Uncertainty-Aware and Data-Efficient Fine-Tuning and Application of Foundation Models

Yinghao Li, ML @ Georgia Tech; April 18, 2025

Committee:

Dr. Chao Zhang (advisor); Dr. Rampi Ramprasad (co-advisor);
Dr. Tuo Zhao; Dr. Srijan Kumar; Dr. Victor Fung; Dr. Ali Torkamani;

Georgia Tech

Background

- Pre-trained foundation models show impressive zero- or few-shot ability

Please identify the LOCATION entities in
"A Trump tower is located on the 5th avenue in New York".

LOCATION entities:
- "Trump tower";
- "5th avenue";
- "New York"

Background

- For niche domains, such as materials science
  - Training data are sparse --> foundation models fail to learn enough/precise knowledge

Please identify the MATERIAL PROPERTIES in "The **domain sizes** estimated by crosssection profiles are about 10-20 nm".

There is no property mentioned

Incorrect!

Address

Uncertainty Quantification

There is no property mentioned

Model's <u>confidence</u> to its answer:

0.01

Decision:

<u>Ignore</u>

Fine-Tuning

In-domain data

MATERIAL PROPERTIES entities:

- **domain sizes**

Challenge: Discriminative Uncertainty Quantification

- Discriminative : output space is a low-dimensional categorical/Gaussian distribution
  - Text classification, material property prediction, etc.

- Larger pre-trained foundation models are more prone to overfit

- Numerous UQ methods exist, each with different characteristics

Which/How to select?

Georgia Tech

Challenge: LM Uncertainty Quantification

- Language model (LM): output response is a sequence of interdependent tokens.
- We are not equally interested in (the confidence of) every token in the response

Answer

`Q : 1 + 1 = ? A : A wise man once said “ 1 + 1 = 2 ”. The answer is \boxed{ 2 } .`

Instruction

Not Interested

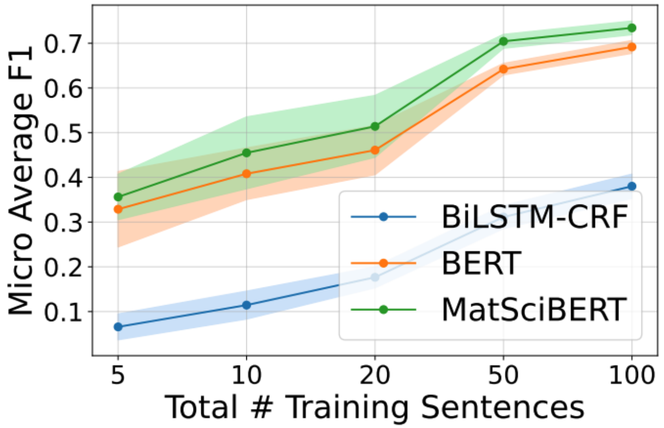
Interested

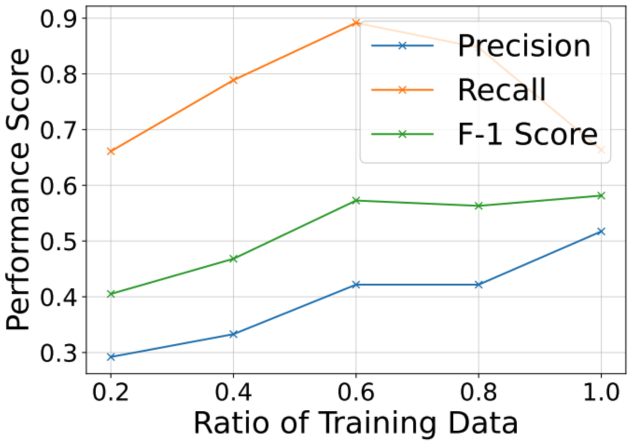
- How to get the marginal probability of the answer tokens we are interested in?

$$P(x_{\text{answer}} | x_{\text{instruction}})$$

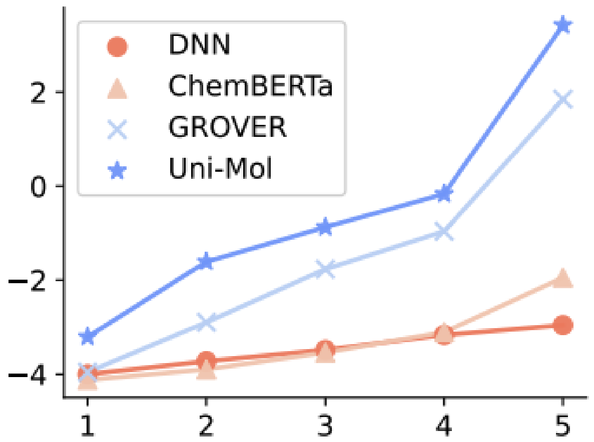
Challenge: Impact of Label Quantify & Quality

- Model optimization requires large, in-domain labeled data





Model performance is impacted by training data size. Dataset: PolyIE. From Cheung et al. (2023)

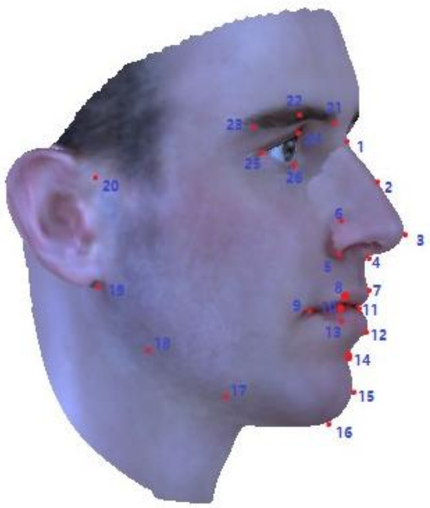


(c) NLL ↓

Model performance is impacted by training-test distribution shift. X-axis represents the difference between training and test distribution; larger number indicates greater distribution gap. From Li et al. (2024)

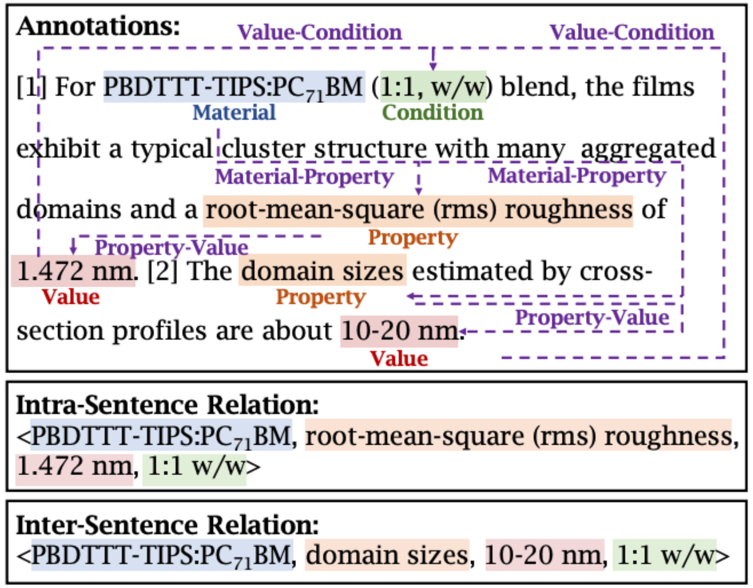
Challenge: Data Collection is Hard

Tedious and repetitive



26 marks × 1,000 images or more

Side face landmark annotation.
From 3DMM-fitting GitHub repo.



PolyIE annotation example. From Cheung et al. (2023).

Requires domain expertise

- **Other Issues**
  - Costly if crowd-sourced, potentially low-quality
  - Extended time-period
  - …

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

04 ELREA

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

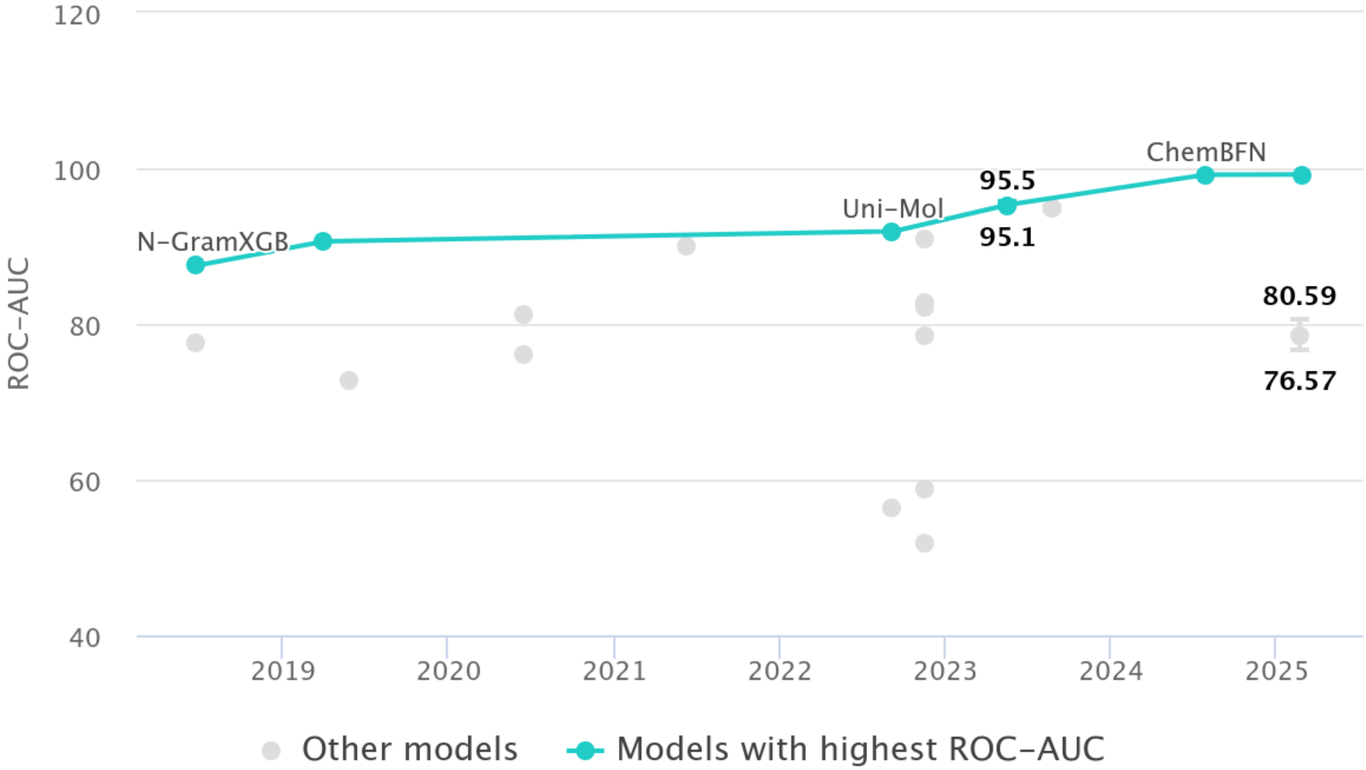
04 ELREA

In *TMLR*: https://openreview.net/forum?id=qYceFeHgm4

Molecular Representation Models

- Pre-trained large molecular representation models achieve SOTA performance on a variety of property prediction tasks through fine-tuning.

QM9

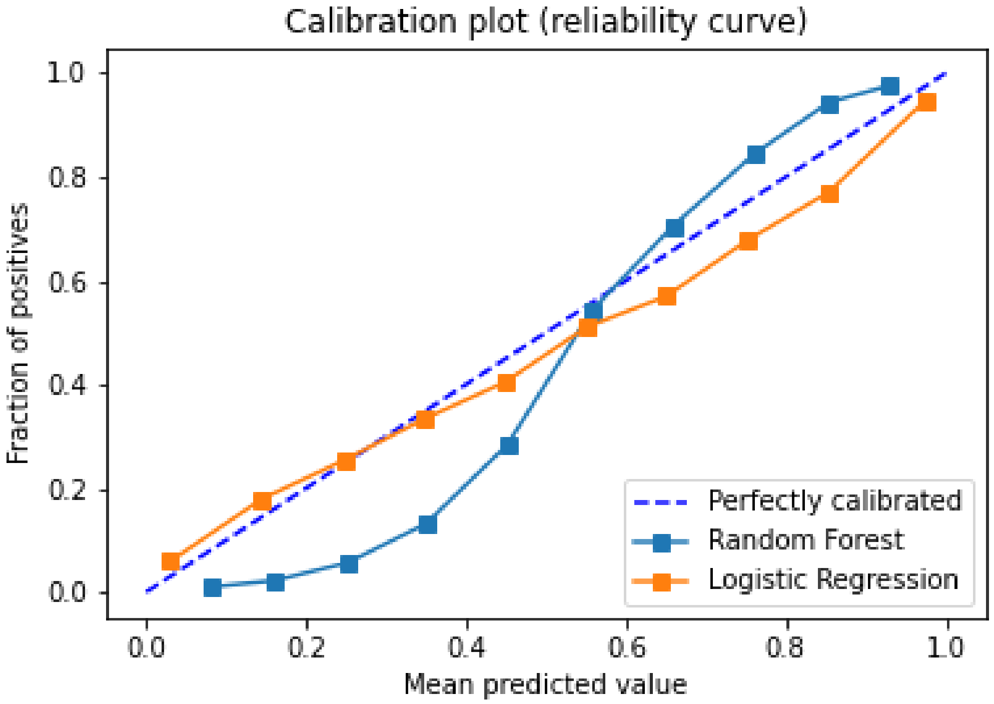


ClinTox

Georgia Tech

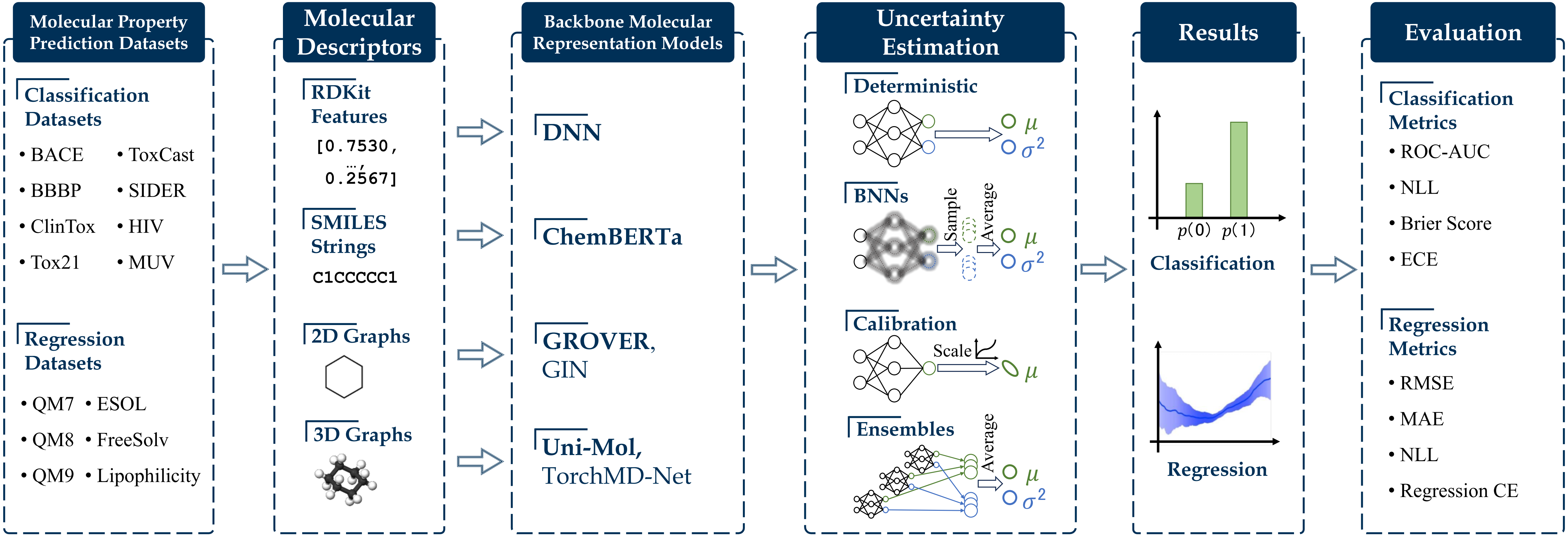
Uncertainty-Aware Property Prediction

- It is desirable for predictions to be not only precise, but also *well-calibrated*
- Distinguish noisy predictions and improve model robustness.
- Applications: active learning; high throughput screening; wet-lab experimental design.



Calibration Plot; from Medium post

MUBen Components



Models and UQ Methods

Model	# Parameters (M)	Average Time per Training Step (ms)[a]
DNN	0.158	5.39
ChemBERTa	3.43	30.18
GROVER	48.71	334.47
Uni-Mol	47.59	392.55
TorchMD-NET	7.23	217.29
GIN	0.26	7.21

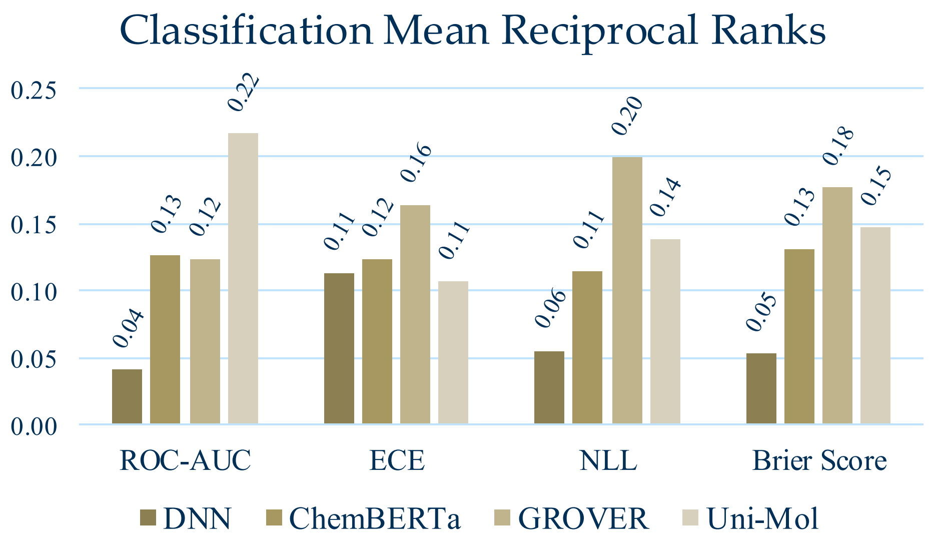
Molecular Representation Models

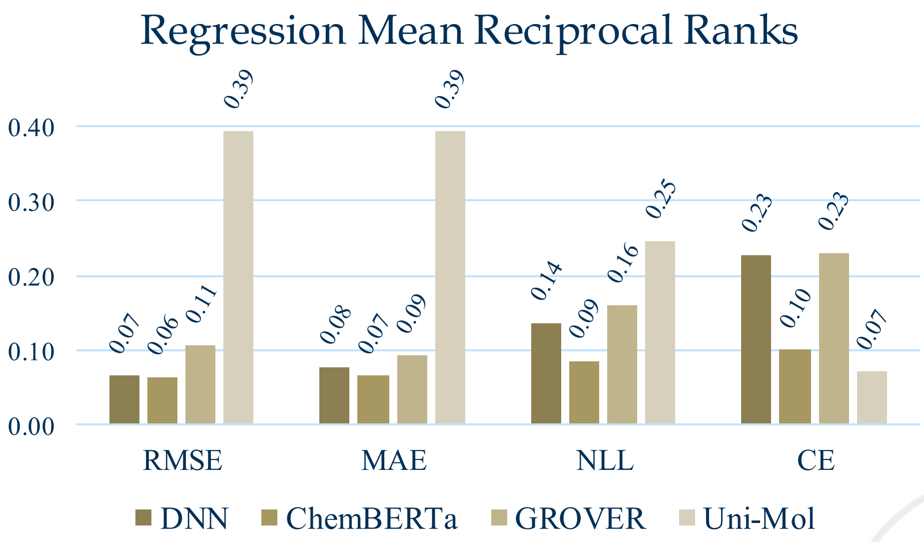
UQ Method	Training Starting Checkpoint	Additional Cost[a]
Deterministic	-	0
Temperature	from fine-tuned backbone	$(T_{\text{infer}} + T_{\text{train-FFN}}) \times M_{\text{train-extra}}$
Focal Loss	from scratch	$T_{\text{train}} \times M_{\text{train}}$
MC Dropout	no training	$T_{\text{infer}} \times M_{\text{infer}}$
SWAG	from fine-tuned backbone	$T_{\text{train}} \times M_{\text{train-extra}} + T_{\text{infer}} \times M_{\text{infer}}$
BBP	from scratch	$T_{\text{train}} \times M_{\text{train}} + T_{\text{infer}} \times M_{\text{infer}}$
SGLD	from scratch	$T_{\text{train}} \times (M_{\text{train}} + M_{\text{train-extra}}) + T_{\text{infer}} \times M_{\text{infer}}$
Ensembles	from scratch	$T_{\text{train}} \times M_{\text{train}} \times (N_{\text{ensembles}} - 1)$

Uncertainty Quantification Methods

Comparison of Backbone Models

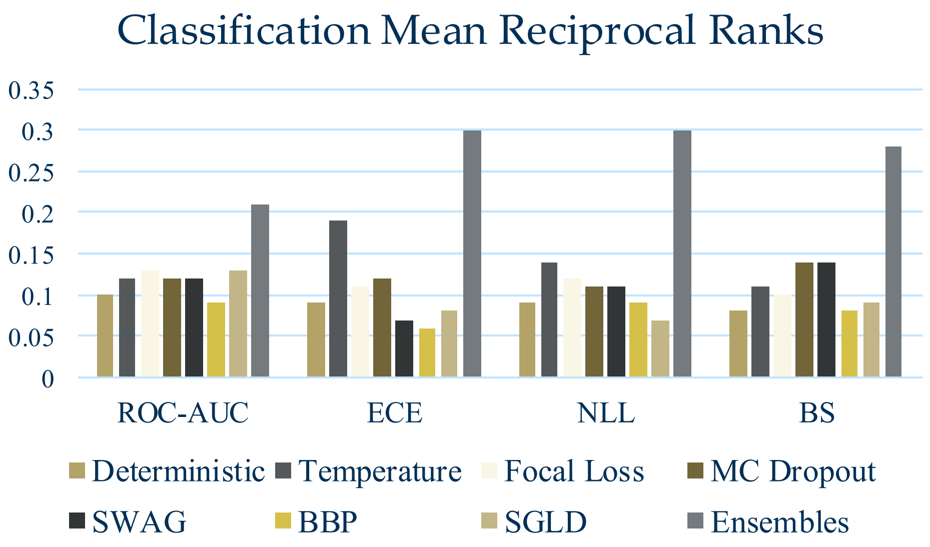
- Uni-Mol performs the best for property prediction (ROC-AUC, RMSE and MAE), but tend to be over-confident, yielding sub-optimal calibration (ECE and CE).
- GROVER is a safer choice when both prediction and UQ performance are required.
- Pre-trained models do not invariably surpass heuristic features, as shown in the comparison between DNN & ChemBERTa for regression.

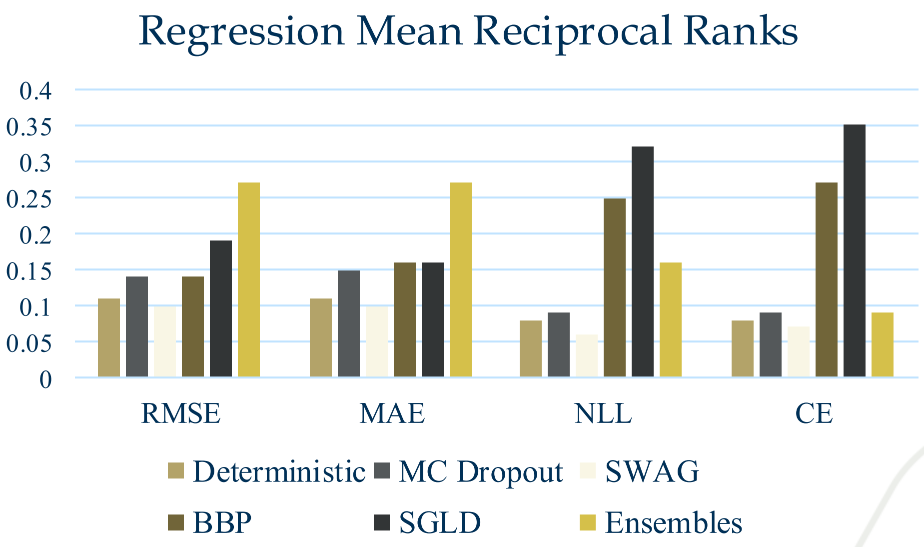




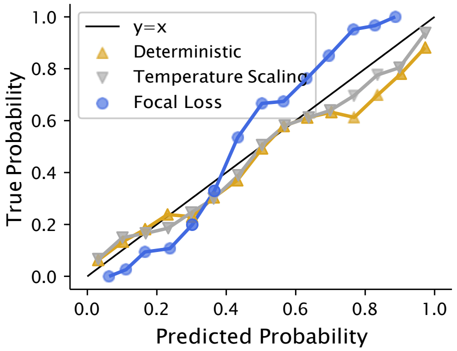
Comparison of UQ Methods

- Most UQ methods enhance both value prediction and uncertainty estimation.
- BBP and SGLD fail on classification but deliver the greatest improvement on regression.
- Deep Ensembles guarantees to improve the prediction and UQ results, but at a cost of heavy computational consumption.
- MC Dropout is cheap to adopt and theoretically does not risk model performance under any circumstances, making it a first-pick when computation resource is limited.
- Temperature Scaling is also cheap for classification calibration, but it may fail when the held-out calibration dataset has a large distribution shift from the test set.

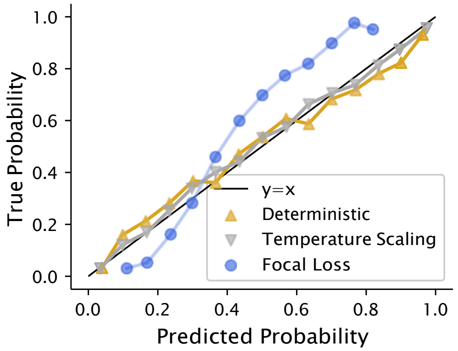




Case Studies

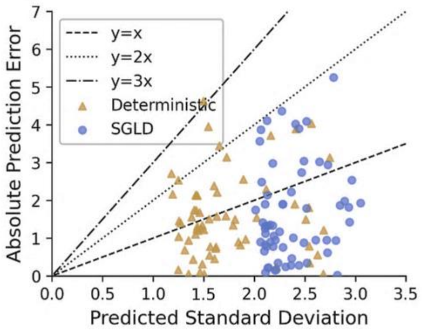


DNN on SIDER

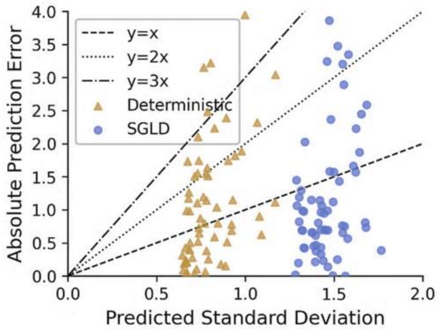


Uni-Mol on SIDER

Calibration plots for classification tasks.



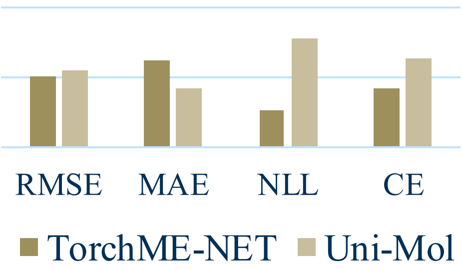
DNN on FreeSolv



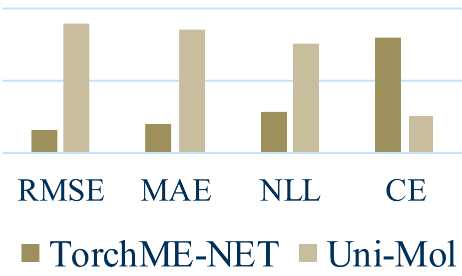
Uni-Mol on FreeSolv

Absolute error v.s. predicted std on regression tasks.

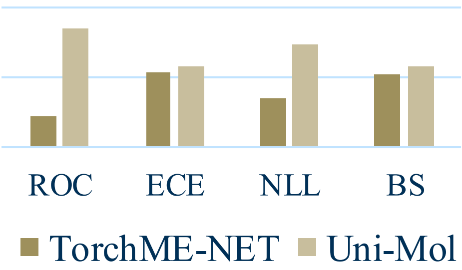
Quantum Mechanics



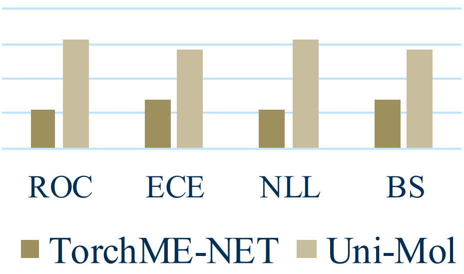
Physical Chemistry



Biophysics



Physiology



The Mean Reciprocal Ranks (larger is better) of TorchMD-NET and Uni-Mol on datasets with different features. TorchMD-NET is mainly pre-trained for predicting QM properties.

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

04 ELREA

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

04 ELREA

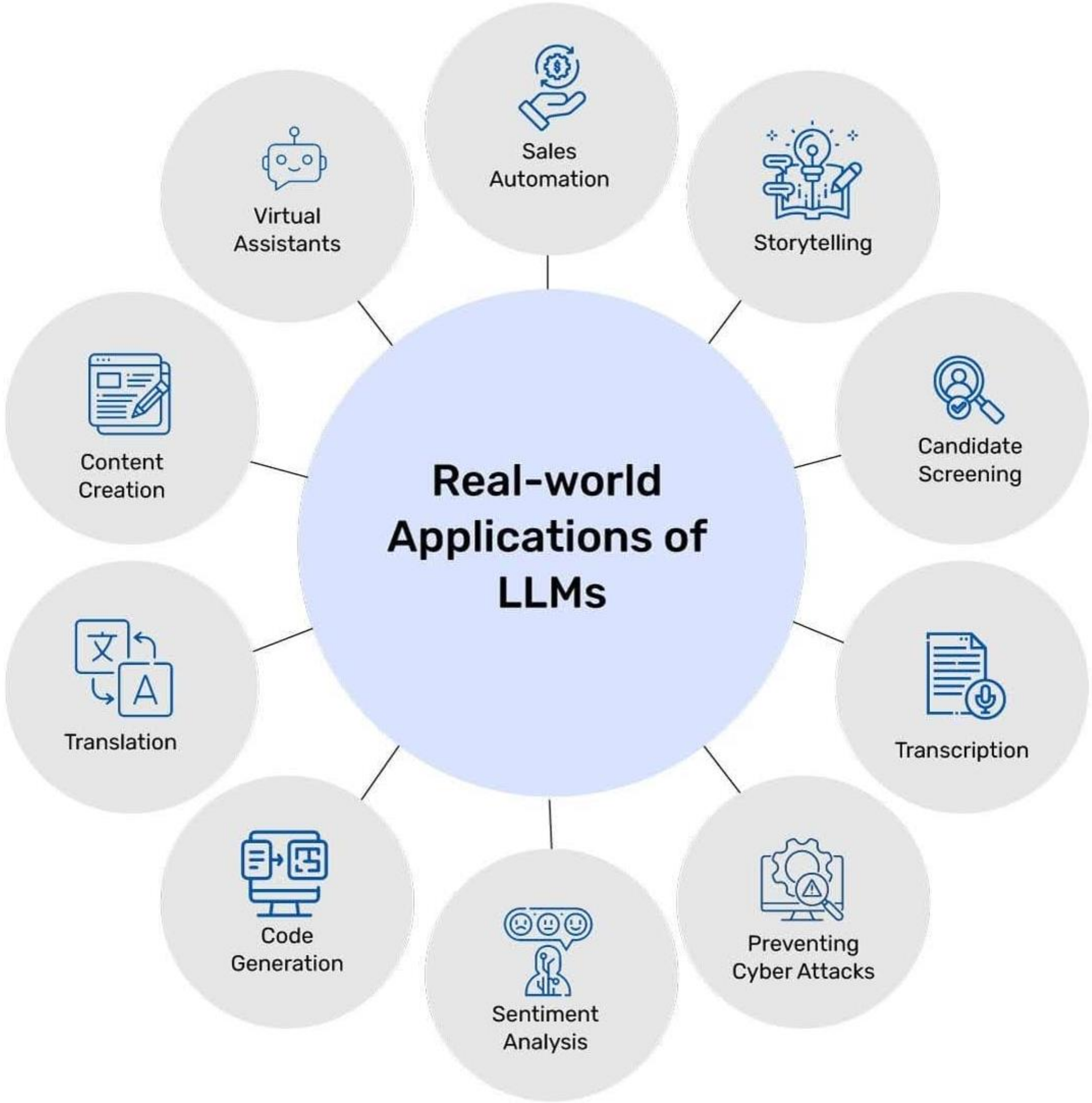
In *Submission of COLM 2025*: https://arxiv.org/abs/2503.19168

Large Language Models

- Many applications; used everyday

- Tend to hallucinate
- Black-box

- How to know answer is accurate?



Source: https://www.mindbowser.com/llm-application-development/

Language Model Uncertainty Quantification

$1 + 1 = ?$

3

$$P(3|\cdot) = 41\%$$

$$P(2|\cdot) = 38\%$$

$$P(1|\cdot) = 20\%$$

...

Impact of the Reasoning Sequence

$1 + 1 = ?$

A wise man once said: "$1 + 1 = 3$ when it is not calculated correctly". So, the answer is 3.

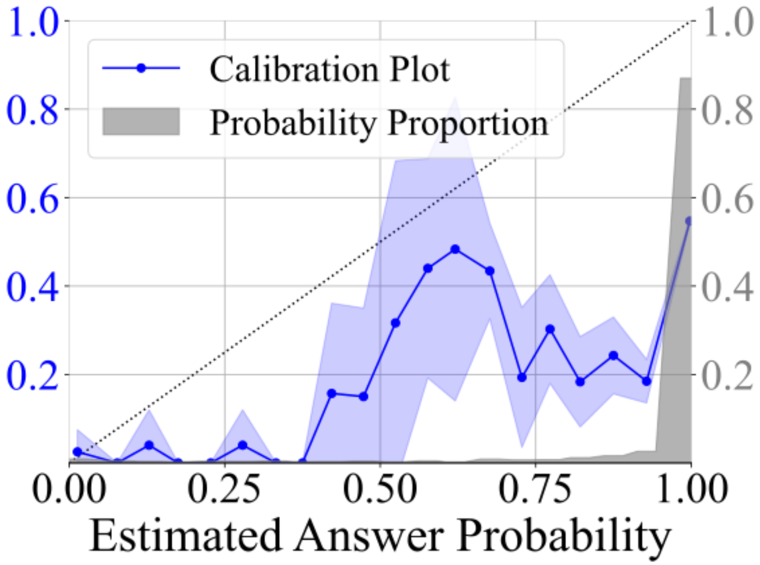
$P(3|\text{reasoning sequence}, \ \cdot) = \mathbf{99.9}\ \%$

$P(2|\text{reasoning sequence}, \ \cdot) = 0.09\ \%$

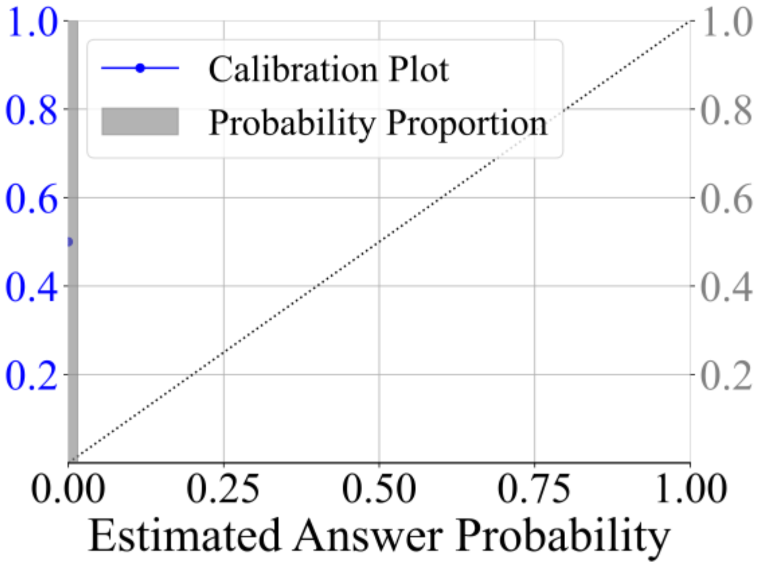
$P(4|\text{reasoning sequence}, \ \cdot) = 10^{-6}\ \%$

…

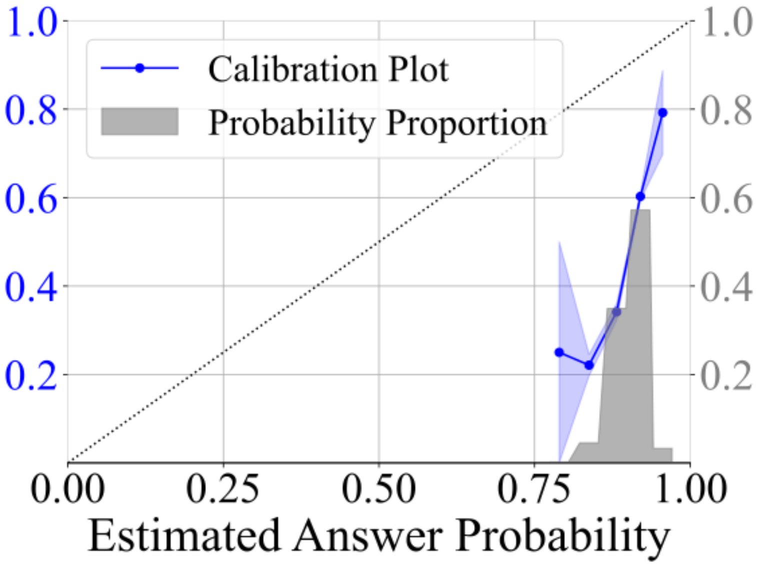
Probability Aggregation Methods?



Joint probability of answer tokens



Joint probability of all response tokens



Average conditional probability of all response tokens

Georgia Tech

The Proper Way

$$P(x_{\text{ans}}|x_{\text{instr}}) = \sum_{x_{\text{cot}}} P(x_{\text{ans}}|x_{\text{cot}}, x_{\text{instr}})P(x_{\text{cot}}|x_{\text{instr}})$$

- x_{ans}: Answer tokens; x_{cot}: reasoning tokens; x_{instr}: instruction tokens

- But $\sum_{x_{\text{cot}}} \cdot$ is intractable

```
A wise man once said …
 The     woman
There    elder
 Wise    monke
  …        …
```

The Observation

- Not all tokens in the reasoning sequence contribute equally to the final answer

```
Q : 1 + 1 = ?
A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
```

↓

```
Q : 1 + 1 = ?
A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
```

Attention Backtracking

```
Q : 1 + 1 = ?
```

```
A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
```

Attention Backtracking

`Q : 1 + 1 = ?`

Target tokens

Answer token(s)

`A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .`

0.7

0.2

Attention Backtracking

Q : 1 + 1 = ?

0.48

0.4

A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .

Source tokens

Attention Backtracking

`Q : 1 + 1 = ?`

0.8

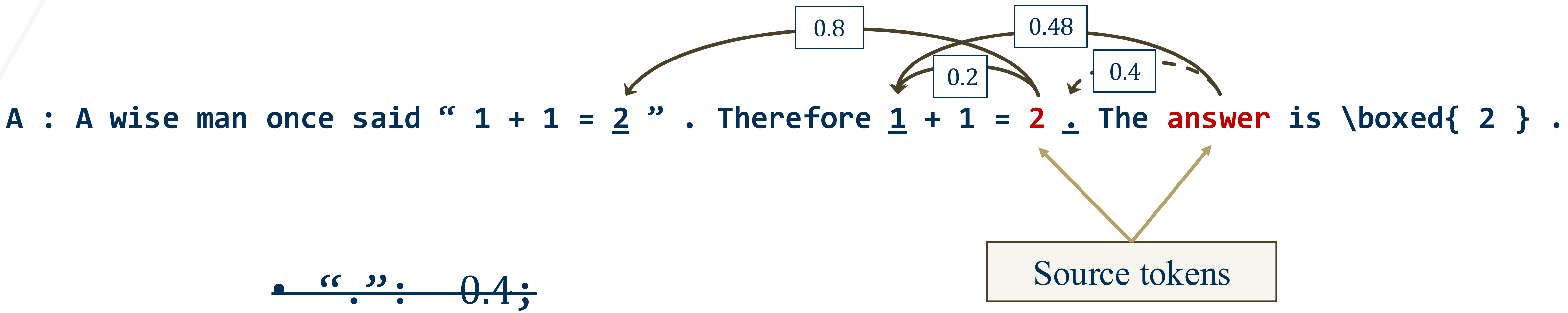
0.2

`A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .`

Source tokens

Attention Backtracking

```
Q : 1 + 1 = ?
```



- ~~“.”: 0.4;~~
- “1”: $0.48 + 0.2 = 0.68$;
- “2”: 0.8;

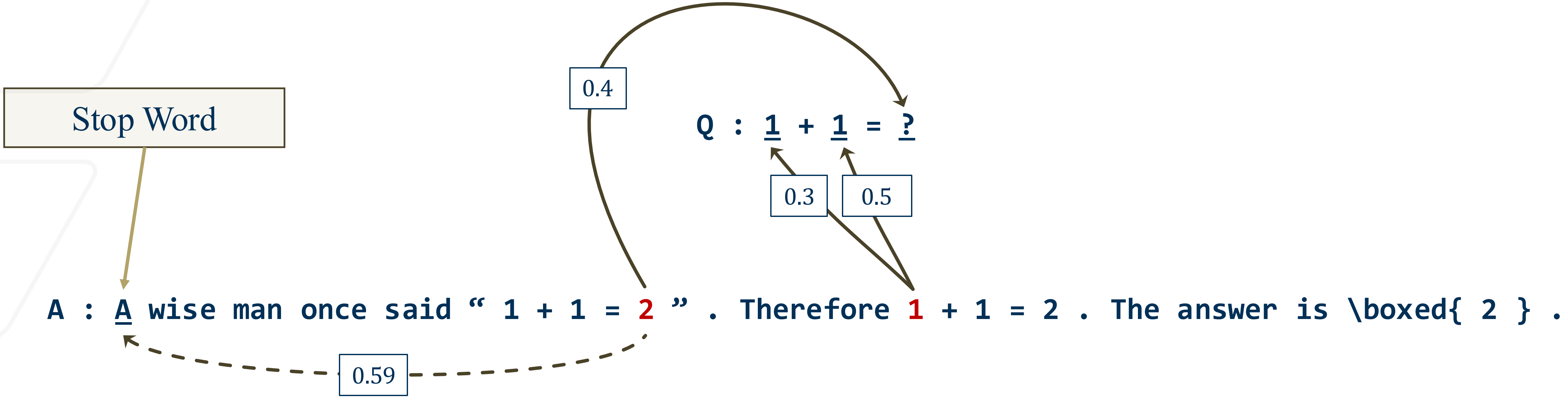
Attention Backtracking

Q : 1 + 1 = ?

0.3 0.5

A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .

Attention Backtracking



Attention Chain

```
Q : 1 + 1 = ?
```

```
A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
```

Attention Chain: [“2”, “1”, “2”, “answer”]

Georgia Tech

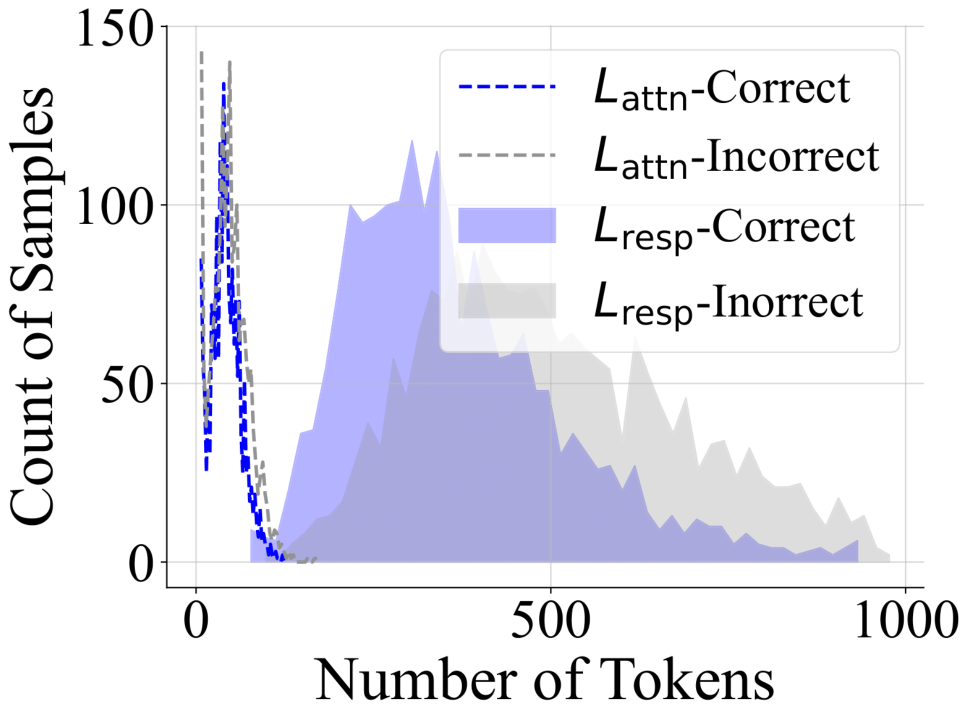
Similarity Filtering

- Attention chain is much shorter than the response sequence ($\sim$10%)
  - May still bee too long for marginal calculation
  - Not control over length

- Similarity filtering

… “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 .

The answer is \boxed{ 2 } .



Attention Chain (filtered): [“2”, “1”, “2”]

Probability Thresholding

- Reasoning sequence is shorter, but how about the vocabulary space?

```
A wise man once said …
      woman
      elder
      monke

        …
```

Complexity: $|\mathcal{V}|^{\text{number of tokens}}$

Probability Thresholding

- Reasoning sequence is shorter, but how about the vocabulary space?

```
A wise man once said …
      woman
      elder
      monke
        …
```

```
A wise man once said …
      Woman: 0.3
      Elder: 0.02
```
~~Monke: 0.002~~
~~Other: < 0.01~~

- Keep only candidate tokens with conditional probability higher than 0.01

Reasoning Space

- Only substitute one token at a time. Do not consider candidate token combinations.
  - i.e., Hamming distance of the original response sequence and other sequences in the space is always 1.

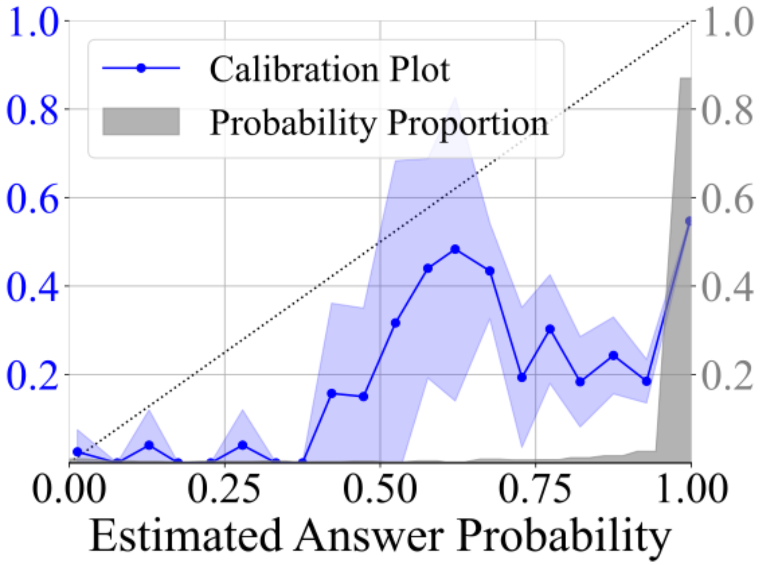
- reasoning space can be generally reduced to a size of 6 – 7.

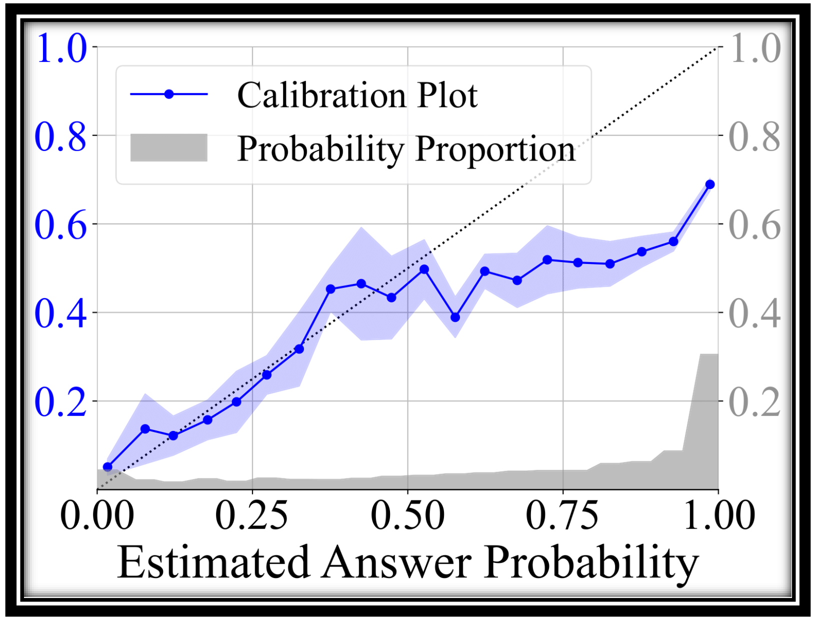
```
A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
A : A wise man once said “ 1 + 1 = 3 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
A : A wise man once said “ 1 + 1 = 2 ” . Therefore the + 1 = 2 . The answer is \boxed{ 2 } .
A : A wise man once said “ 1 + 1 = 1 ” . Therefore 1 + 1 = 2 . The answer is \boxed{ 2 } .
A : A wise man once said “ 1 + 1 = 2 ” . Therefore 1 + 1 = 1 . The answer is \boxed{ 2 } .
```

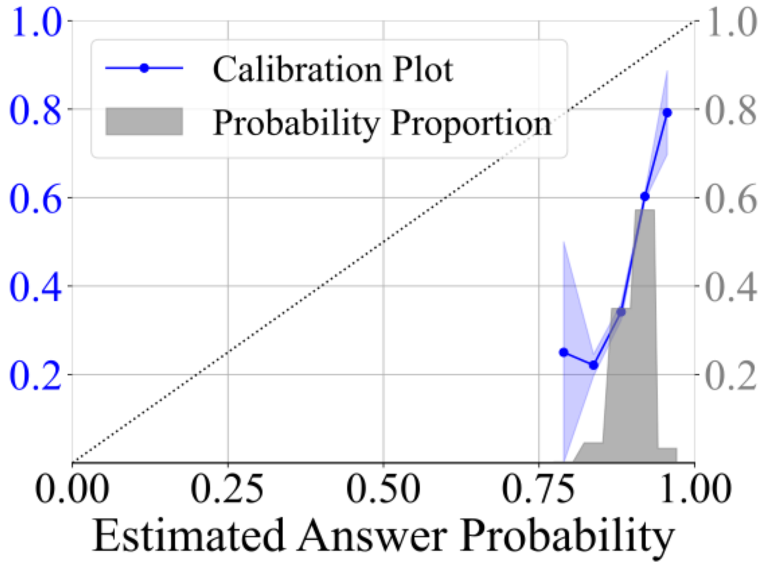

Georgia Tech

Calibration Plots

UQAC







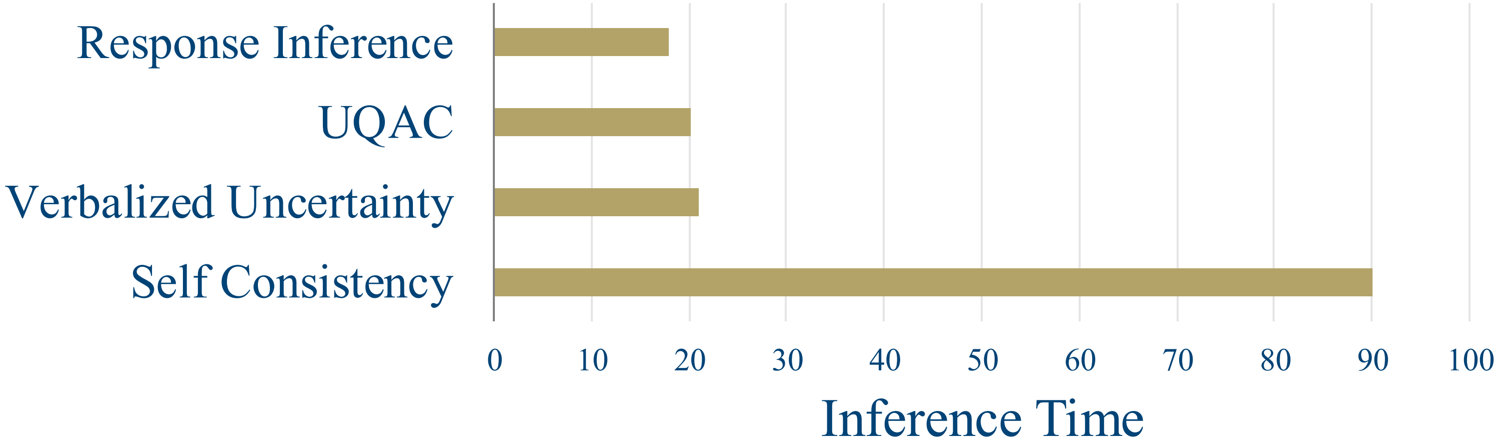
Joint probability of answer tokens

Average conditional probability of all response tokens

Llama-3.1-8B-Instruct on MATH dataset; positive and negatives are balanced

Comparison with Other Methods

	GSM8k		MATH		BBH	
	AUROC↑	ECE↓	AUROC↑	ECE↓	AUROC↑	ECE↓
Self-Consistency	66.4±1.9	28.9±0.8	79.5±1.0	15.8±0.8	79.5±1.0	31.6±0.7
Verbalized Uncertainty	54.9±0.5	42.9±0.2	57.4±0.7	45.1±0.2	58.2±1.2	39.7±0.3
UQAC	61.3±0.9	33.6±0.4	69.5±1.2	25.8±0.9	66.7±1.2	24.2±0.9



UQAC Characteristics

- Efficient
  - Attention backtracking needs attention scores and last layer embeddings, which are already calculated for inference.
  - Do not rely on external models.
  - No recurrent generation; marginalization can be computed in parallel.
- Applicable
  - Working on any Transformer-based white-box autoregressive LLMs.
- Calibrated
  - Marginal probability ranging from 0 – 1;

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

04 ELREA

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

04 ELREA

CHMM in *ACL 2021*: https://aclanthology.org/2021.acl-long.482/
Wrench in *NeurIPS 2021 Benchmark*: https://openreview.net/forum?id=Q9SKS5k8io
Sparse CHMM in *KDD 2022*: https://dl.acm.org/doi/10.1145/3534678.3539247
G&O in *ACL 2024 Findings (Short)*: https://aclanthology.org/2024.findings-acl.947/

Georgia Tech

Named Entity Recognition (NER)

- Subtask of information extraction, seeks to find pre-defined named entities in a sequence
  - Named entities: such as "person", "location", "organization", *etc*.

On the 15th of September DATE , Tim Cook PERSON announced that Apple ORG wants to acquire ABC Group ORG from New York GPE for 1 billion dollars MONEY

- NER is usually formulated as a token classification task
  - Assigns one label to each token in the sequence

Sentence:	On	the	15th	of	September	,	Tim	Cook	announced	that	Apple	wants	to	…
Labels:	O	B-DATE	I-DATE	I-DATE	I-DATE	O	B-PER	I-PER	O	O	B-ORG	O	O	…

Weakly Supervised Named Entity Recognition

Fully Supervised

	Rockefeller	Center	in	New	York	was...
Target	B-LOC	I-LOC	O	B-LOC	I-LOC	O...

	Rockefeller	Center	in	New	York	was...
Target	B-LOC	I-LOC	O	B-LOC	I-LOC	O...

Weakly Supervised

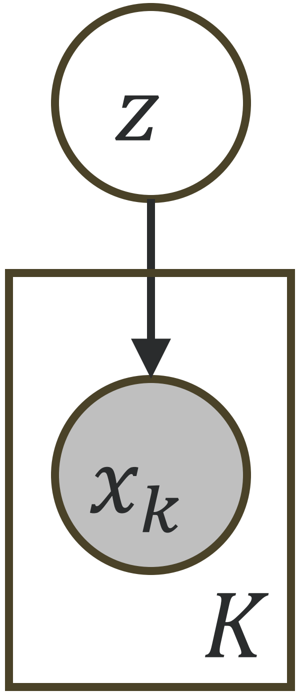
	Rockefeller	Center	in	New	York	was...
LF 1	B-PER	O	O	B-LOC	I-LOC	O...
LF 2	B-LOC	I-LOC	O	O	B-LOC	O...

	Rockefeller	Center	in	New	York	was...
Target	B-LOC	I-LOC	O	B-LOC	I-LOC	O...

A Principled Method with Graphical Models

Text Classification

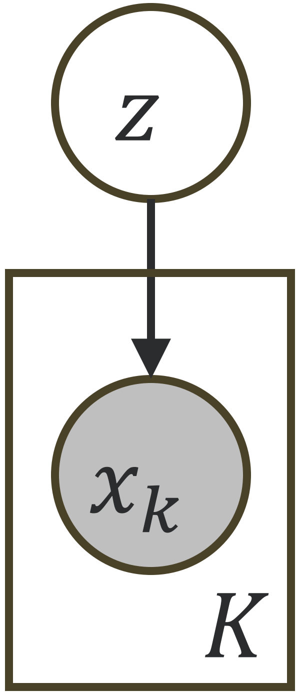
Naïve Bayes



A Principled Method with Graphical Models

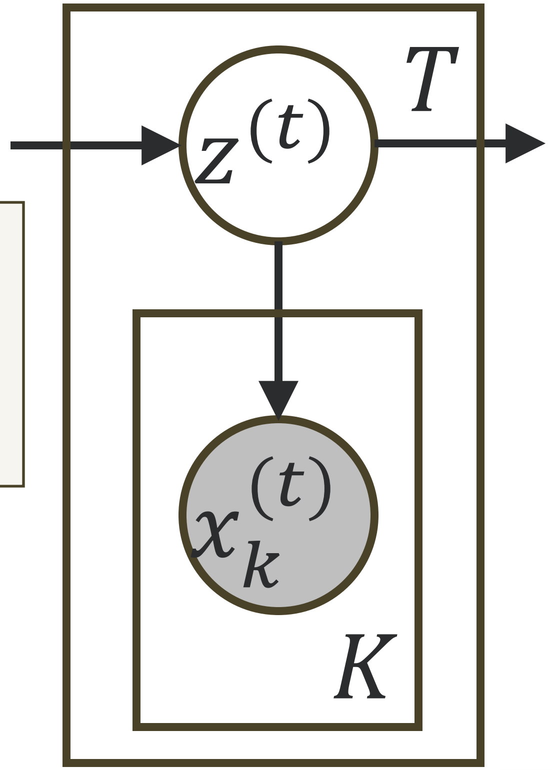
Text Classification

Naïve Bayes

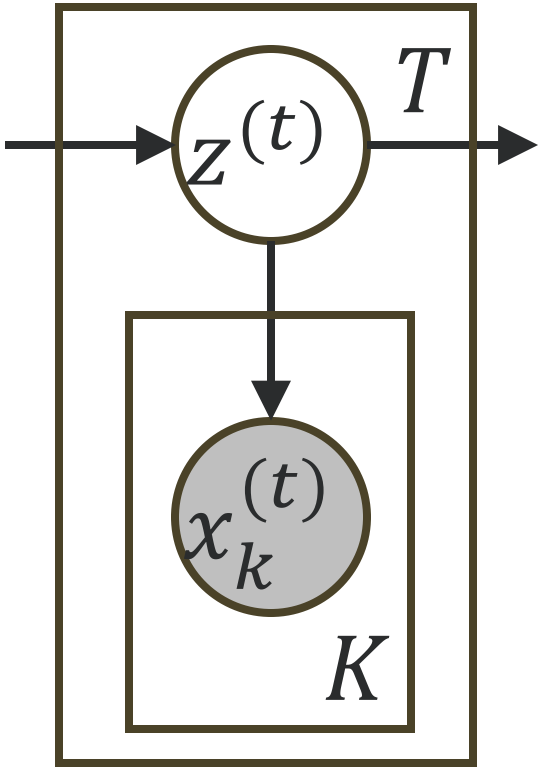


Named Entity Recognition

Hidden Markov Model (HMM)



HMM's Disadvantage



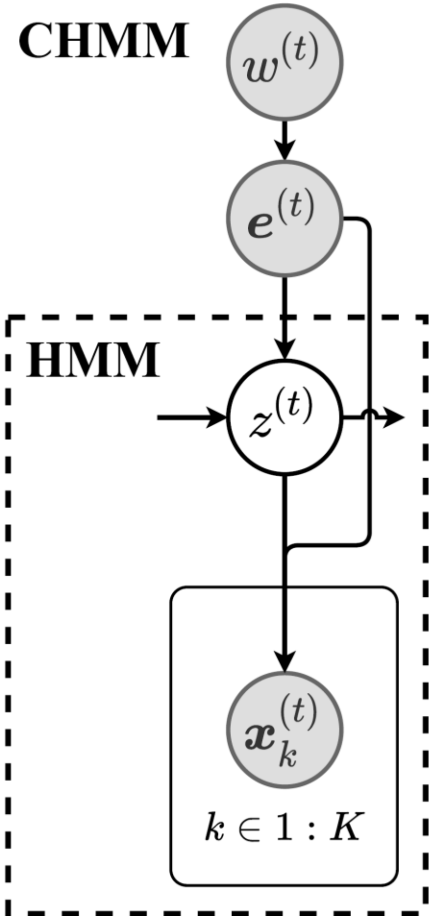
- The transitions and emissions remain constant for all time-steps
- Does not directly consider token information

- Fails to properly incorporate the sentence & token semantics

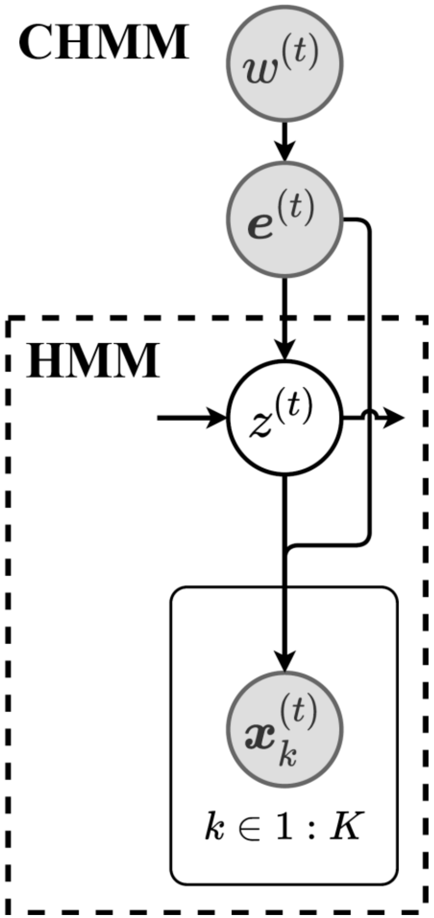
The house of Barack Obama…

Ideal:	$P(\text{PER}\mid\text{others}) = 0.1$	$P(\text{PER}\mid\text{others}) = 0.8$	Different ✓
HMM:	$P(\text{PER}\mid\text{others}) = 0.2$	$P(\text{PER}\mid\text{others}) = 0.2$	Same ✗

Conditional Hidden Markov Model (CHMM)

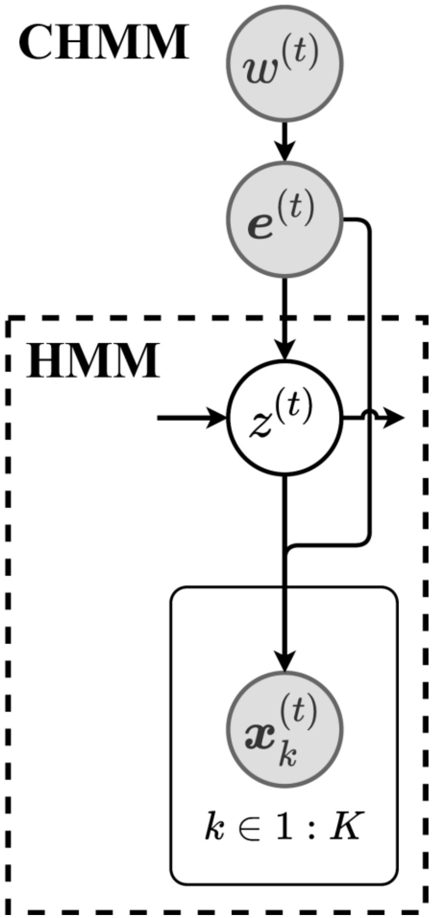


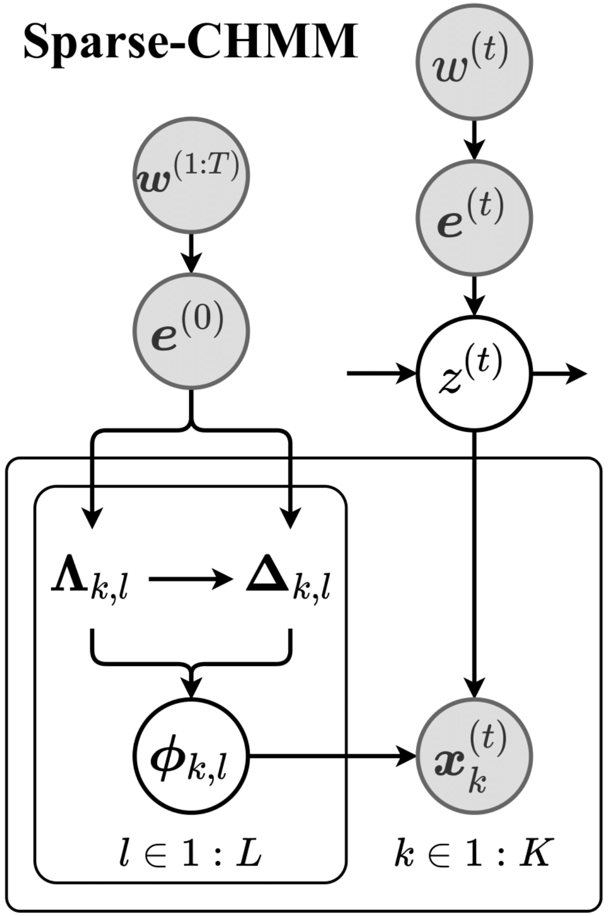
CHMM's Disadvantage



- CHMM directly predicts all elements in the emission matrix
  - $\Phi \in [0,1]^{K \times L \times L}$, linear layer to predict emission:
    $$\mathbb{R}^{d_{\text{model}} \times K \cdot L \cdot L}$$
- Large number of emission NN parameters
  - High degrees of freedom
  - More local optima
  - Slow training & inference
- **Solution**: restrict the number of trainable emission parameters

Sparse CHMM


CHMM
$w^{(t)}$
$e^{(t)}$
HMM
$z^{(t)}$
$x_k^{(t)}$
$k \in 1:K$


Sparse-CHMM
$w^{(t)}$
$w^{(1:T)}$
$e^{(t)}$
$e^{(0)}$
$z^{(t)}$
$\Lambda_{k,l}$
$\Delta_{k,l}$
$\phi_{k,l}$
$x_k^{(t)}$
$l \in 1:L$
$k \in 1:K$

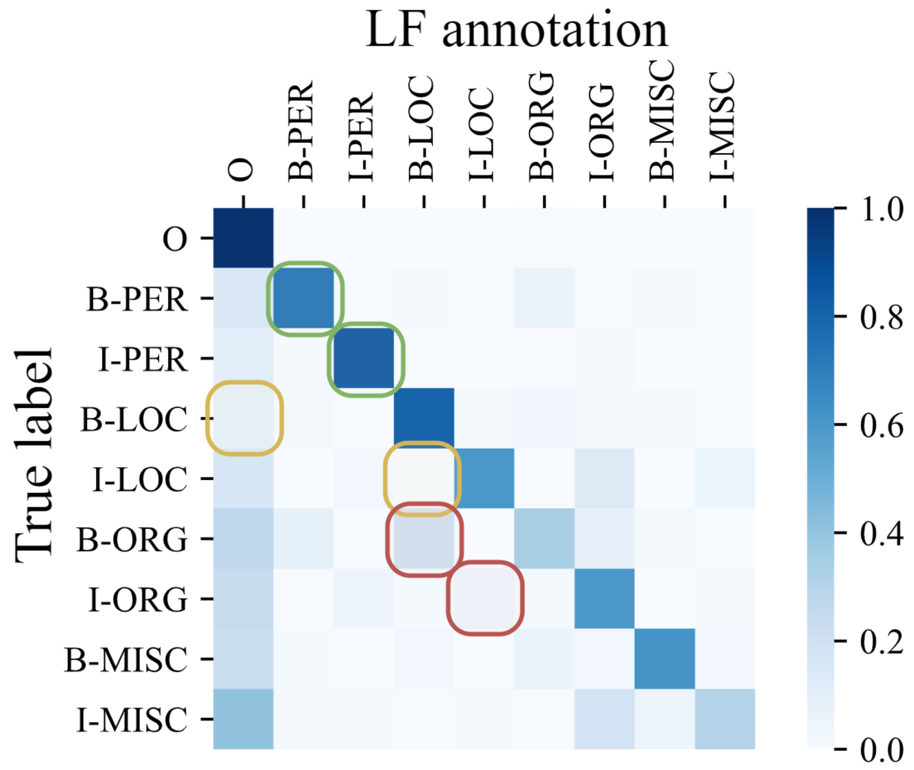
Expected Emission Matrix

Sentence	Joe	Biden	,	joined	by	White	House	staff	, ...
True label	B-PER	I-PER	O	O	O	B-ORG	I-ORG	O	O ...
LF annotation	B-PER	I-PER	O	O	O	B-LOC	I-LOC	O	O ...

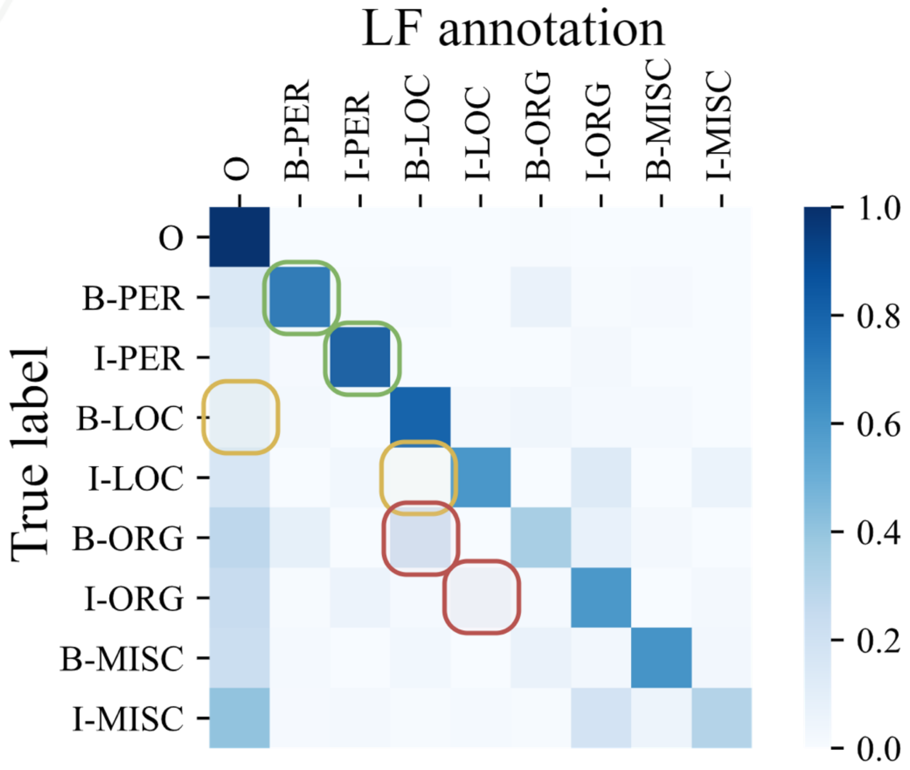
Sentence	The	White	House	is	the	official	residence	of ...
True label	O	B-LOC	I-LOC	O	O	O	O	O ...
LF annotation	O	O	B-LOC	O	O	O	O	O ...

: **Correct** annotations & corresponding probabilities

: **Incorrect** annotations & corresponding probabilities



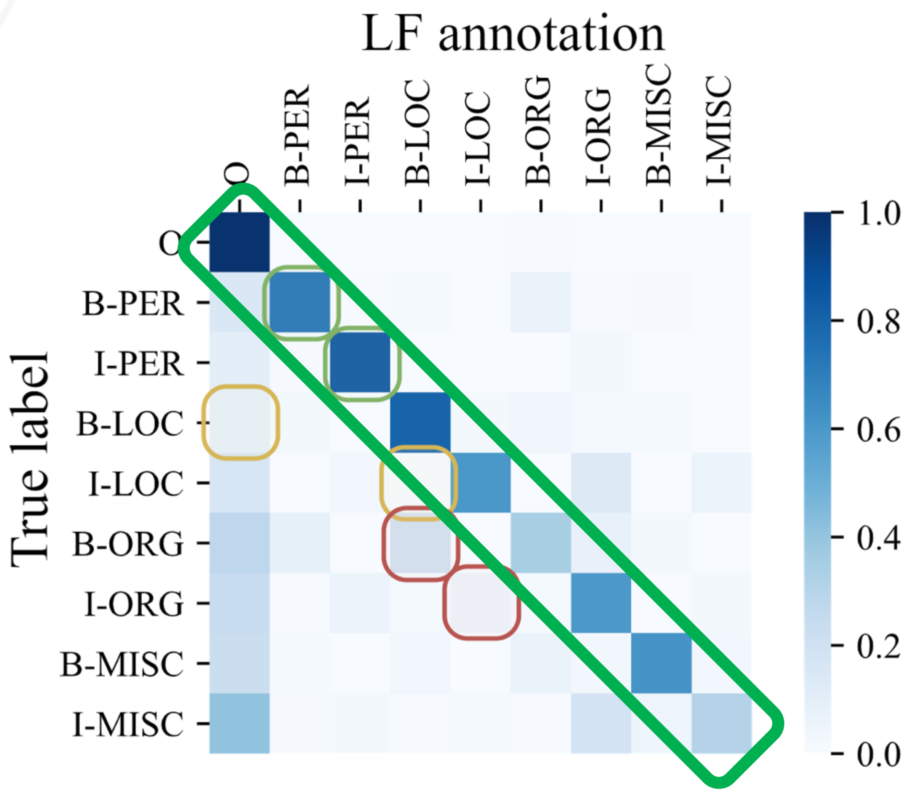
Emission Elements



- The emission: $\Phi_{k,i,j} \triangleq p(x_{k,j}^{(t)} = 1 | z^{(t)} = i)$
- Diagonal elements
  - the probabilities of LF k observing the true label
  - This can be regarded as LF k's reliability score
  - $\Phi_{k,l,l}$ are large → LF k is reliable; vice versa

- *If we know how LF k performes, can we construct the emission from it?*

Sparse CHMM



- Focus on predicting the emission diagonal, *i.e.*, LF reliability
- Expand the diagonal to matrix with heuristics

- Reduced trainable parameters
- Faster convergence rate
- Better overall model performance

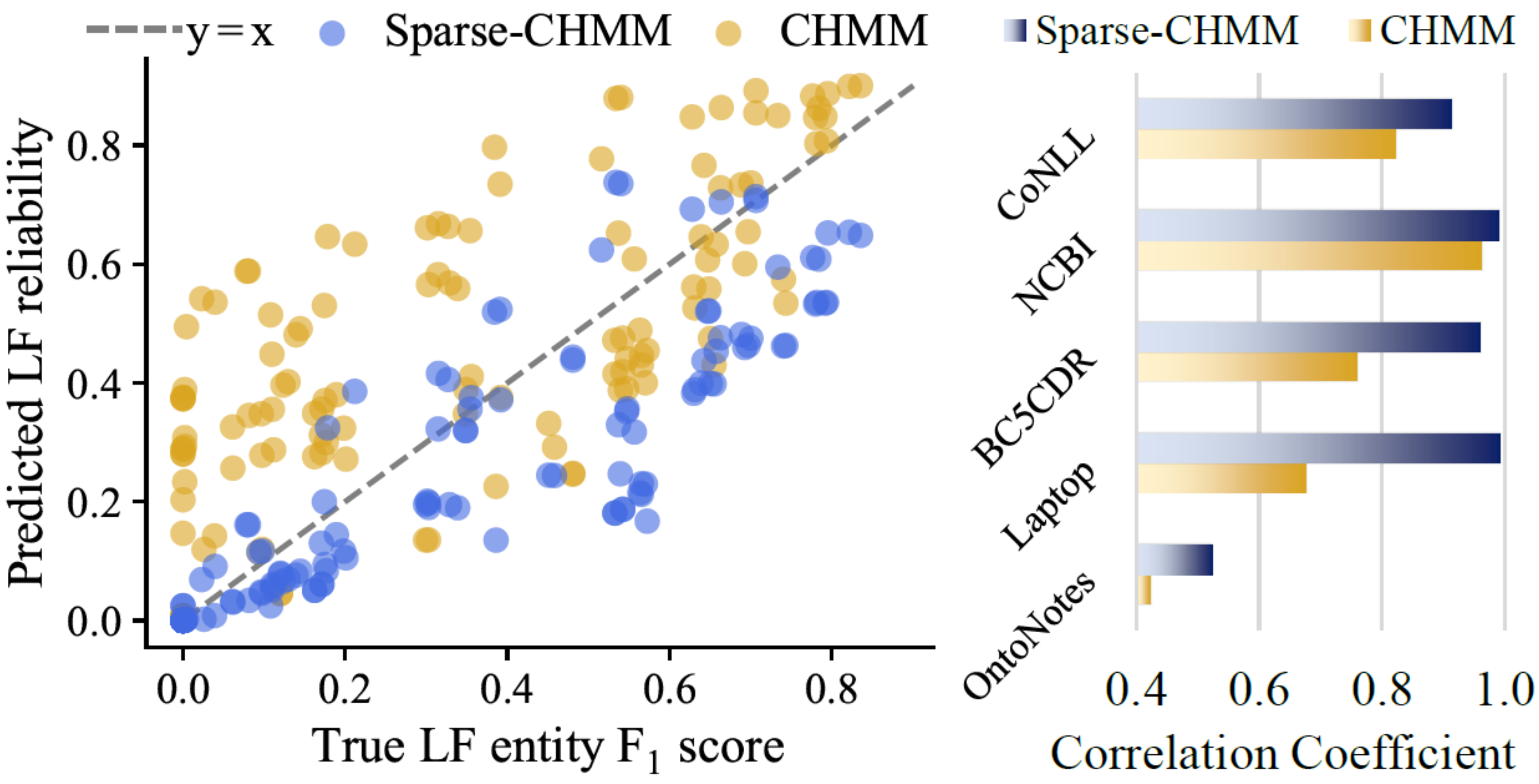
Main Results

Models		CoNLL 2003	NCBI-Disease	BC5CDR	LaptopReview	OntoNotes 5.0
Supervised Methods	BERT-NER	90.74 (90.37 / 91.10)	88.89 (87.05 / 90.82)	88.81 (87.12 / 90.57)	81.34 (82.02 / 80.67)	84.11 (83.11 / 85.14)
	Best consensus	86.73 (98.62 / 77.39)	81.65 (99.85 / 69.06)	88.42 (99.86 / 79.33)	77.60 (100.0 / 63.40)	85.11 (97.35 / 75.61)
	CHMM-FE	71.43 (72.89 / 70.02)	81.86 (90.75 / 74.55)	86.45 (91.73 / 81.75)	72.38 (88.13 / 61.41)	67.99 (65.23 / 71.00)
Weakly Supervised Models	ConNet*	66.02 (67.98 / 64.19)	63.04 (74.55 / 55.16)	72.04 (77.71 / 67.18)	50.36 (63.04 / 42.73)	60.58 (59.43 / 61.83)
	MV*	60.36 (59.06 / 61.72)	78.44 (93.04 / 67.79)	80.73 (83.79 / 77.88)	73.27 (88.86 / 62.33)	58.85 (54.17 / 64.40)
	Snorkel*	62.43 (61.62 / 63.26)	78.44 (93.04 / 67.79)	83.50 (91.69 / 76.65)	73.27 (88.86 / 62.33)	61.85 (57.44 / 66.99)
	HMM*	62.18 (66.42 / 58.45)	66.80 (**96.79** / 51.00)	71.57 (**93.48** / 57.98)	73.63 (89.30 / 62.63)	55.67 (57.95 / 53.57)
	CHMM*	63.22 (61.93 / 64.56)	78.74 (93.21 / 68.15)	83.66 (91.76 / 76.87)	73.26 (88.79 / 62.36)	64.06 (59.70 / **69.09**)
	Sparse-CHMM	**71.53** (**73.80** / **69.39**)	**82.24** (93.18 / **73.60**)	**86.63** (89.56 / **83.88**)	**75.90** (**91.94** / **64.62**)	**64.85** (**61.26** / 68.88)

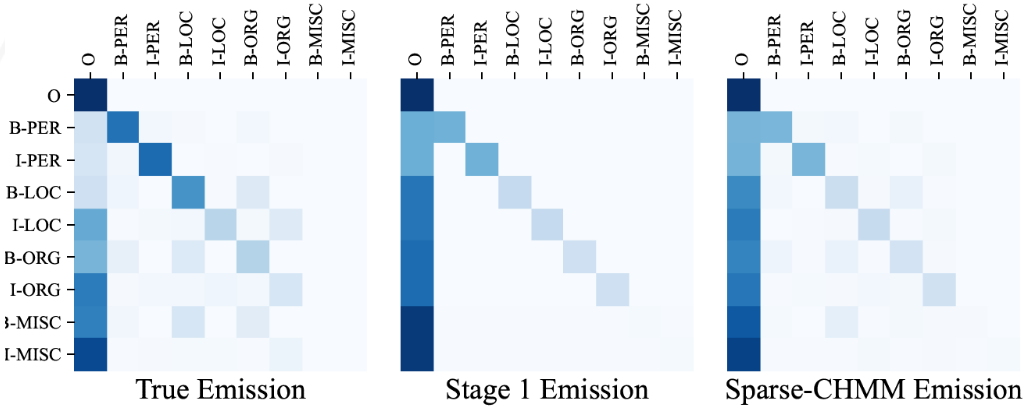
* Results are from the Wrench benchmark [31]. All weakly supervised models are evaluated with identical data and weak annotations.

- The models are trained on the training set (no labels) and tested on the test set (w/ gt, only for evaluation)
- The validation set is for early stopping and hyper-parameter fine-tuning

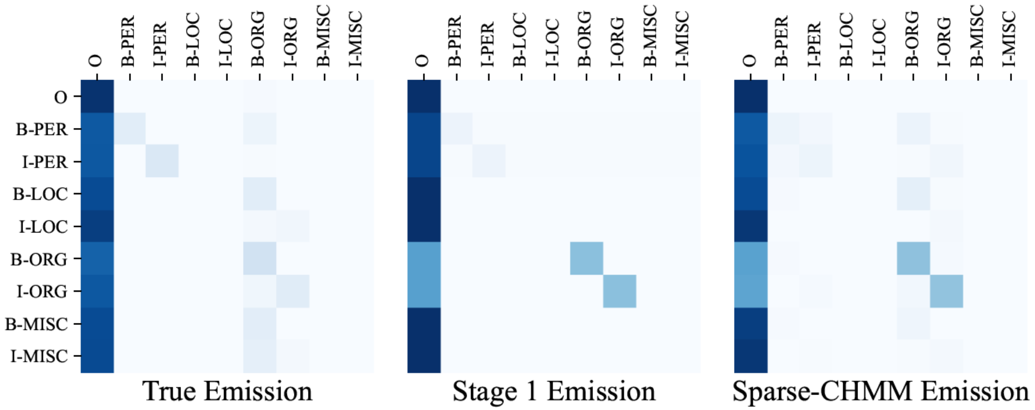
Reliability Prediction



Case Study



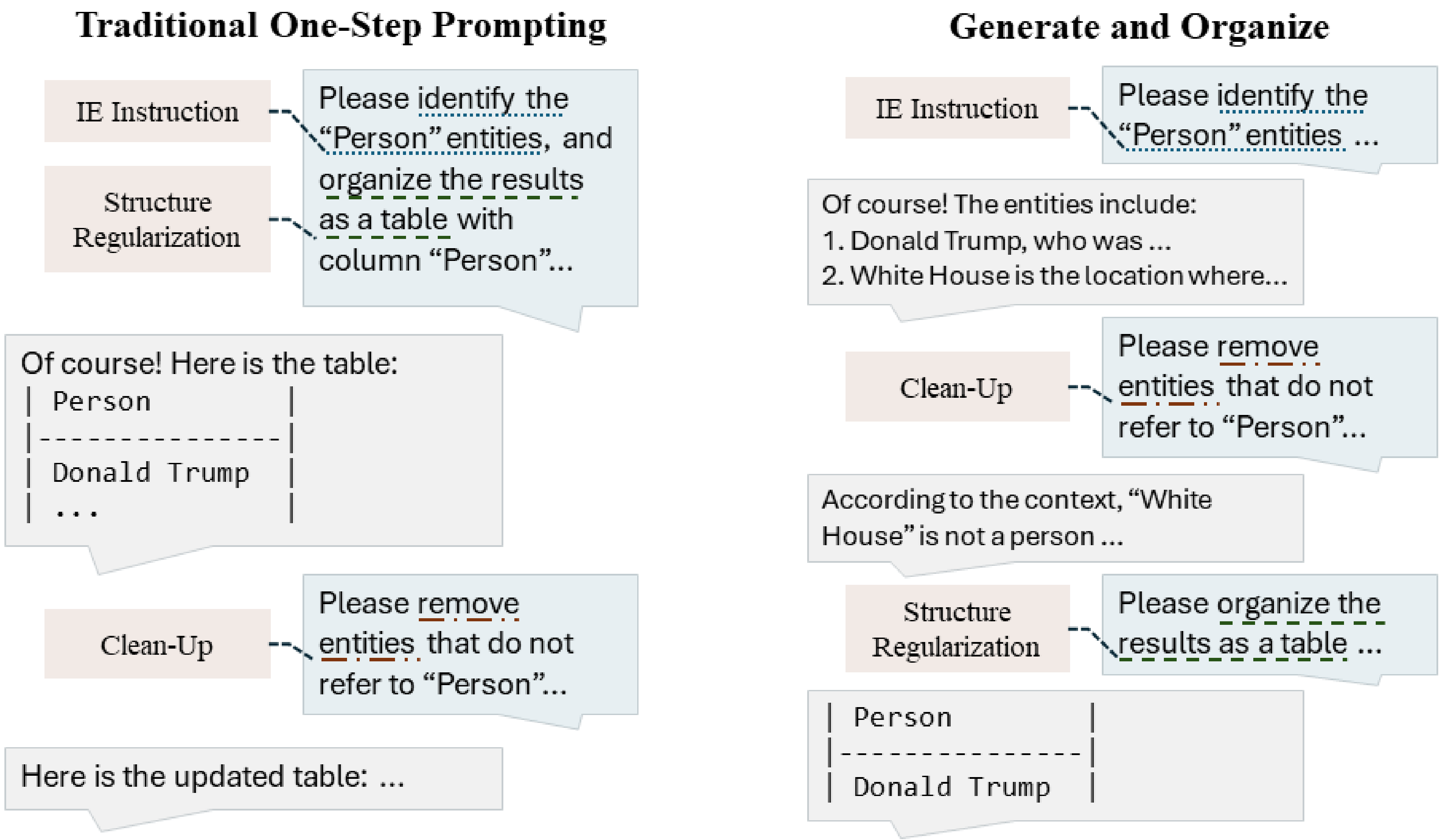
(a) Emissions of LF BTC



(b) Emissions of LF crunchbase_uncased

- Sparse-CHMM focuses on the diagonal and emit-to-O at stage 1
- It then refines the emission by adding the prominent off-diagonal back to the matrix
- Sparse-CHMM fits the LF reliabilities well without using any clean labeled data.

Zero-Shot IE with LLMs



Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

03 Information Extraction

04 ELREA

Agenda

Reliable Uncertainty Quantification

01 MUBen

02 UQAC

Data-Efficient Model Learning

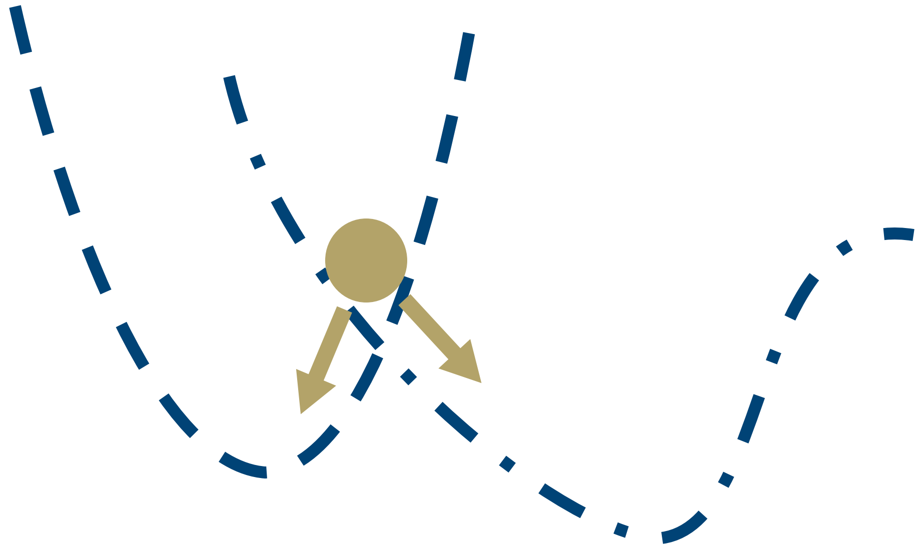
03 Information Extraction

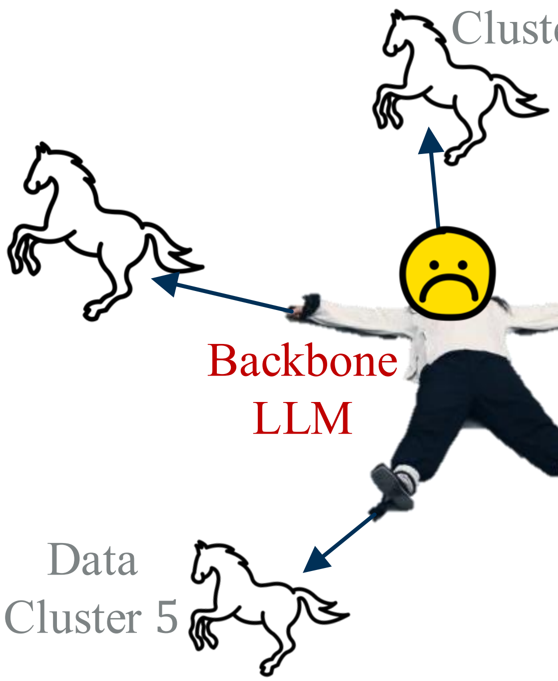
04 ELREA

In *ICLR 2025*: https://openreview.net/forum?id=l0gZS0sAlf

Motivation

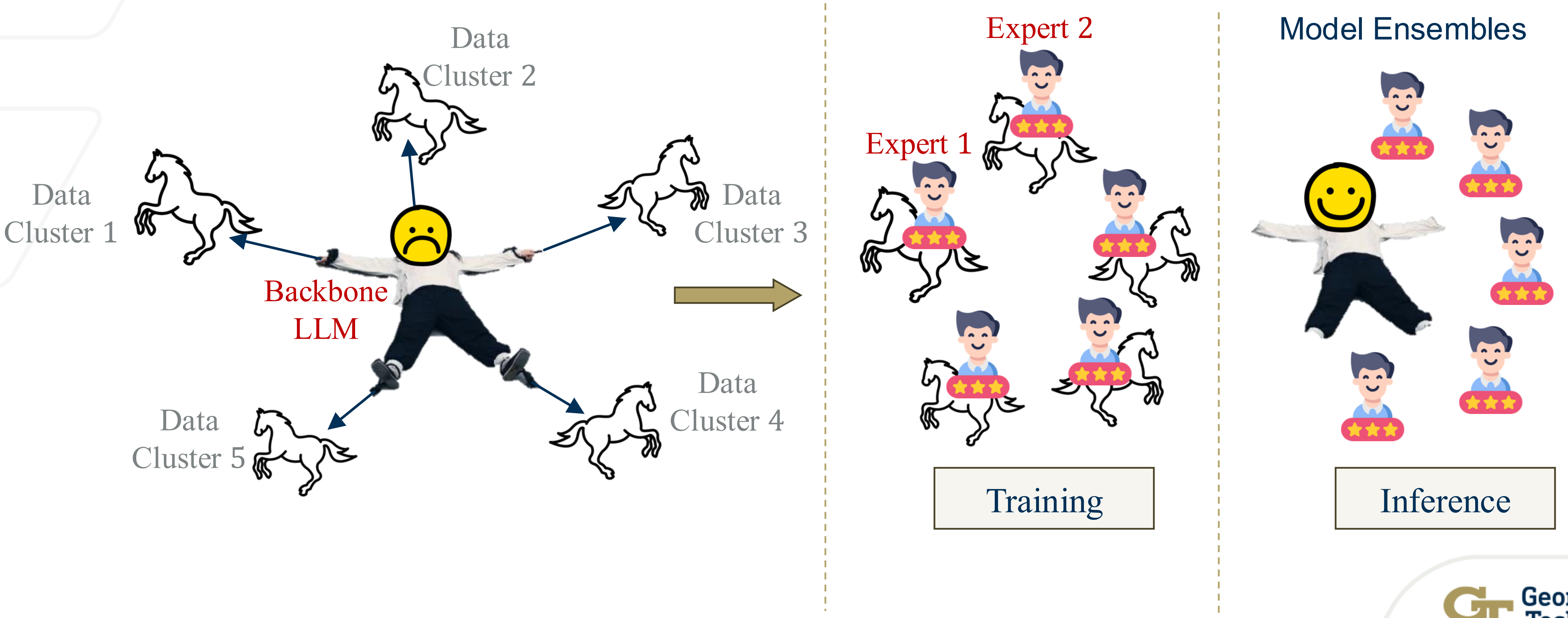
- Complicated task w/ diverse training data, data points may lead to different update direction
  - Resulting in under-optimized models
  - Especially for smaller models





Expert Training

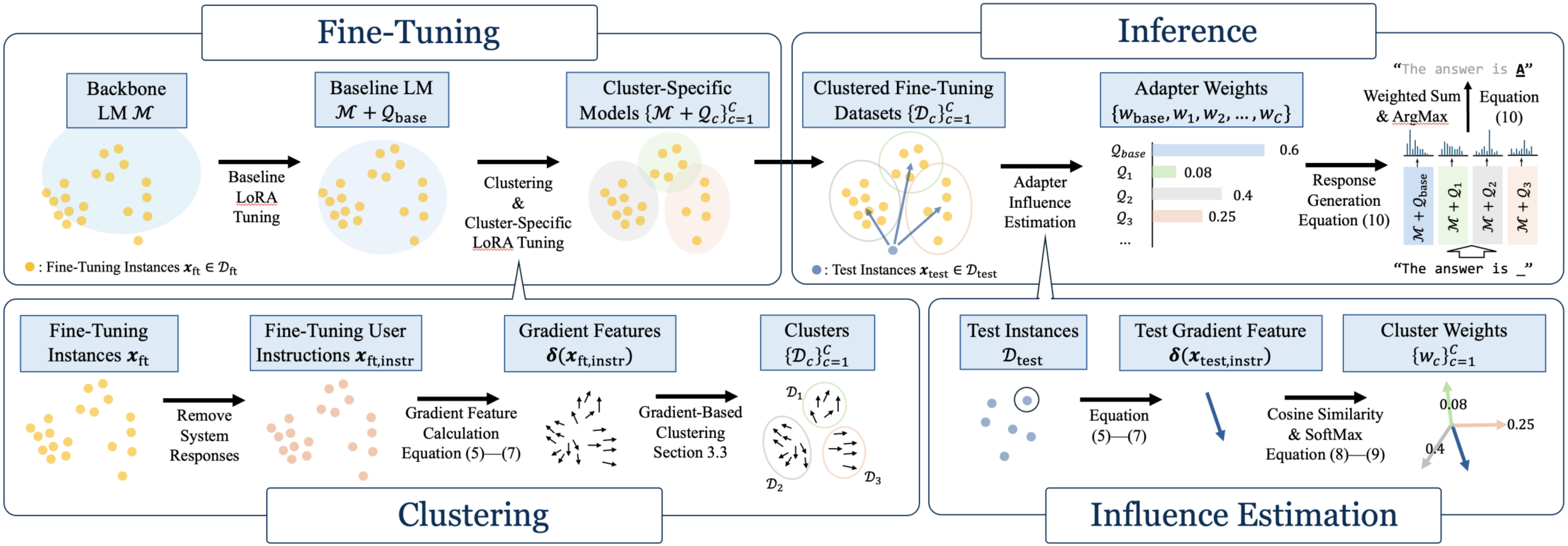
- Fit different expert models to different data clusters



Ensembles of Low-Rank Expert Adapters (ELREA)

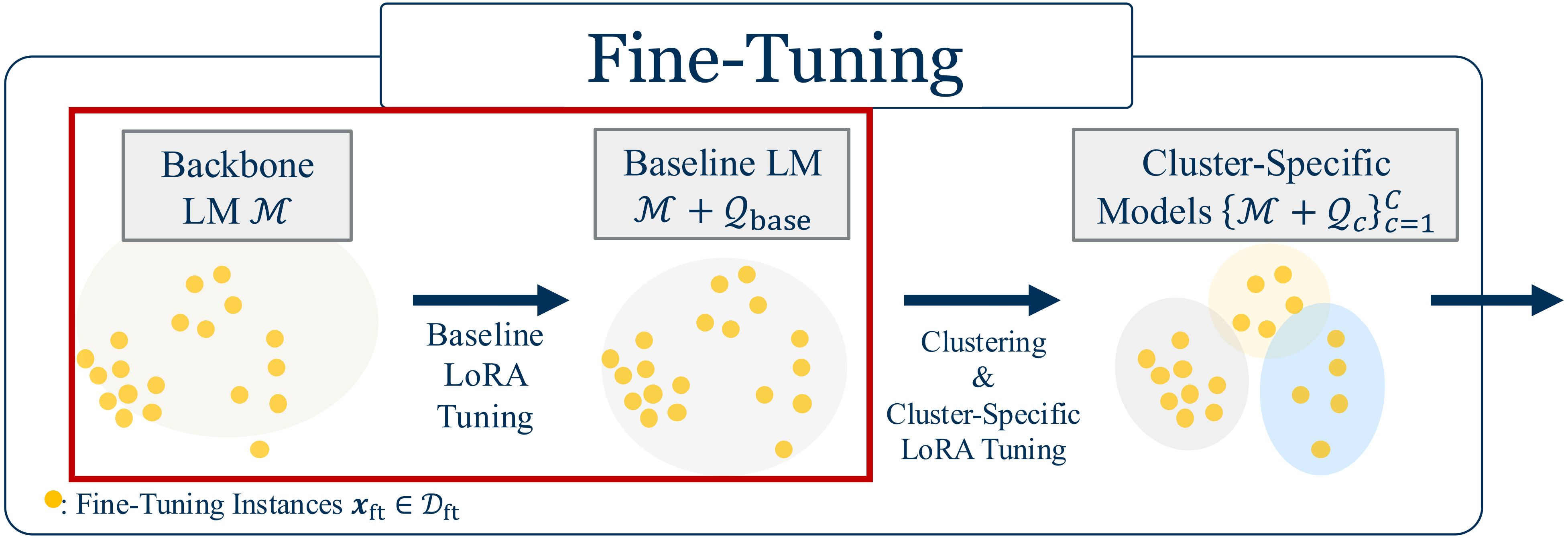
Data Clustering → Expert Training → Expert Ensembling

Pipeline

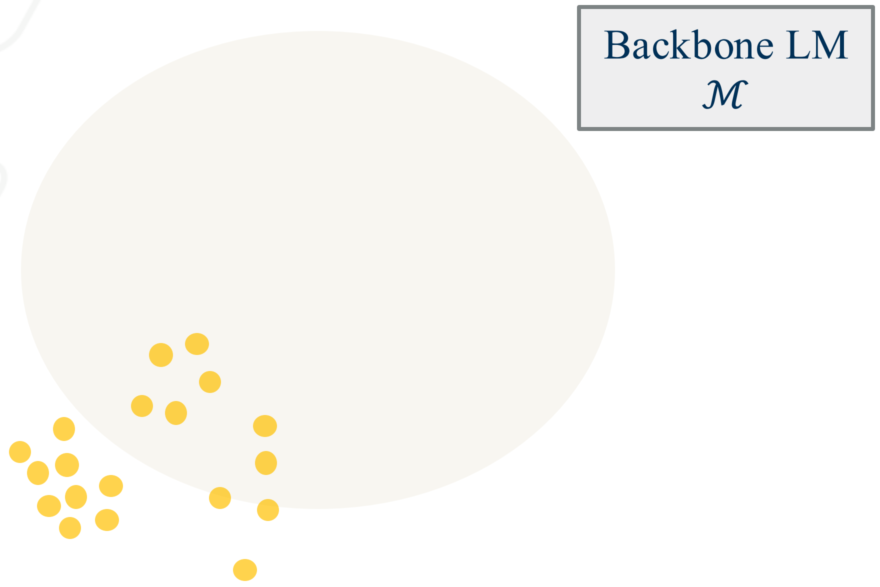


Fine-Tuning
Backbone LM $\mathcal{M}$
Baseline LoRA Tuning
Baseline LM $\mathcal{M} + \mathcal{Q}_{\text{base}}$
Clustering & Cluster-Specific LoRA Tuning
Cluster-Specific Models $\{\mathcal{M} + \mathcal{Q}_c\}_{c=1}^{C}$
: Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}} \in \mathcal{D}_{\text{ft}}$
Inference
Clustered Fine-Tuning Datasets $\{\mathcal{D}_c\}_{c=1}^{C}$
Adapter Influence Estimation
Adapter Weights $\{w_{\text{base}}, w_1, w_2, \dots, w_C\}$
$\mathcal{Q}_{base}$ 0.6
$\mathcal{Q}_1$ 0.08
$\mathcal{Q}_2$ 0.4
$\mathcal{Q}_3$ 0.25
...
Response Generation Equation (10)
Weighted Sum & ArgMax
Equation (10)
"The answer is A"
$\mathcal{M} + \mathcal{Q}_{\text{base}}$
$\mathcal{M} + \mathcal{Q}_1$
$\mathcal{M} + \mathcal{Q}_2$
$\mathcal{M} + \mathcal{Q}_3$
"The answer is _"
: Test Instances $\boldsymbol{x}_{\text{test}} \in \mathcal{D}_{\text{test}}$
Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}}$
Remove System Responses
Fine-Tuning User Instructions $\boldsymbol{x}_{\text{ft,instr}}$
Gradient Feature Calculation Equation (5)—(7)
Gradient Features $\boldsymbol{\delta}(\boldsymbol{x}_{\text{ft,instr}})$
Gradient-Based Clustering Section 3.3
Clusters $\{\mathcal{D}_c\}_{c=1}^{C}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_3$
Clustering
Test Instances $\mathcal{D}_{\text{test}}$
Equation (5)—(7)
Test Gradient Feature $\boldsymbol{\delta}(\boldsymbol{x}_{\text{test,instr}})$
Cosine Similarity & SoftMax Equation (8)—(9)
Cluster Weights $\{w_c\}_{c=1}^{C}$
0.08
0.25
0.4
Influence Estimation

Fine-Tuning


Fine-Tuning
Backbone LM $\mathcal{M}$
Baseline LoRA Tuning
Baseline LM $\mathcal{M} + \mathcal{Q}_{\text{base}}$
Clustering & Cluster-Specific LoRA Tuning
Cluster-Specific Models $\{\mathcal{M} + \mathcal{Q}_c\}_{c=1}^{C}$
: Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}} \in \mathcal{D}_{\text{ft}}$

Baseline LoRA Tuning



: Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}} \in \mathcal{D}_{\text{ft}}$

The backbone LM (Gemma, GPT, etc.) may not initially fit our fine-tuning data very well or lack domain knowledge

User Instruction

```
User: 1 + 1 = ?
```

System Response

```
System: A wise man once said: "…
```

Baseline LoRA Tuning

Baseline LM
$\mathcal{M} + \mathcal{Q}_{\text{base}}$

: Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}} \in \mathcal{D}_{\text{ft}}$

Fine-tune the backbone LM on **all** data to inject domain knowledge or to adjust model distribution

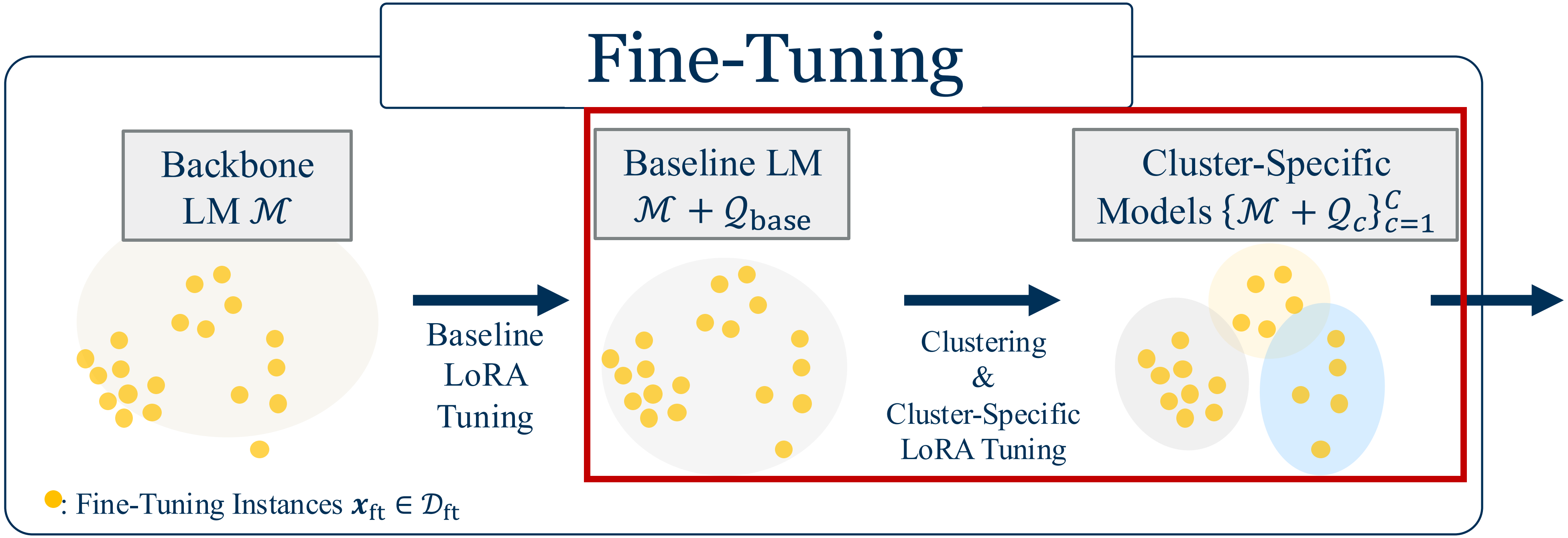
User Instruction

```
User: 1 + 1 = ?
```

System Response

```
System: A wise man once said: "…
```

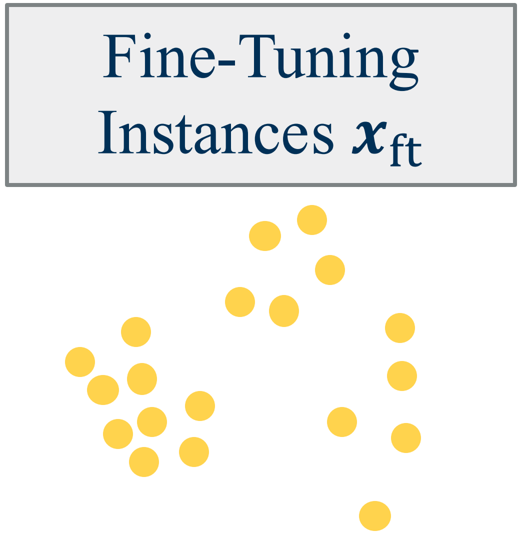
Fine-Tuning



Fine-Tuning
Backbone
LM $\mathcal{M}$
Baseline
LoRA
Tuning
Baseline LM
$\mathcal{M} + \mathcal{Q}_{\text{base}}$
Clustering
&
Cluster-Specific
LoRA Tuning
Cluster-Specific
Models $\{\mathcal{M} + \mathcal{Q}_c\}_{c=1}^{C}$
: Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}} \in \mathcal{D}_{\text{ft}}$

Data Clustering

- Remove system responses in the training data



User Instruction

```
User: 1 + 1 = ?
```

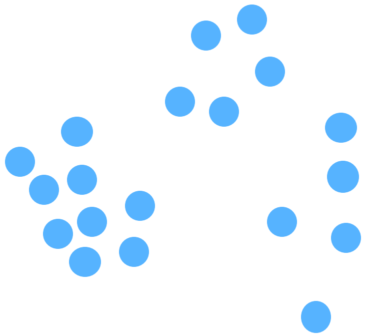
System Response

```
System: A wise man once said: "…
```

Data Clustering

- Remove system responses in the training data

Fine-Tuning User Instructions $\boldsymbol{x}_{\text{ft,instr}}$



User Instruction

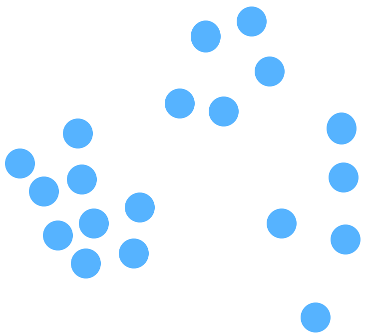
```
User: 1 + 1 = ?
```

System Response

```
System: A wise man once said: "…
```

Gradient Features

Fine-Tuning User Instructions $\boldsymbol{x}_{\text{ft,instr}}$

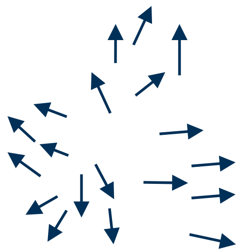


Gradient Feature Calculation

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}((\underbrace{\mathcal{M}+\mathcal{Q}_{\text{base}})}_{\text{Baseline Adapter}}(\boldsymbol{x}_{\text{ft,instr}}))$$

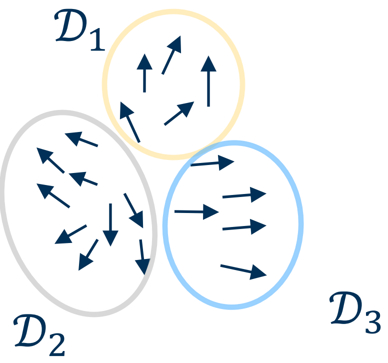
Baseline Adapter parameters

Unit Gradient Features $\boldsymbol{\delta}(\boldsymbol{x}_{\text{ft,instr}})$



Clustering Gradient Features

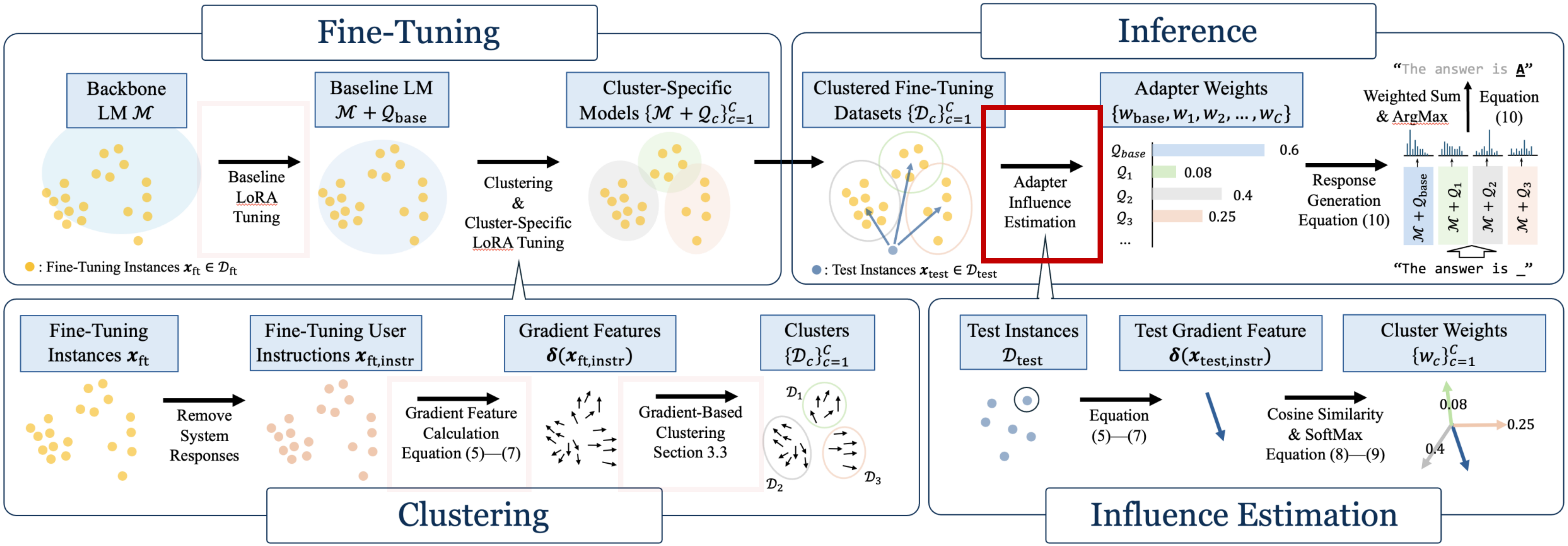
Unit Gradient Features $\boldsymbol{\delta}(\boldsymbol{x}_{\text{ft,instr}})$



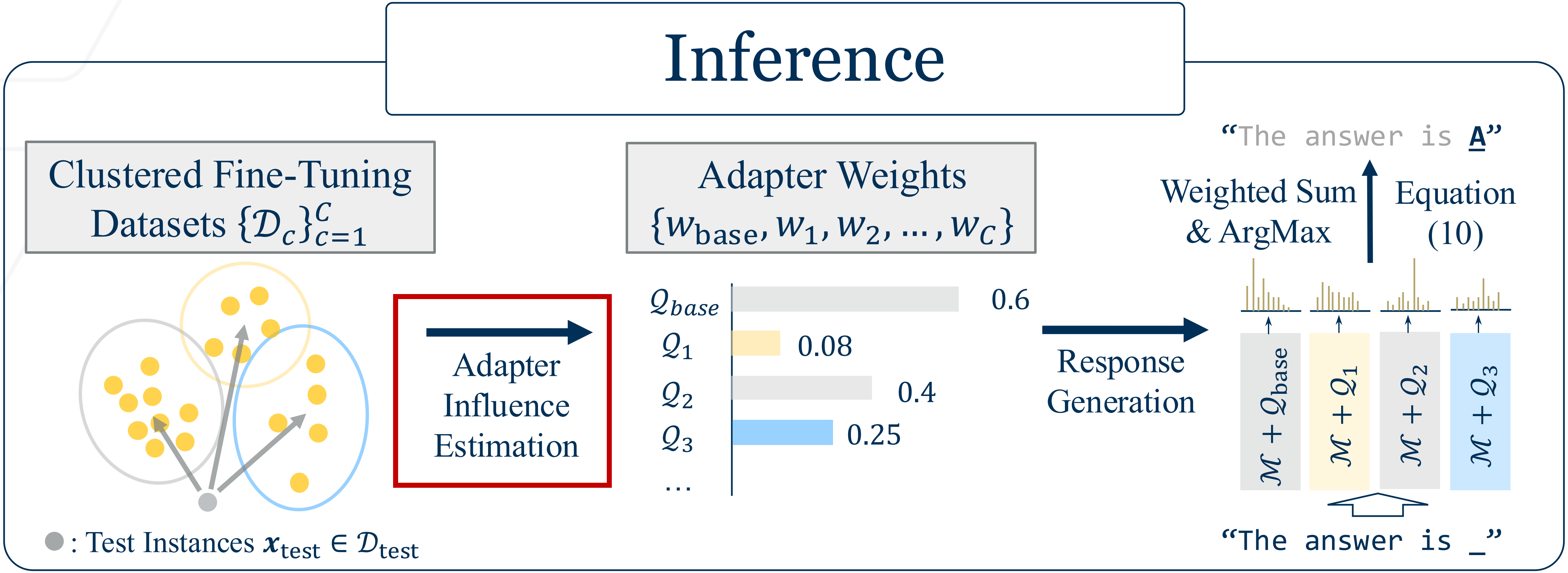
Clusters $\{\mathcal{D}_c\}_{c=1}^{C}$

Fitting Expert Adapters

Pipeline


Fine-Tuning
Backbone LM $\mathcal{M}$
Baseline LoRA Tuning
Baseline LM $\mathcal{M} + \mathcal{Q}_{\text{base}}$
Clustering & Cluster-Specific LoRA Tuning
Cluster-Specific Models $\{\mathcal{M} + \mathcal{Q}_c\}_{c=1}^{C}$
: Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}} \in \mathcal{D}_{\text{ft}}$
Inference
Clustered Fine-Tuning Datasets $\{\mathcal{D}_c\}_{c=1}^{C}$
Adapter Influence Estimation
Adapter Weights $\{w_{\text{base}}, w_1, w_2, \dots, w_C\}$
$\mathcal{Q}_{base}$ 0.6
$\mathcal{Q}_1$ 0.08
$\mathcal{Q}_2$ 0.4
$\mathcal{Q}_3$ 0.25
...
Response Generation Equation (10)
Weighted Sum & ArgMax
Equation (10)
"The answer is A"
$\mathcal{M} + \mathcal{Q}_{\text{base}}$
$\mathcal{M} + \mathcal{Q}_1$
$\mathcal{M} + \mathcal{Q}_2$
$\mathcal{M} + \mathcal{Q}_3$
"The answer is _"
: Test Instances $\boldsymbol{x}_{\text{test}} \in \mathcal{D}_{\text{test}}$
Fine-Tuning Instances $\boldsymbol{x}_{\text{ft}}$
Remove System Responses
Fine-Tuning User Instructions $\boldsymbol{x}_{\text{ft,instr}}$
Gradient Feature Calculation Equation (5)—(7)
Gradient Features $\boldsymbol{\delta}(\boldsymbol{x}_{\text{ft,instr}})$
Gradient-Based Clustering Section 3.3
Clusters $\{\mathcal{D}_c\}_{c=1}^{C}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_3$
Clustering
Test Instances $\mathcal{D}_{\text{test}}$
Equation (5)—(7)
Test Gradient Feature $\boldsymbol{\delta}(\boldsymbol{x}_{\text{test,instr}})$
Cosine Similarity & SoftMax Equation (8)—(9)
Cluster Weights $\{w_c\}_{c=1}^{C}$
0.08
0.25
0.4
Influence Estimation

Inference



Inference
Clustered Fine-Tuning Datasets $\{\mathcal{D}_c\}_{c=1}^{C}$
Adapter Weights $\{w_{\text{base}}, w_1, w_2, \ldots, w_C\}$
Adapter Influence Estimation
$\mathcal{Q}_{base}$ 0.6
$\mathcal{Q}_1$ 0.08
$\mathcal{Q}_2$ 0.4
$\mathcal{Q}_3$ 0.25
…
Response Generation
$\mathcal{M} + \mathcal{Q}_{\text{base}}$
$\mathcal{M} + \mathcal{Q}_1$
$\mathcal{M} + \mathcal{Q}_2$
$\mathcal{M} + \mathcal{Q}_3$
Weighted Sum & ArgMax
Equation (10)
"The answer is A"
"The answer is _"
: Test Instances $\boldsymbol{x}_{\text{test}} \in \mathcal{D}_{\text{test}}$

Adapter Influence Estimation (Routing)

- To select the most appropriate adapter(s) for the test data point

Test Instances
$\mathcal{D}_{\text{test}}$

Gradient Feature Calculation

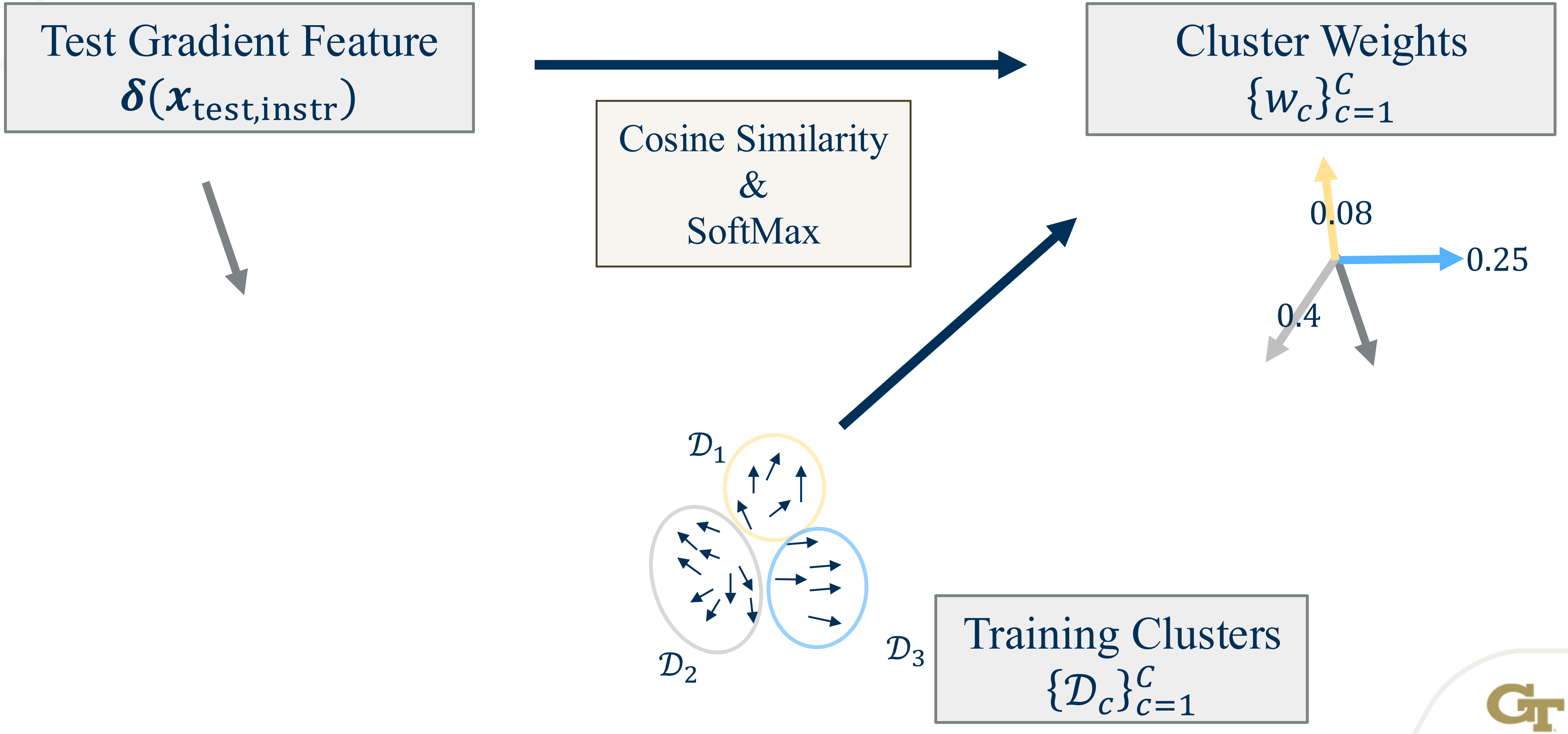
Test Gradient Feature
$\boldsymbol{\delta}(\boldsymbol{x}_{\text{test,instr}})$

User Instruction

```
User: 1 + 2 = ? Explain.
```

Test data points **only** contain user instructions

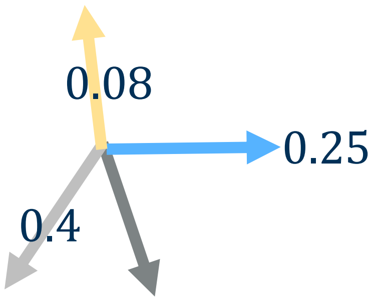
Adapter Influence Estimation (Routing)



Base Adapter

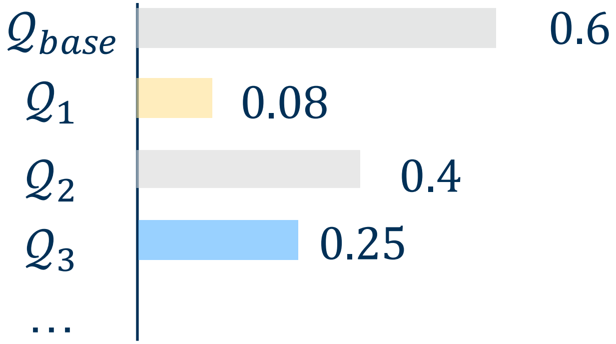
Cluster Weights

$\{w_c\}_{c=1}^{C}$



Adapter Weights

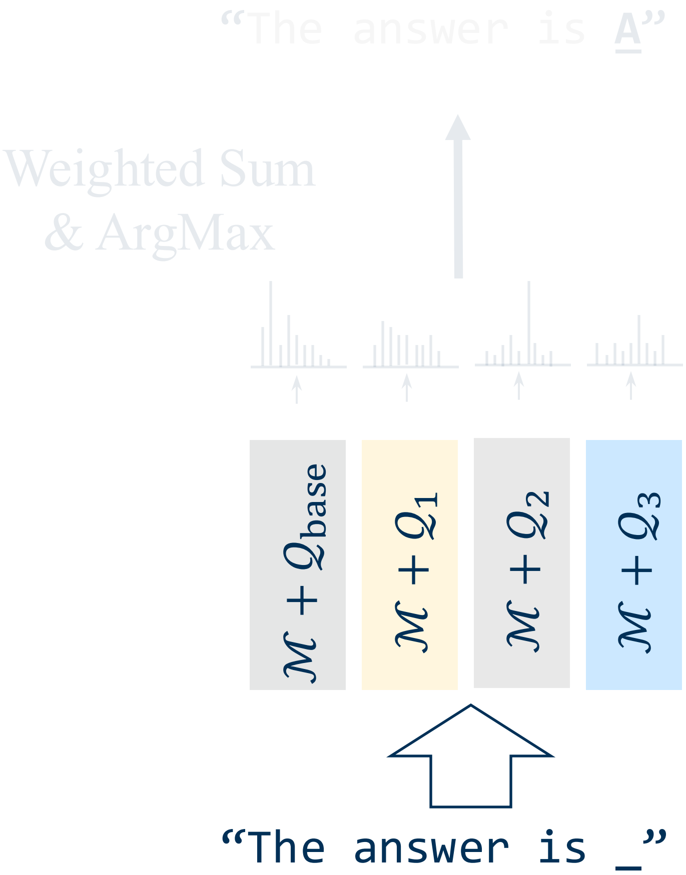
$\{w_{\text{base}}, w_1, w_2, \ldots, w_C\}$



Response Generation as Ensembles

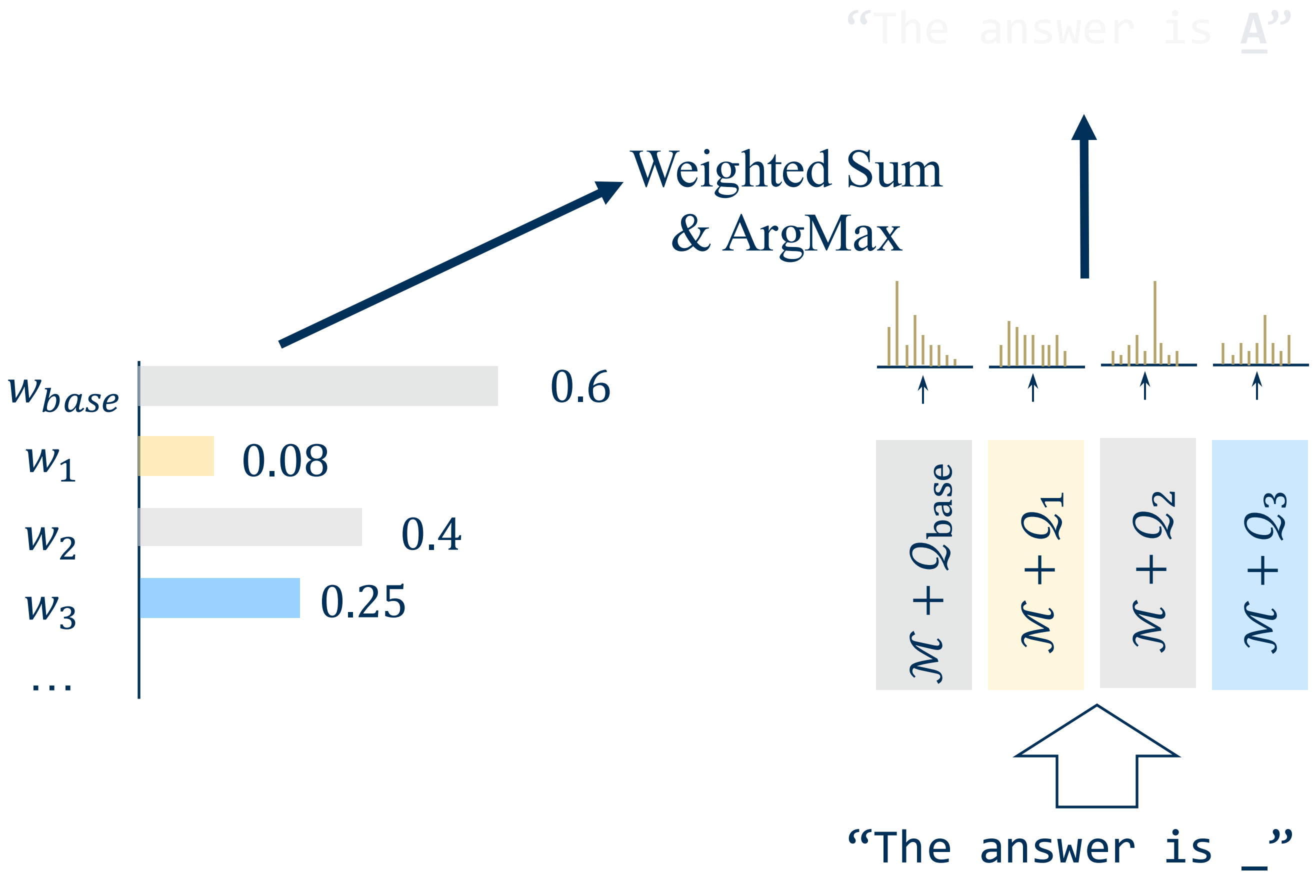
“The answer is _”

Response Generation as Ensembles

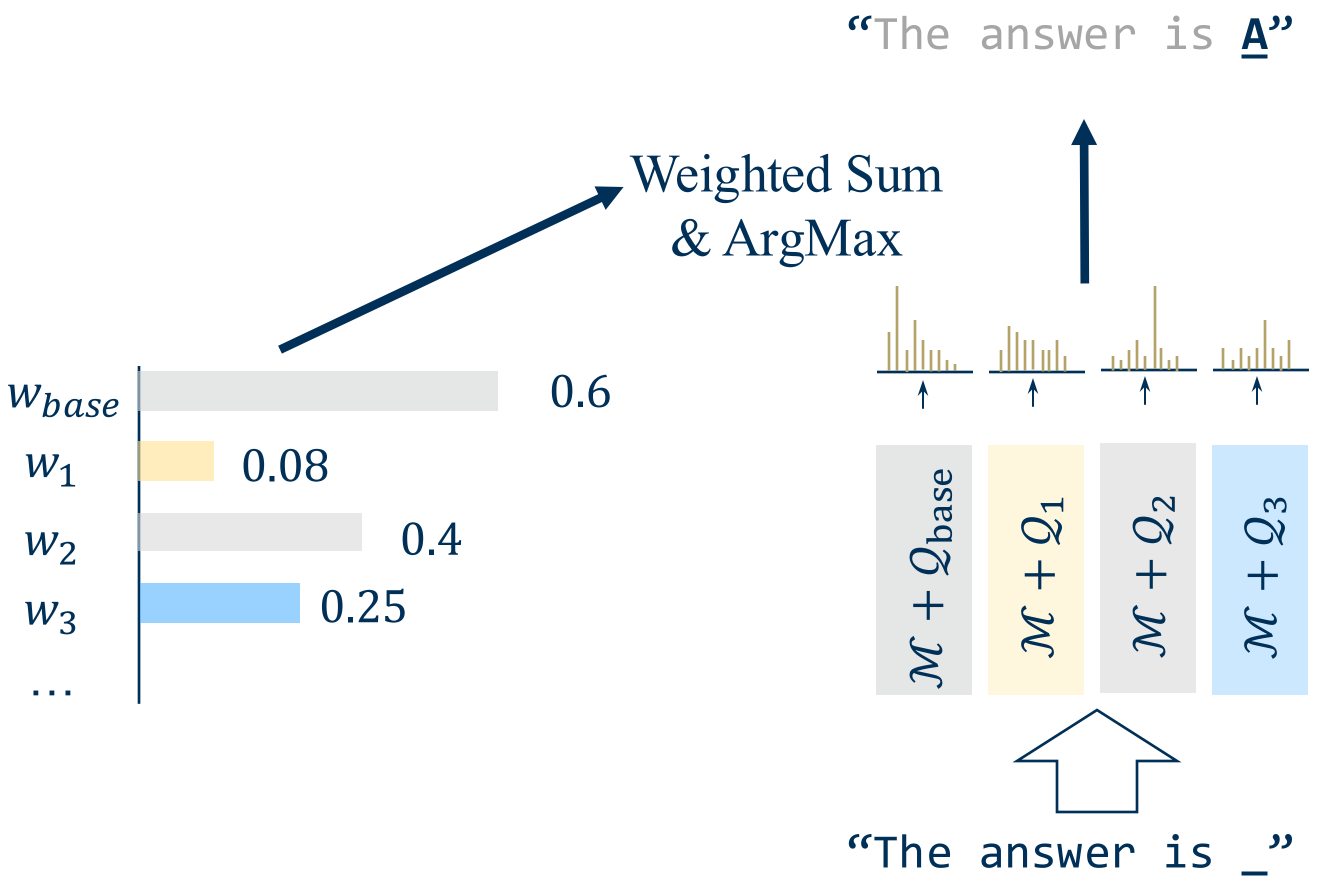


Response Generation as Ensembles

Response Generation as Ensembles



Response Generation as Ensembles



Main Results: Mathematical Reasoning

Methods	MATH	GSM8k	SVAMP	MathQA	Average (+Δ)
$\mathcal{M} + Q_{\text{base}}$	9.2	22.1	46.07	16.83	18.61
$\mathcal{M} + Q_{\text{dataset}}$	7.3	25.7	45	16.73	19.01 (+0.40)
MoE Routing	9.2	22.7	48.21	16.23	18.79 (+0.18)
MoE Merging	9.1	23.1	48.21	15.73	18.73 (+0.12)
MoLE	8.8	21.6	46.43	15.53	17.99 (-0.62)
LoRA Ensembles	9.3	24.7	47.5	16.73	19.55 (+0.94)
Self-Consistency	5.9	14.3	44.64	10.32	13.12 (-5.49)
Instruction Embedding	9.8	24.1	46.79	16.83	19.46 (+0.85)
ELREA	9.1	25.9	49.64	18.04	20.41 (+1.80)
Random Cluster	9.1	25.1	48.21	18.84	20.30 (+1.69)
Uniform Weights	9.6	25.2	47.5	18.04	20.16 (+1.55)

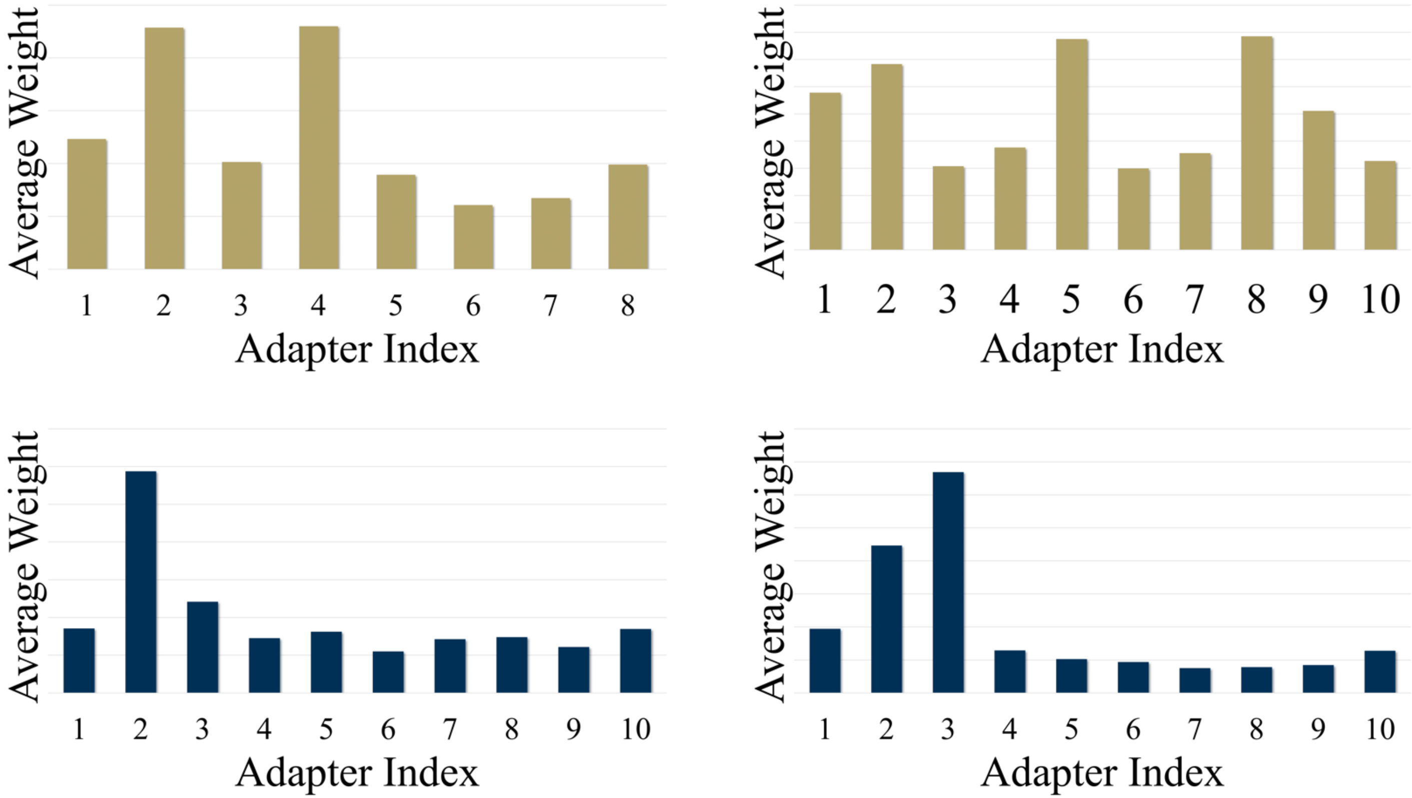
Expert Weight Distribution

Rank = 8

Rank = 64

Mathematical Reasoning

General Language Understanding (OOD Test)

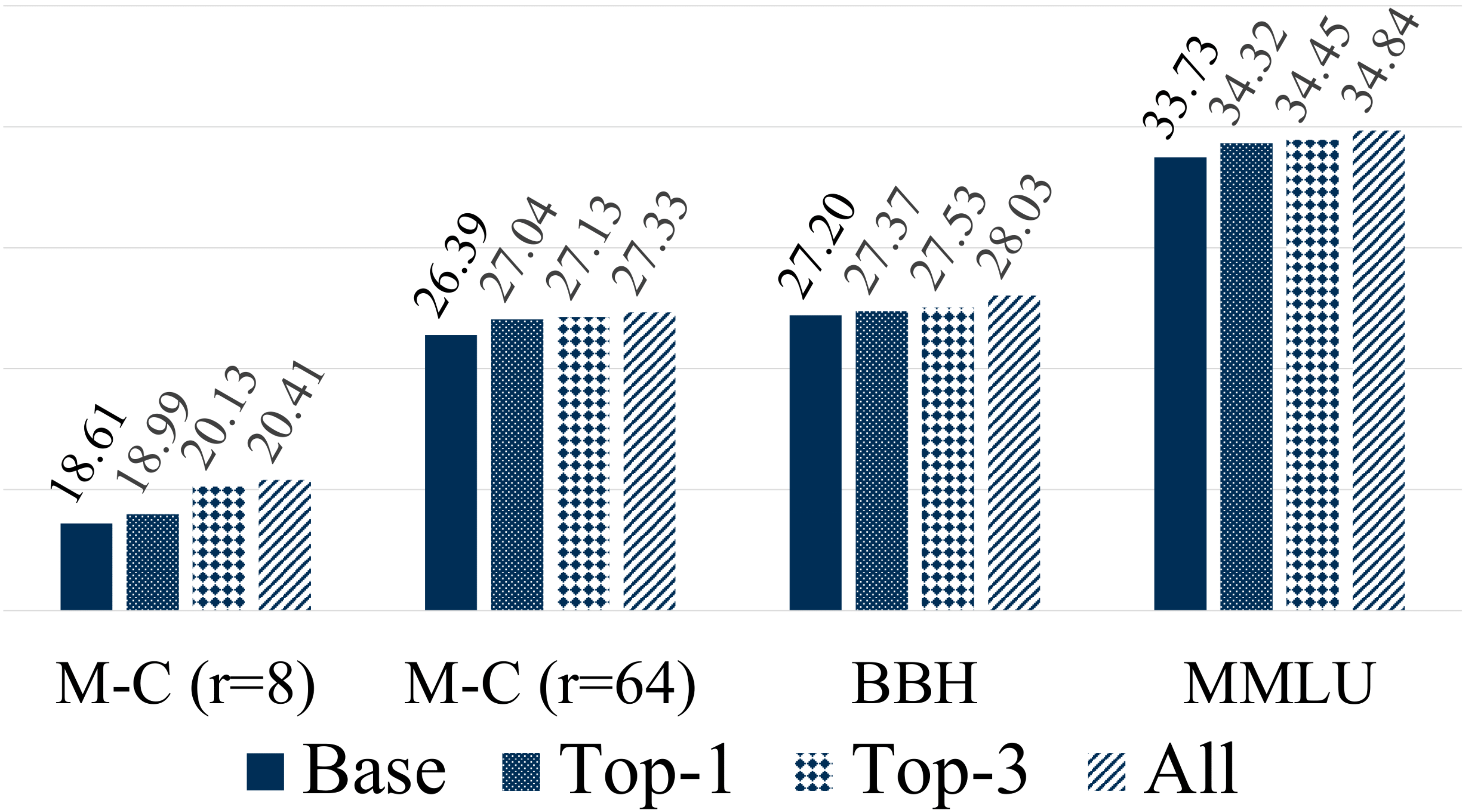


BBH

MMLU

Georgia Tech

Top-k Experts



Conclusion and Future Works

Summary

- Characteristics of molecular foundation models; and how to select appropriate UQ methods
- How to more reliably estimate the confidence of LLM responses
- How to conduct information extraction without relying on manual labels
- How to improve model performance without additional training data

Future Works

- Tighter connection between UQ and model learning

Thanks for Attending!