: Knowledge Fusion



Transport from one token distribution to another!

3. New Alternative: Optimal Transport

Gaspard Monge
French Geometer
The founding fathers of optimal transport



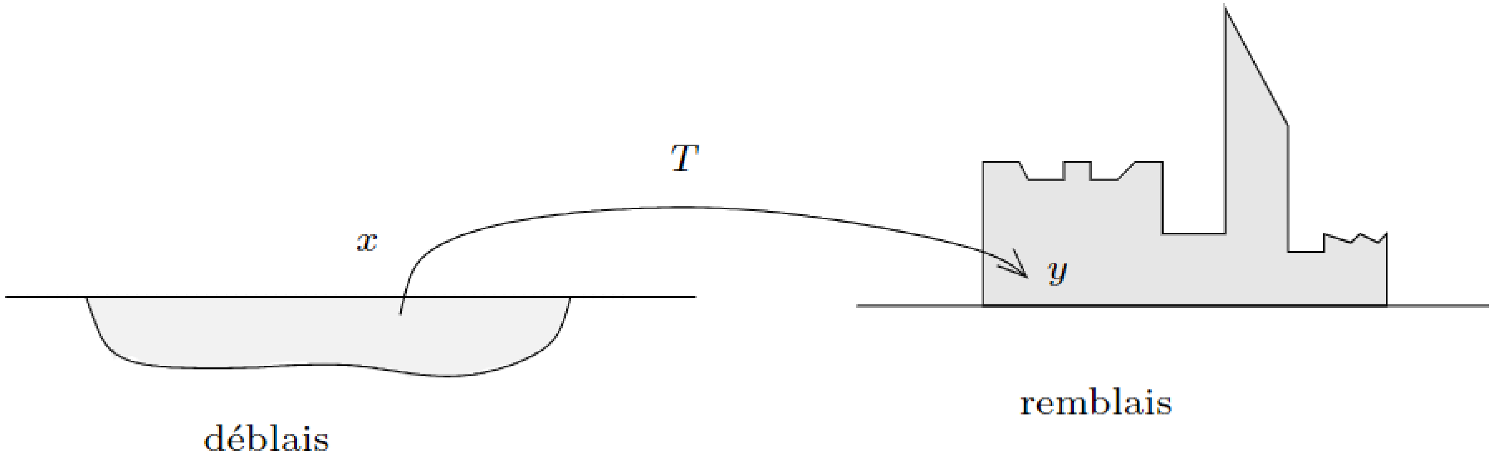
Monge's problem of déblais and remblais

$$\mathbf{T} = \arg\min \mathcal{L}\left(\mathbf{X}, \mathbf{Y}\right)$$

d éblai: an amount of material that is extracted from the earth or a mine

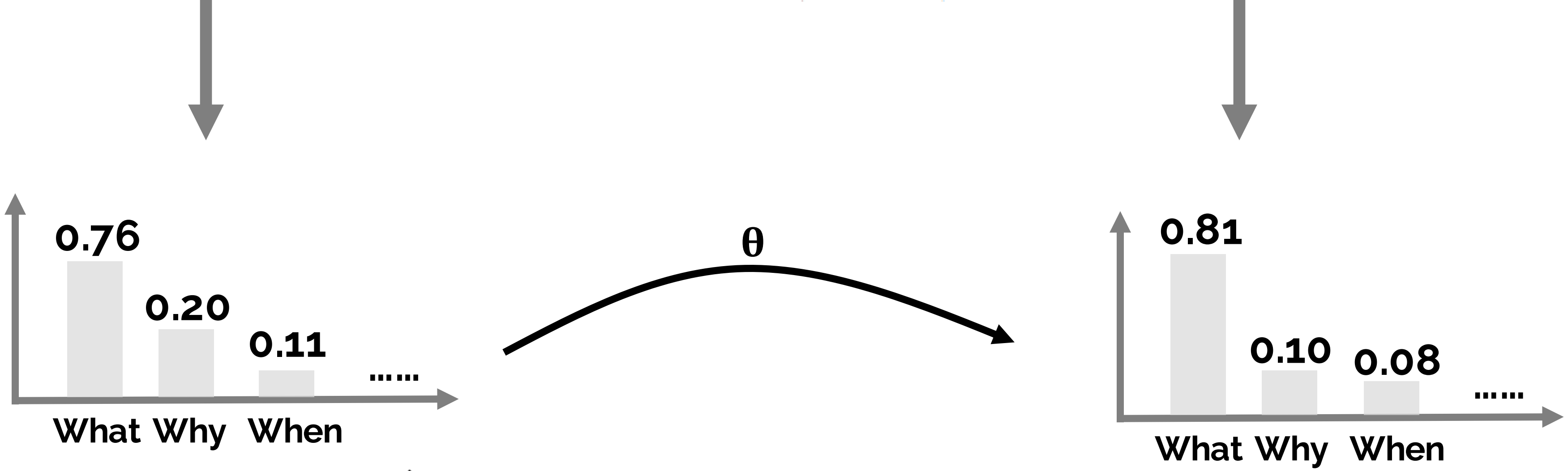
remblai: a material that is input into a new construction

3. New Alternative: Optimal Transport



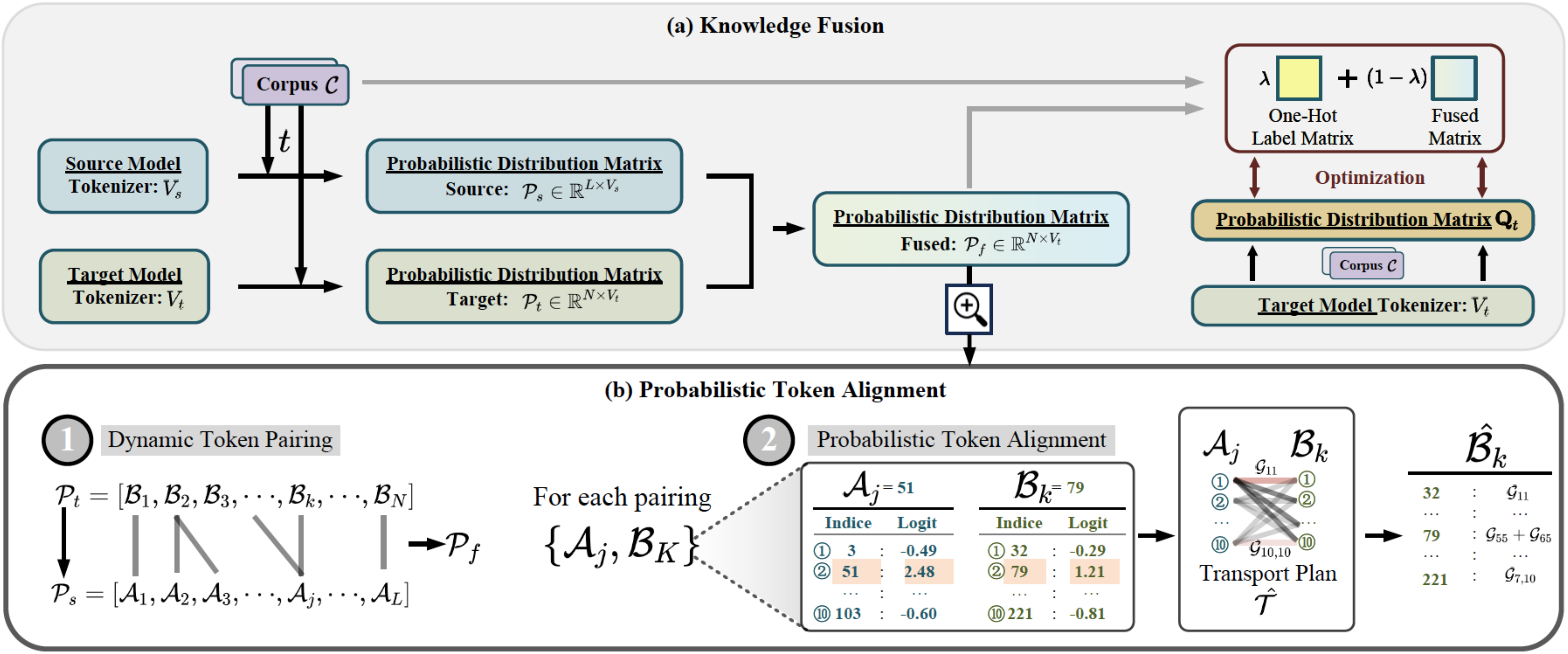
Monge's problem of déblais and remblais

$$\mathbf{T} = \arg\min \mathcal{L}(\mathbf{X}, \mathbf{Y})$$



$$\hat{\mathsf{P}} = \arg\min \mathcal{L}(\mathbf{S}_j, \mathbf{T}_k)$$

3. New Alternative: PTA-LLM



3. **New Alternative:** Dynamic Token Pairing

1 Dynamic Token Pairing

2 Probabilistic Token Alignment

$$\mathcal{P}_t = [\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \cdots, \mathcal{B}_k, \cdots, \mathcal{B}_N]$$

$$\rightarrow \mathcal{P}_f$$

$$\mathcal{P}_s = [\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_L]$$

Brute-force methods to explore all possible combinations?

Computation Intensive!

1 Dynamic Token Pairing

2 Probabilistic Token Alignment

Question 3:

How to evaluate the “best“ combinations?
How can we find the “best“ combinations?

3. New Alternative: Dynamic Token Pairing

1 Dynamic Token Pairing

2 Probabilistic Token Alignment

$$\mathcal{P}_t = [\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3, \cdots, \mathcal{B}_k, \cdots, \mathcal{B}_N] \rightarrow \mathcal{P}_f$$

$$\downarrow$$

$$\mathcal{P}_s = [\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \cdots, \mathcal{A}_j, \cdots, \mathcal{A}_L]$$

Editing one sequence to the other with minimum edit cost using dynamic programming

3. New Alternative: Probabilistic Token Alignment

① Dynamic Token Pairing

② Probabilistic Token Alignment

② Probabilistic Token Alignment

For each pairing

$\{\mathcal{A}_j, \mathcal{B}_K\}$

$\mathcal{A}_j$ = 51

Indice		Logit
① 3	:	-0.49
② 51	:	2.48
…	:	…
⑩ 103	:	-0.60

$\mathcal{B}_k$ = 79

Indice		Logit
① 32	:	-0.29
② 79	:	1.21
…	:	…
⑩ 221	:	-0.81

$$\hat{\mathcal{B}}_k = \text{TokenAlignment}\left(\mathcal{A}_j, \mathcal{B}_k\right)$$

1 Dynamic Token Pairing

2 Probabilistic Token Alignment

Question 4:

How to evaluate the “best“ transportation?
How can we find the “best“ transportation?

3. New Alternative: Probabilistic Token Alignment

1 Dynamic Token Pairing

2 Probabilistic Token Alignment

$$\hat{\mathcal{B}}_k = \mathrm{TokenAlignment}\left(\mathcal{A}_j, \mathcal{B}_k\right)$$

$$\hat{\mathcal{T}} = \arg\min \mathcal{L}\left(\mathcal{A}_j, \mathcal{B}_k\right)$$

Optimal transport problem!

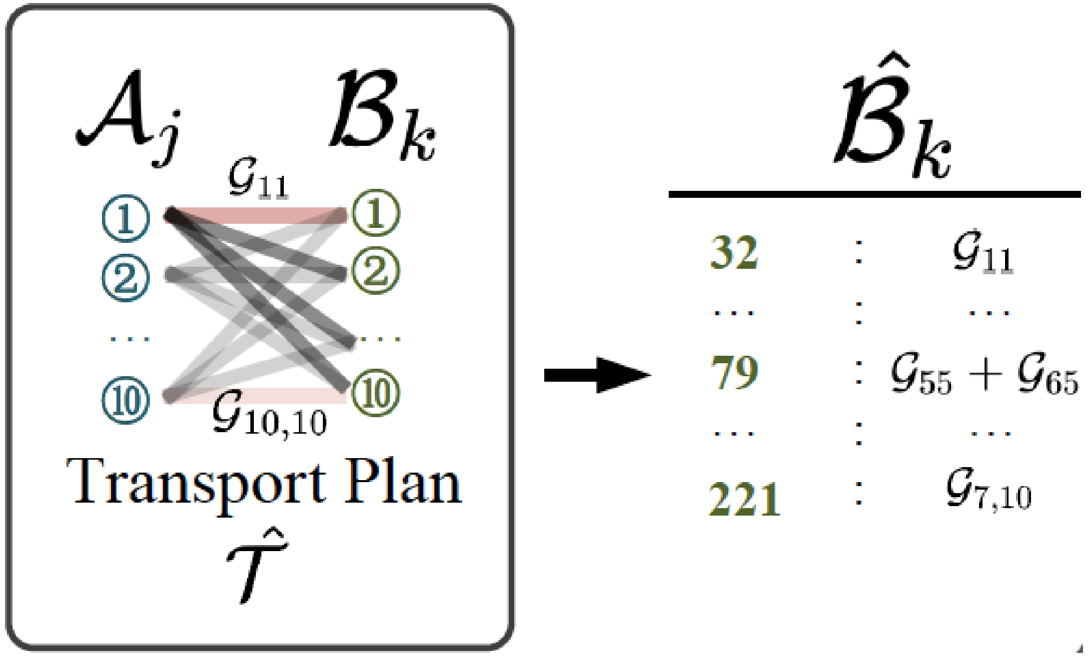
$$\hat{\mathcal{T}} = \arg\min_{\mathcal{T} \geq 0} \left\{ \sum_{x=1}^{n} \sum_{y=1}^{m} c_{xy}\, \mathcal{G}_{xy} \;\middle|\; \sum_{y=1}^{m} \mathcal{G}_{xy} = \mathcal{A}_j[x] \forall x, \quad \sum_{x=1}^{n} \mathcal{G}_{xy} = \mathcal{B}_k[y] \forall y \right\}, n = m = 10,$$

Per logit minimum edit distance cost

3. New Alternative: Probabilistic Token Alignment

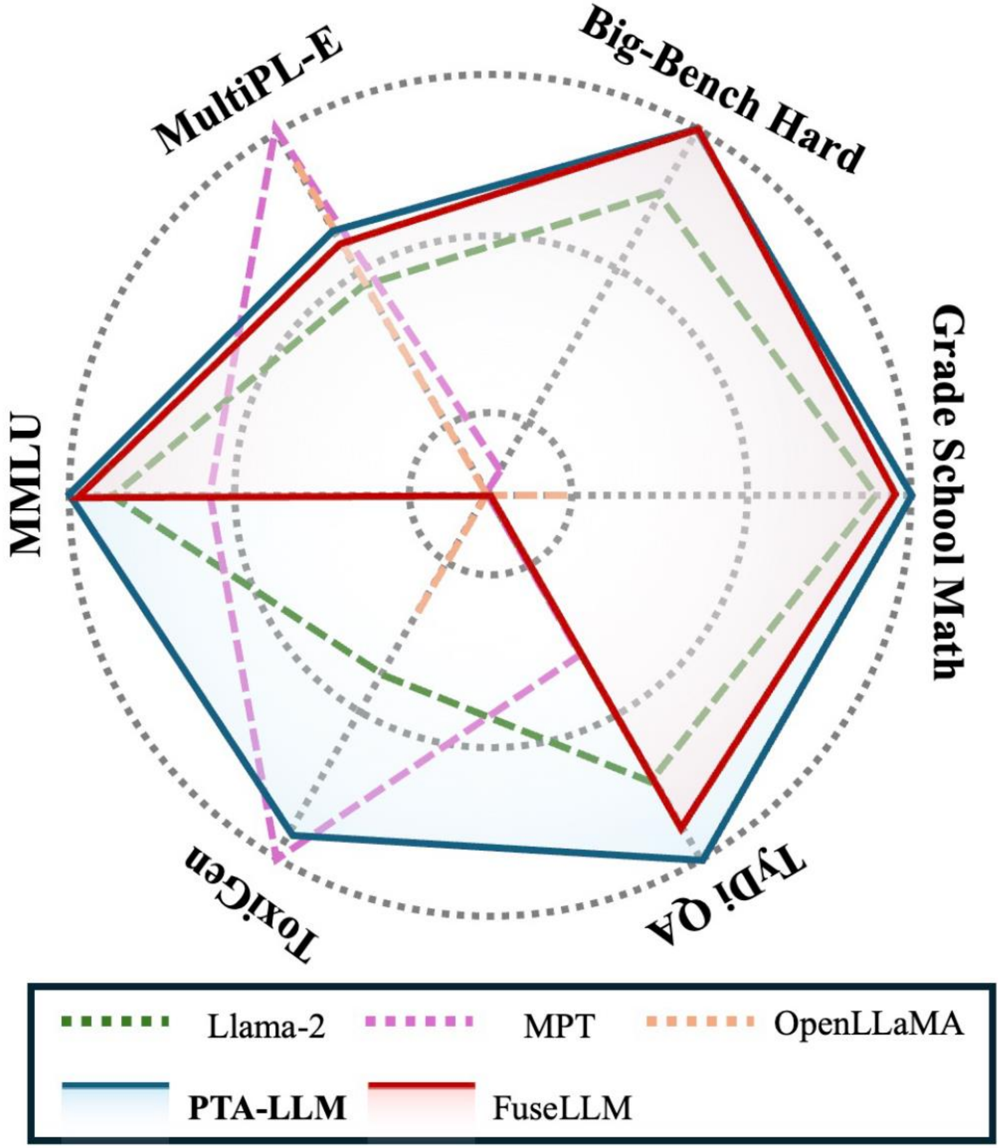
① Dynamic Token Pairing

② Probabilistic Token Alignment



Empirical Evidence: Does PTA work?

4. Empirical Evidence: Generalization



The **global probabilistic distribution transport** enhances the coherence of the representations, thereby improving the model's ability to **generalize across a wide range of tasks**.

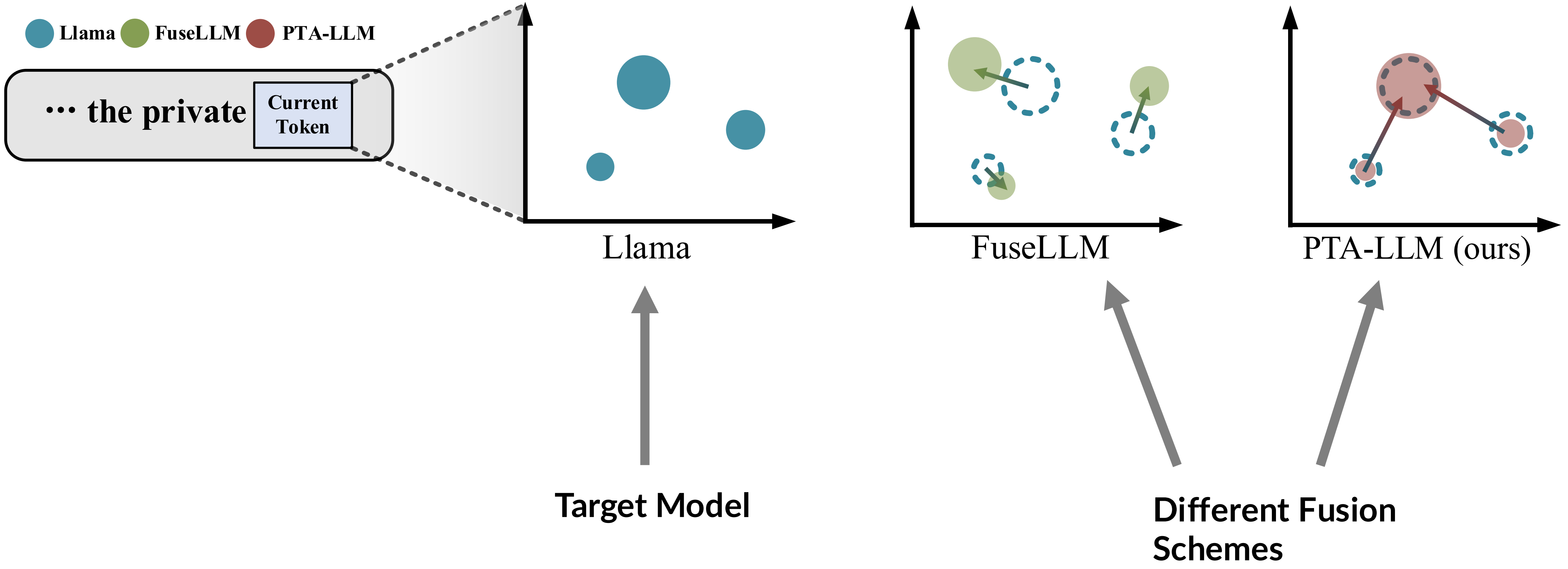
4. Empirical Evidence: Robustness

Table 7: Case studies on the Tracking Shuffled Objects (7 objects) task.

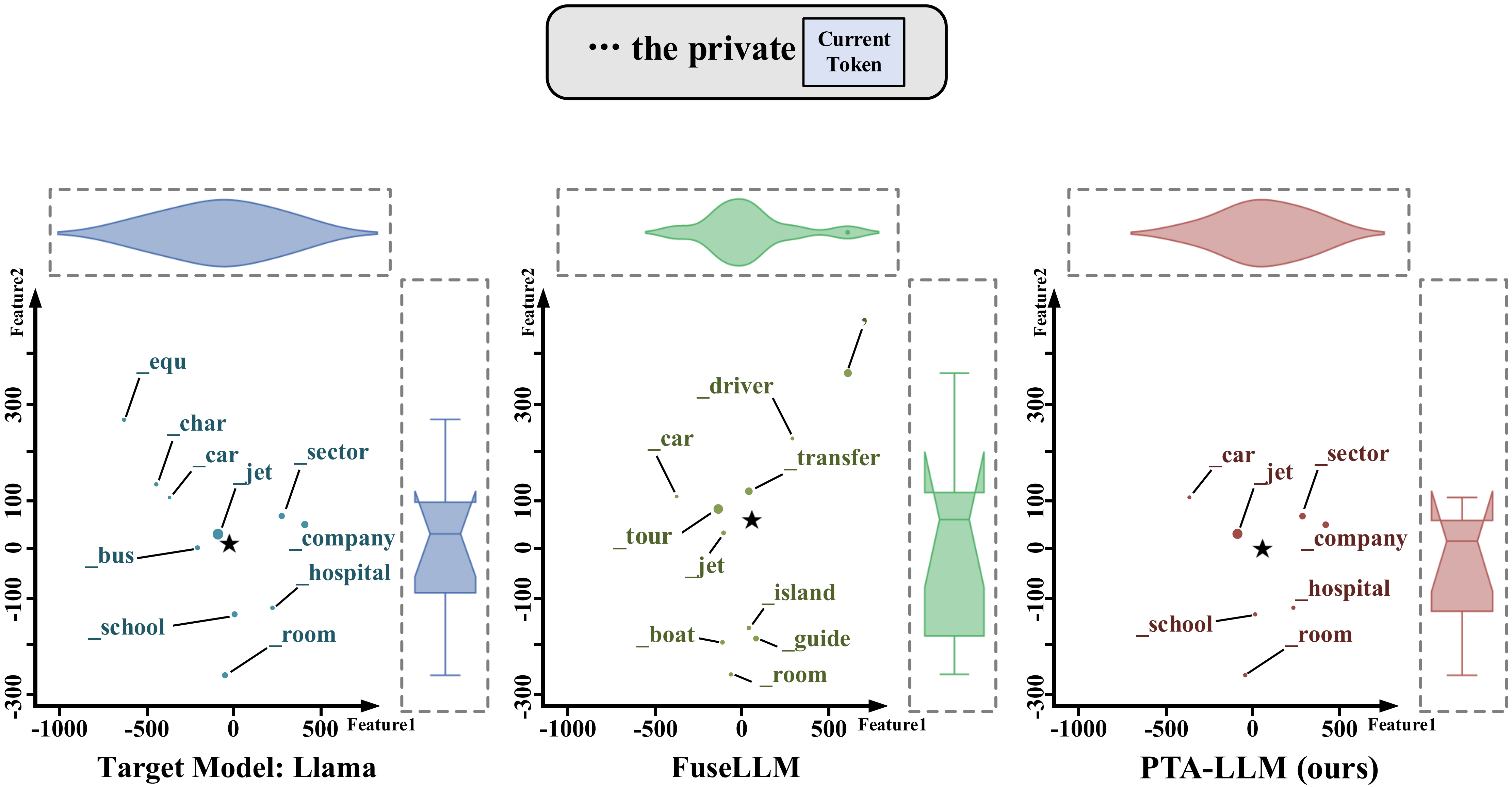
BBH; Tracking Shuffled Objects (7 objects);
Question: Alice, Bob, Claire, Dave, Eve, Fred, and Gertrude are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Lola, Bob is dancing with Rodrigo, Claire is dancing with Jamie, Dave is dancing with Izzi, Eve is dancing with Sam, Fred is dancing with Karl, and Gertrude is dancing with Helga. Throughout the song, the dancers often trade partners. First, Alice and Claire switch partners. Then, Bob and Fred switch partners. Then, Claire and Dave switch partners. Then, Alice and Fred switch partners. Then, Claire and Alice switch partners. Then, Bob and Eve switch partners. Finally, Eve and Gertrude switch partners. At the end of the dance, Eve is dancing with Options: (A) Lola (B) Rodrigo (C) Jamie (D) Izzi (E) Sam (F) Karl (G) Helga
Target: (G)
FUSELLM: Let's think step by step. (0) At the start: Alice: Lola, Bob: Rodrigo, Claire: Jamie, Dave: Izzi, Eve: Sam, Fred: Karl, Gertrude: Helga. (1) Alice and Claire switch partners: Alice: Rodrigo, Claire: Lola, Dave: Izzi, Eve: Sam, Fred: Karl, Gertrude: Helga. (2) Bob and Fred switch partners: Alice: Rodrigo, Claire: Lola, Dave: Izzi, Eve: Sam, Fred: Karl, Gertrude: Helga. (3) Claire and Dave switch partners: Alice: Rodrigo, Claire: Lola, Bob: Izzi, Eve: Sam, Fred: Karl, Gertrude: Helga. (4) Alice and Fred switch partners: Alice: Izzi, Claire: Lola, Bob: Rodrigo, Dave: Sam, Eve: Karl, Gertrude: Helga. (5) Claire and Alice switch partners: Alice: Izzi, Claire: Lola, Bob: Rodrigo, Dave: Sam, Eve: Karl, Gertrude: Helga. (6) Bob and Eve switch partners: Alice: Izzi, Claire: Lola, Bob: Rodrigo, Dave: Sam, Eve: Karl, Gertrude: Helga. (7) Eve and Gertrude switch partners: Alice: Izzi, Claire: Lola, Bob: Rodrigo, Dave: Sam, Eve: Karl, Gertrude: Helga. At the end of the dance, Eve is dancing with Karl. So the answer is (F).
PTA-LLM: Let's think step by step. (0) At the start: Alice: Lola, Bob: Rodrigo, Claire: Jamie, Dave: Izzi, Eve: Sam, Fred: Karl, Gertrude: Helga. (1) Alice and Claire switch partners: Alice: Jamie, Claire: Lola, Dave: Izzi, Eve: Sam, Fred: Karl, Gertrude: Helga. (2) Bob and Fred switch partners: Alice: Izzi, Claire: Lola, Bob: Rodrigo, Dave: Jamie, Eve: Sam, Fred: Karl, Gertrude: Helga. (3) Claire and Dave switch partners: Alice: Izzi, Claire: Rodrigo, Bob: Jamie, Dave: Lola, Eve: Sam, Fred: Karl, Gertrude: Helga. (4) Alice and Fred switch partners: Alice: Sam, Claire: Rodrigo, Bob: Izzi, Dave: Lola, Eve: Jamie, Fred: Karl, Gertrude: Helga. (5) Claire and Alice switch partners: Alice: Sam, Claire: Izzi, Bob: Rodrigo, Dave: Lola, Eve: Jamie, Fred: Karl, Gertrude: Helga. (6) Bob and Eve switch partners: Alice: Sam, Claire: Izzi, Bob: Rodrigo, Dave: Lola, Eve: Jamie, Fred: Karl, Gertrude: Helga. (7) Eve and Gertrude switch partners: Alice: Sam, Claire: Izzi, Bob: Rodrigo, Dave: Lola, Eve: Jamie, Fred: Karl, Gertrude: Helga. At the end of the dance, Eve is dancing with Helga. So the answer is (G).

Wrong tracking!

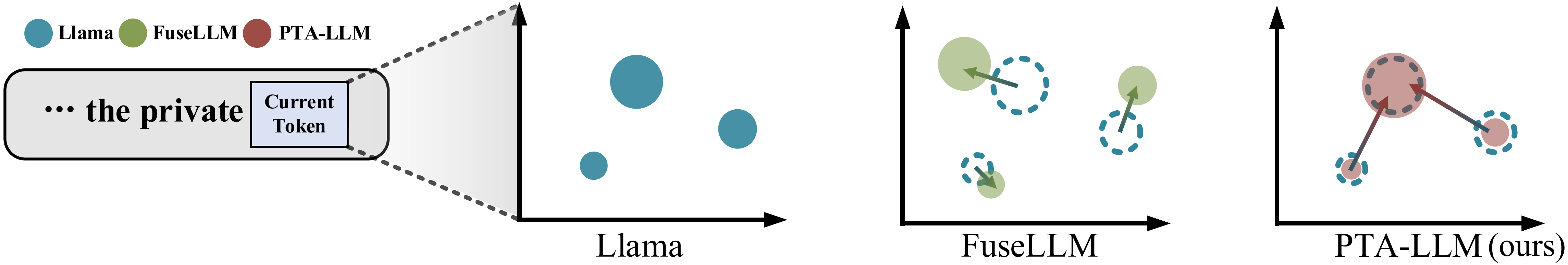
4. Empirical Evidence: Interpretability



4. Empirical Evidence: Interpretability



4. Empirical Evidence: Interpretability



PCA Variance Ratio: 95.60%

	Compactness↓	Similarity↓
Llama	283.24	0.00
FuseLLM	257.83	136.95
PTA-LLM	239.44	22.25

Takeaway: Insight for future works

5. Takeaway

- **Knowledge Fusion**

  ⭐The potential paradigm of building a stronger baseline model.

  💭The challenge of homogenization?

  💭End-to-end fusion pipeline?

- **Probabilistic Token Alignment**

  ⭐Coherently fusing each logit benefits the model learning.

  💭N-1 and 1-N mapping?

  💭The potential extension to multilingual applications?

THANK YOU!