NEURAL INFORMATION PROCESSING SYSTEMS

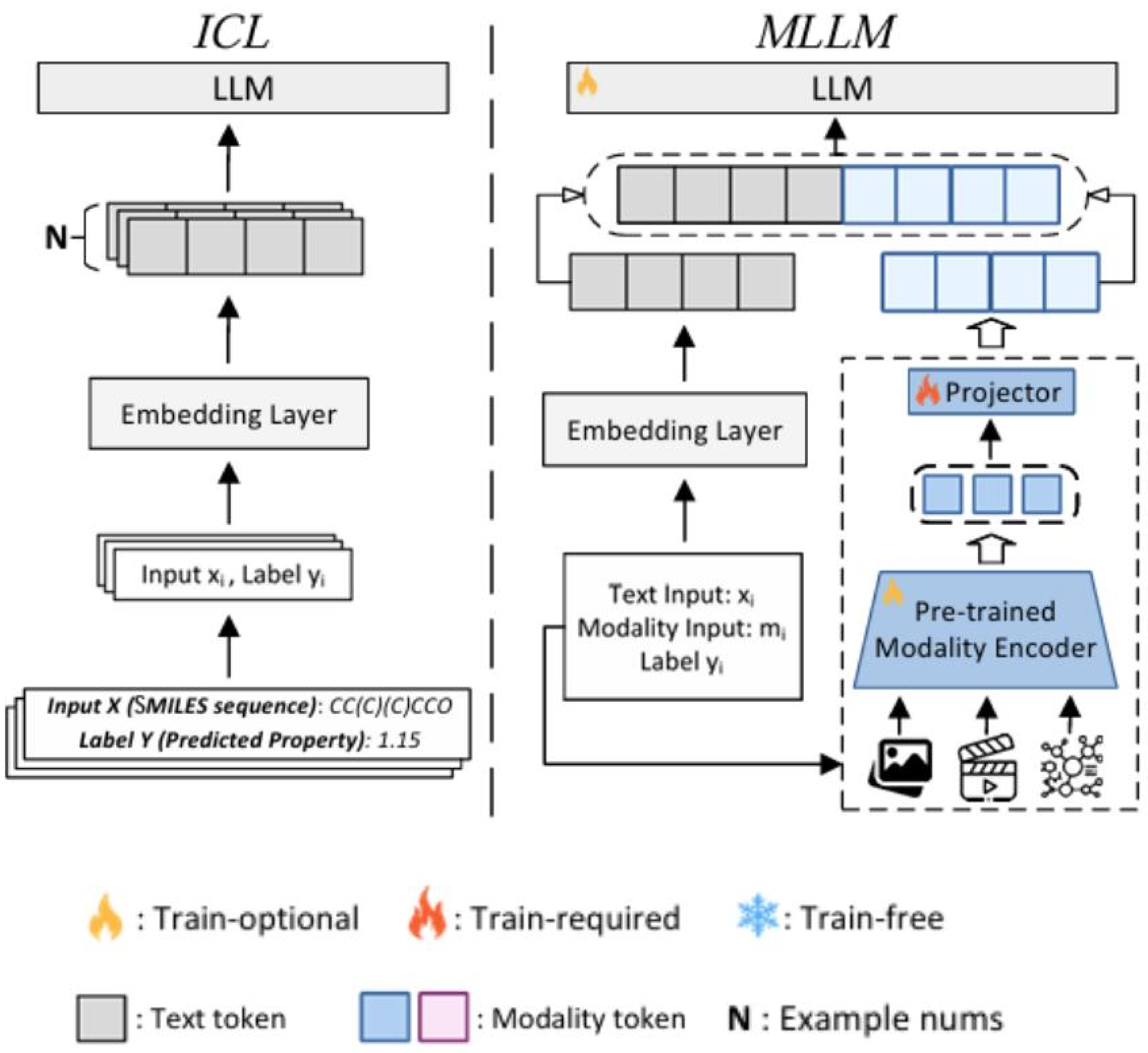
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Can LLMs Reason Over Non-Text Modalities in a Training-Free Manner? A Case Study with In-Context Representation Learning

Tianle Zhang, Wanlong Fang*, Jonathan Woo*, Paridhi Latawa, Deepak A. Subramanian, Alvin Chan.*

NeurIPS 2025

Challenges of Text-Only LLMs in Leveraging Non-Text Modalities



Motivations:

- Many domains such as molecules, proteins, vision, and speech rely on non-text data.
- Most multimodal methods rely on **costly supervised training**, limiting adaptation to new domains.

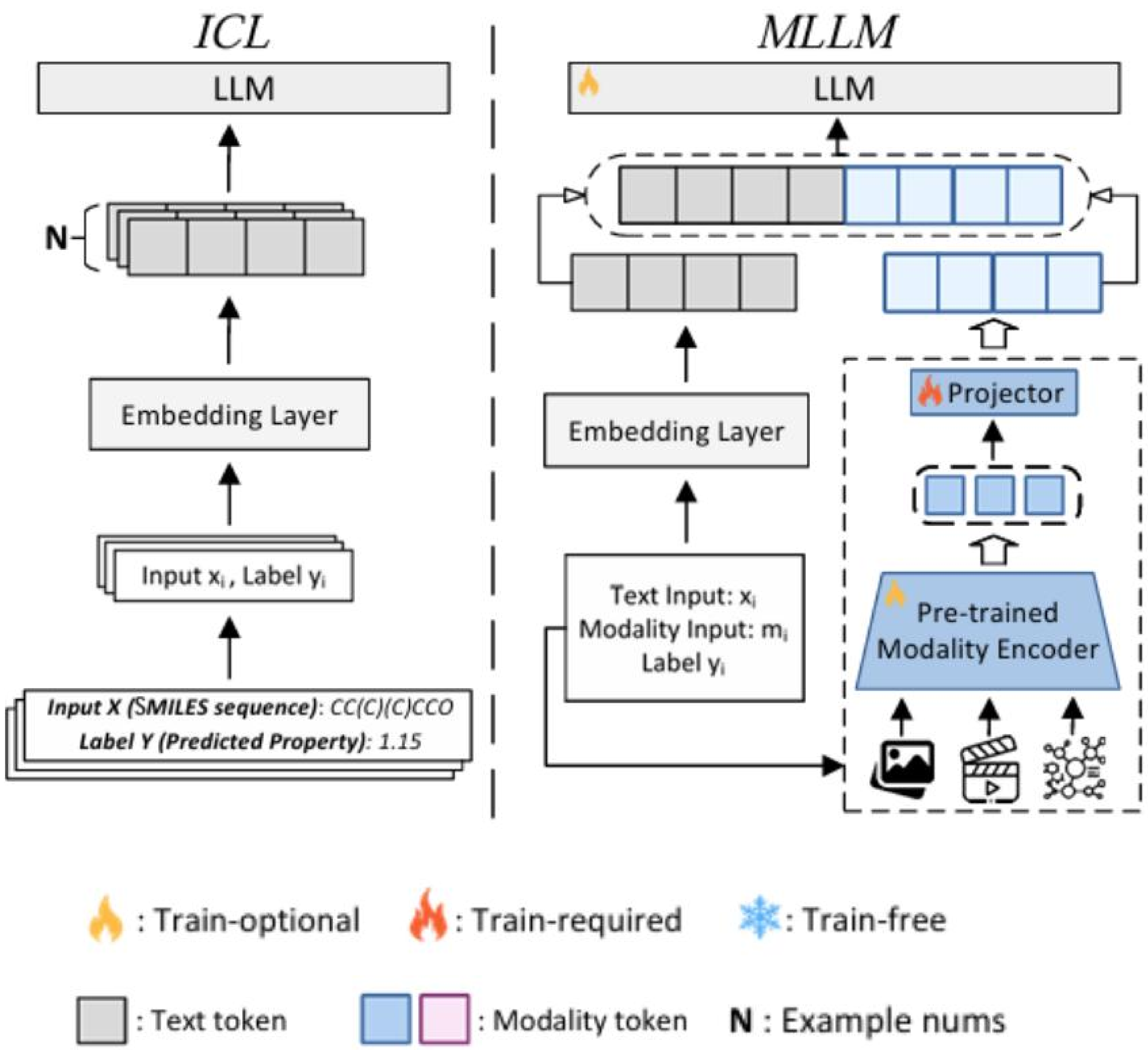
Current solutions

Multi-Modal Large Language Models:

✓ Capable of integrating **diverse modalities**.

× Require **additional and costly** training.
-- even for lightweight projector tuning.

Challenges of Text-Only LLMs in Leveraging Non-Text Modalities



Motivations:

- Many domains such as molecules, proteins, vision, and speech rely on non-text data.
- Most multimodal methods rely on **costly supervised training**, limiting adaptation to new domains.

Current solutions

Multi-Modal Large Language Models:

✓ Capable of integrating **diverse modalities**.

× Require **additional and costly** training.

-- even for lightweight projector tuning.

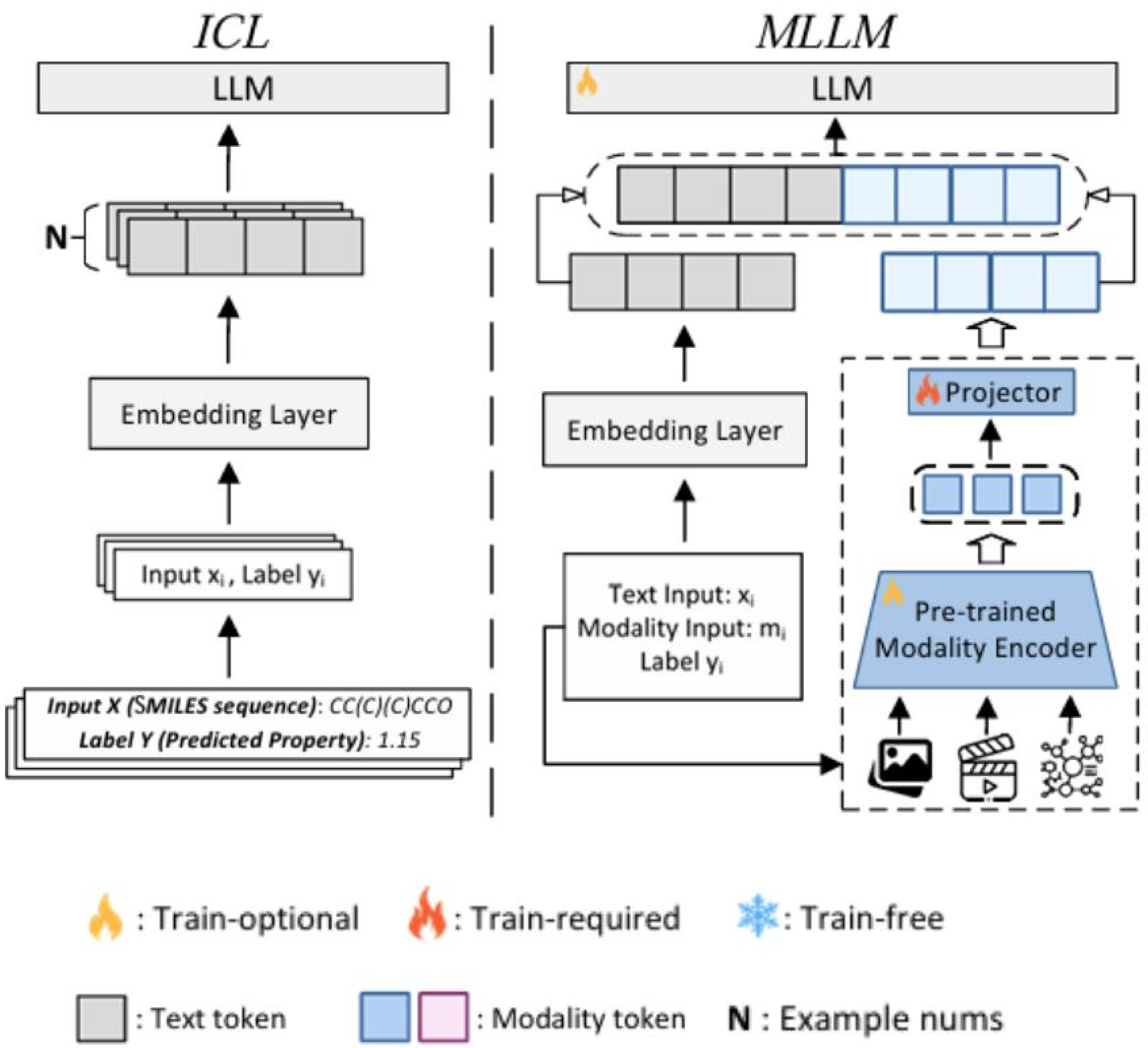
In-Context Learning:

✓ **Training-free** and data-efficient.

× Restricted to **text-only** inputs.

-- cannot directly leverage non-text features.

Challenges of Text-Only LLMs in Leveraging Non-Text Modalities



Motivations:

- Many domains such as molecules, proteins, vision, and speech rely on non-text data.
- Most multimodal methods rely on **costly supervised training**, limiting adaptation to new domains.

Current solutions

Multi-Modal Large Language Models:

✓ Capable of integrating **diverse modalities**.

× Require **additional and costly** training.
-- even for lightweight projector tuning.

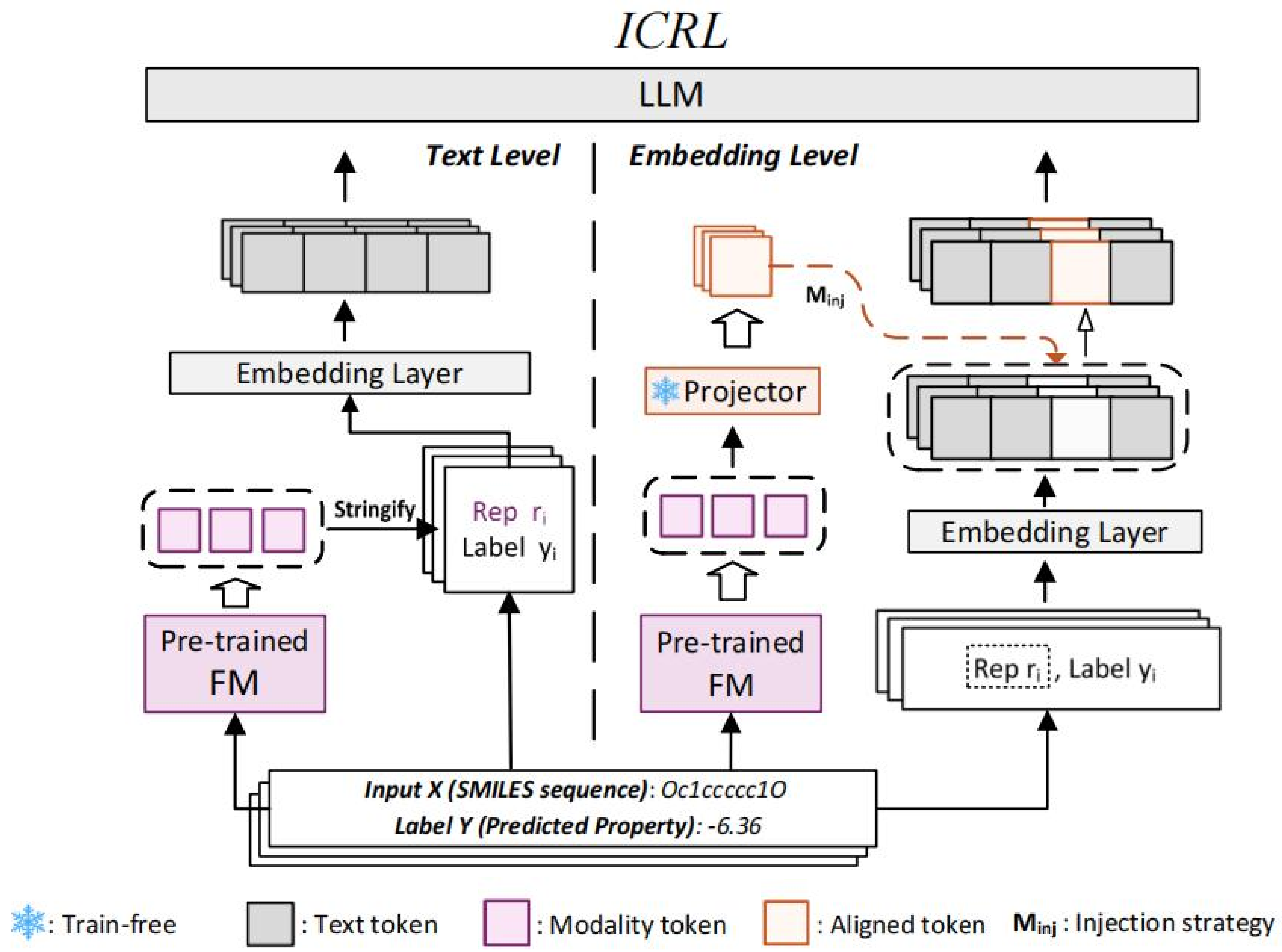
In-Context Learning:

✓ **Training-free** and data-efficient.

× Restricted to **text-only** inputs.
-- cannot directly leverage non-text features.

Can LLMs directly leverage non-text foundation models representations directly at inference time, without training?

ICRL: In-Context Representation Learning

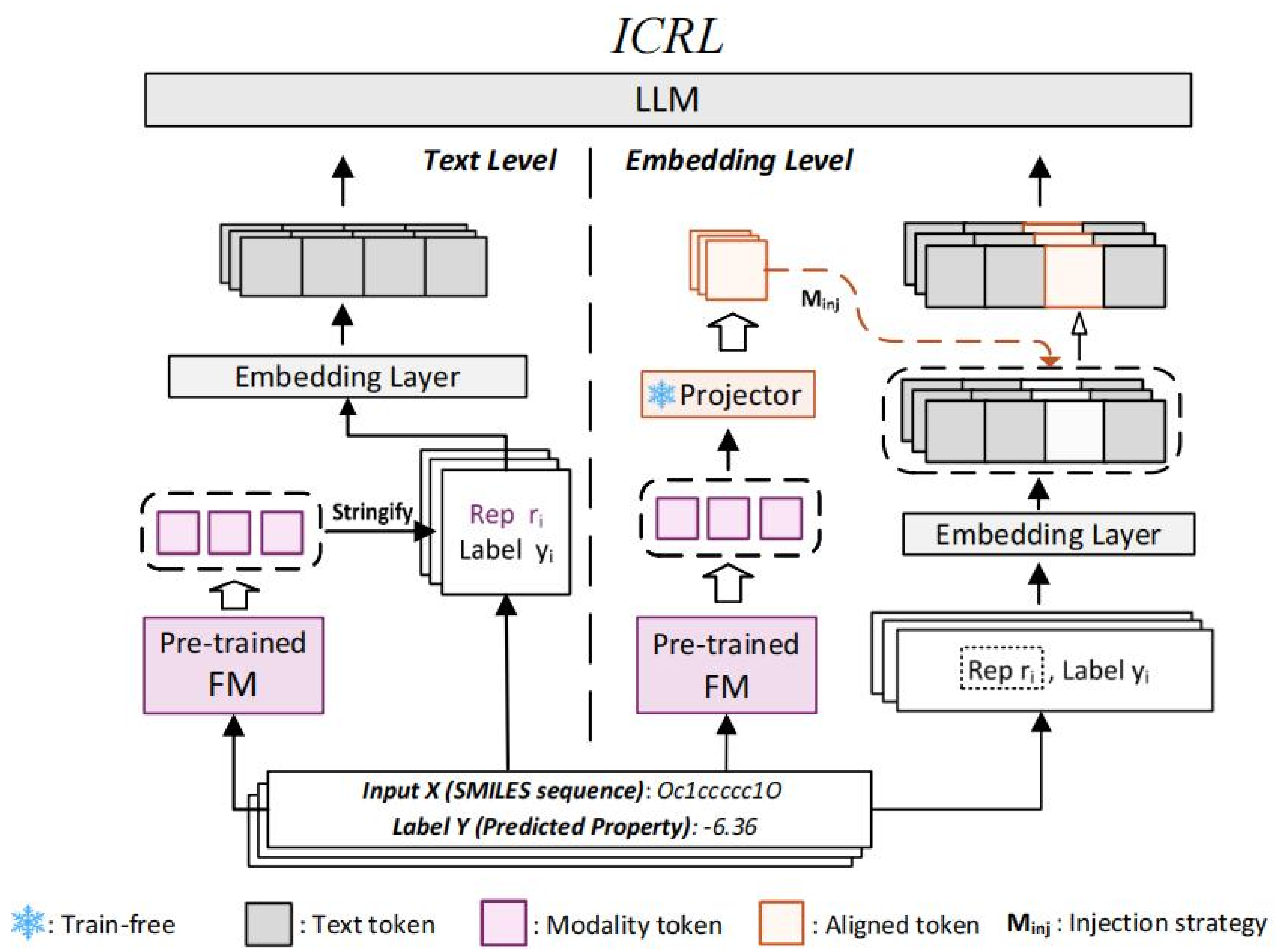


Two Levels of Representation Injection in ICRL

- **Text-Level**: FM feature → PCA (dim. reduction) → input as text → few-shot example → reasoning

Question: What is the Solubility of the drug molecule?
Molecular vector representation: ***[4.26, -6.16, ..., 1.32]***
Answer: -0.258 ***Interpretable but context-inefficient.***

ICRL: In-Context Representation Learning



Two Levels of Representation Injection in ICRL

- **Text-Level**: FM feature → PCA (dim. reduction) → input as text → few-shot example → reasoning

Question: What is the Solubility of the drug molecule?
*Molecular vector representation: **[4.26, -6.16, ..., 1.32]***
Answer: -0.258 ***Interpretable but context-inefficient.***

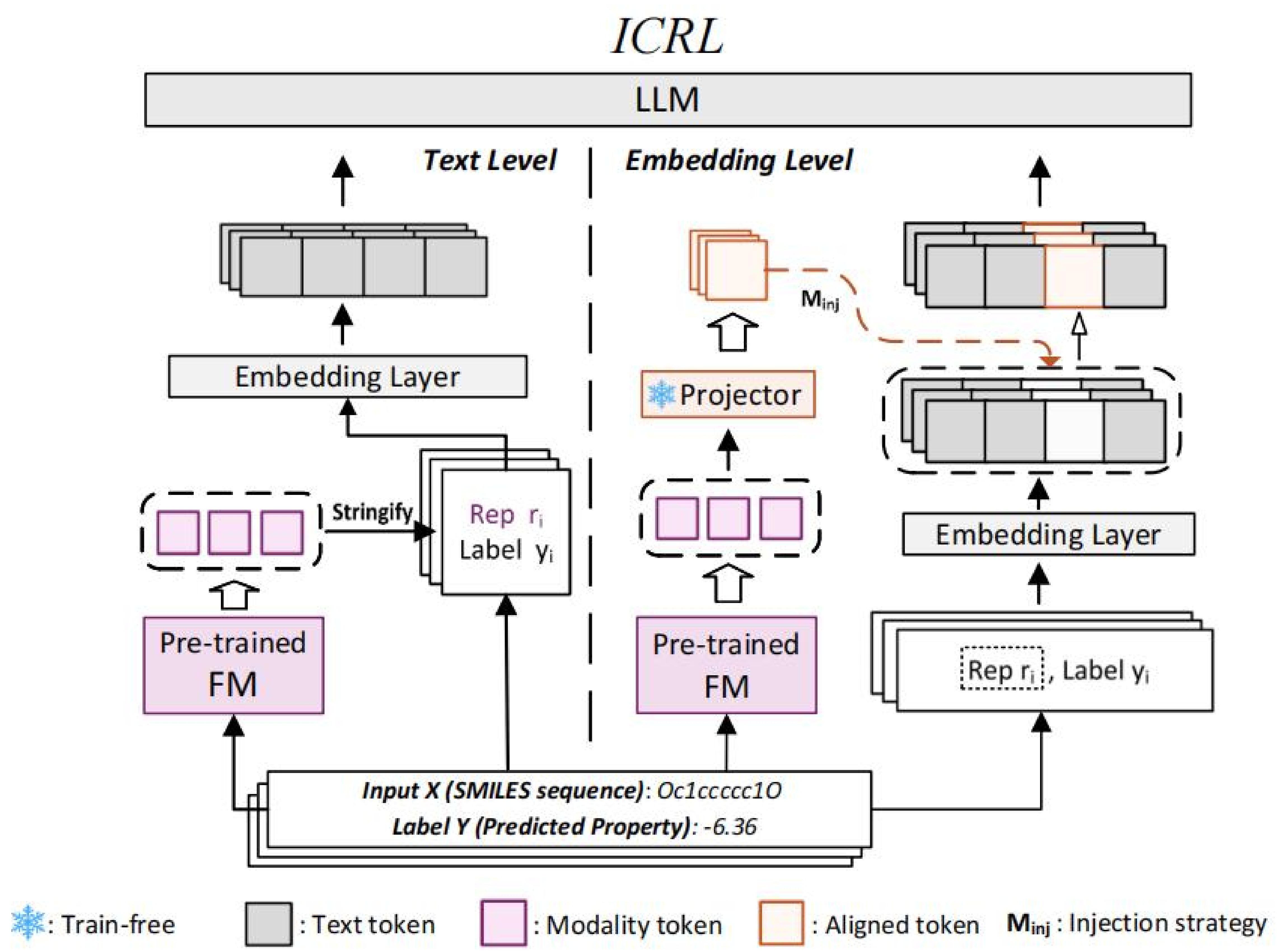
- **Embedding-Level**: FM feature → random projector → Optimal Transport (OT) alignment → input as aligned token → few-shot example → reasoning

Question: What is the Solubility of the drug molecule?
*Molecular vector representation: **[REP]492[/REP]***
Answer: -0.258 ***Requires alignment but token-efficient.***

Candidate Injection Strategies

- Zero-Pad → pad FM features to match LLM dimension.
- Random Projection → map with untrained linear layer.
- OT Alignment → Align the distribution of FM embeddings with the LLM embedding space via OT, using the token embeddings of SMILES text (**OT-Embed**) or PCA strings (**OT-PCA**) as the target distribution.

ICRL: In-Context Representation Learning



Candidate Injection Strategies

- Zero-Pad → pad FM features to match LLM dimension.
- Random Projection → map with untrained linear layer.
- OT Alignment → Align the distribution of FM embeddings with the LLM embedding space via OT, using the token embeddings of SMILES text (**OT-Embed**) or PCA strings (**OT-PCA**) as the target distribution.

Two Levels of Representation Injection in ICRL

- **Text-Level**: FM feature → PCA (dim. reduction) → input as text → few-shot example → reasoning

Question: What is the Solubility of the drug molecule?
Molecular vector representation: ***[4.26, -6.16, ..., 1.32]***
Answer: -0.258 ***Interpretable but context-inefficient.***

- **Embedding-Level**: FM feature → random projector → Optimal Transport (OT) alignment → input as aligned token → few-shot example → reasoning

Question: What is the Solubility of the drug molecule?
Molecular vector representation: ***[REP]492[/REP]***
Answer: -0.258 ***Requires alignment but token-efficient.***

Non-Trivial Results with Representations Only

- Text-level (PCA): Achieves performance **comparable to, or better than**, ICL.
- Embedding-level (OT): Compresses each FM feature into just **1 token**, drastically reducing context usage while significantly improving over naive methods.

Lightweight Yet Powerful: How ICRL Extends ICL

ICRL Boosts ICL with Text Features

Dataset	Baseline	ICRL (Ours)					
	Text	Text	Embedding				
	ICL	PCA+ICL	Zero-Pad+ICL	Ran-Noi+ICL	Ran-Pro+ICL	OT-Embed+ICL	OT-PCA+ICL
ESOL	0.465 ±9.2e-4	0.455 ±1.2e-4	0.526 ±2.1e-4	0.540 ±1.6e-3	0.525 ±6.5e-5	0.508 ±1.7e-4	**0.542** ±5.4e-4
Caco2_Wang	0.411 ±1.3e-3	0.393 ±9.2e-4	0.410 ±4.6e-6	0.420 ±1.1e-4	0.405 ±1.6e-5	**0.429** ±1.1e-3	0.394 ±5.7e-4
AqSolDB	0.596 ±5.1e-5	0.549 ±3.2e-4	**0.606** ±6.8e-6	0.597 ±1.1e-5	0.600 ±2.4e-5	0.569 ±5.7e-4	0.589 ±3.9e-5
LD50_Zhu	0.378 ±1.2e-5	0.356 ±1.9e-4	**0.393** ±8.6e-6	0.379 ±5.4e-6	0.392 ±7.3e-5	0.361 ±1.2e-5	0.362 ±7.8e-5
AstraZeneca	0.266 ±2.3e-5	0.227 ±3.1e-5	**0.272** ±4.8e-5	0.267 ±2.1e-5	0.269 ±1.9e-5	0.269 ±2.1e-4	0.271 ±6.6e-5

- **Overall gain:**
  Combining both **consistently improves** performance (e.g., OT-PCA achieves +16.6% on ESOL over text-only ICL.).
- **Counterintuitive results**:
  With text features, even random noise outperforms the baseline, while zero-padding performs better in most cases.

Lightweight Yet Powerful: How ICRL Extends ICL

ICRL Boosts ICL with Text Features

Dataset	Baseline	ICRL (Ours)					
	Text	Text	Embedding				
	ICL	PCA+ICL	Zero-Pad+ICL	Ran-Noi+ICL	Ran-Pro+ICL	OT-Embed+ICL	OT-PCA+ICL
ESOL	0.465 ±9.2e-4	0.455 ±1.2e-4	0.526 ±2.1e-4	0.540 ±1.6e-3	0.525 ±6.5e-5	0.508 ±1.7e-4	**0.542** ±5.4e-4
Caco2_Wang	0.411 ±1.3e-3	0.393 ±9.2e-4	0.410 ±4.6e-6	0.420 ±1.1e-4	0.405 ±1.6e-5	**0.429** ±1.1e-3	0.394 ±5.7e-4
AqSolDB	0.596 ±5.1e-5	0.549 ±3.2e-4	**0.606** ±6.8e-6	0.597 ±1.1e-5	0.600 ±2.4e-5	0.569 ±5.7e-4	0.589 ±3.9e-5
LD50_Zhu	0.378 ±1.2e-5	0.356 ±1.9e-4	**0.393** ±8.6e-6	0.379 ±5.4e-6	0.392 ±7.3e-5	0.361 ±1.2e-5	0.362 ±7.8e-5
AstraZeneca	0.266 ±2.3e-5	0.227 ±3.1e-5	**0.272** ±4.8e-5	0.267 ±2.1e-5	0.269 ±1.9e-5	0.269 ±2.1e-4	0.271 ±6.6e-5

- **Overall gain:** Combining both **consistently improves** performance (e.g., OT-PCA achieves +16.6% on ESOL over text-only ICL.).
- **Counterintuitive results**: With text features, even random noise outperforms the baseline, while zero-padding performs better in most cases.

ICRL vs. Costly Training

Method	Type	Resource	Training Time	ESOL (RMSE)	Lipo (RMSE)	Avg
MolecularGPT [36]	I-FT	4×A800-80G	<1 day	1.471	1.157	1.314
GIMLET [67]	S-PT + FT	2–4 GPUs	∼1 day	1.132	1.345	1.239
SELFormer [64]	PT	2×A5000	∼2 weeks	1.357	3.192	2.275
	PT + FT	2×A5000	∼2 weeks	0.682	1.005	0.844
GPT-MolBERTa [5]	PT + FT	2–4 GPUs	∼2 weeks	0.477±0.01	0.758±0.01	0.612
OT-PCA (ours)	Training-free	CPU only	∼2 sec	1.140±0.01	1.349±0.01	1.245
OT-PCA + ICL (ours)	Training-free	CPU only	∼2 sec	1.094±0.01	1.277±0.01	1.186

- **Better performance–cost trade-off:** While falling short of full PT+FT performance, ICRL delivers **comparable or even superior** results to most lightweight training methods, with only a **~2 s CPU-based** alignment step required.

Value of the Study

- **Training-Free Multimodal Reasoning**
  -- Introduces a framework that enables text-only LLMs to reason over non-text representations **without** any retraining.

- **Cross-Modality Generalization**
  -- Demonstrates that even **frozen** LLMs can **generalize** across modalities through contextual reasoning alone, revealing their latent representational flexibility.

- **Scalable and Efficient Integration**
  -- Provides a **lightweight, CPU-based** approach to multimodal alignment that completes within seconds, making it practical for resource-limited or retraining-impractical domains.

Thank You!

Code: https://github.com/ztlmememe/LLMxFM_ICRL

Paper Link: https://arxiv.org/abs/2509.17552