





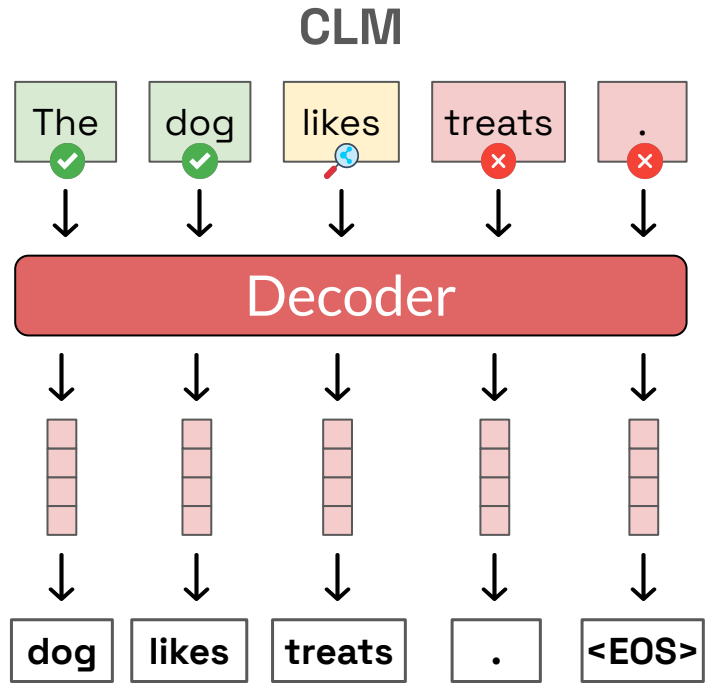
BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs

Nicolas Boizard, Théo Deschamps-Berger
Hippolyte Gisserot-Boukhlef

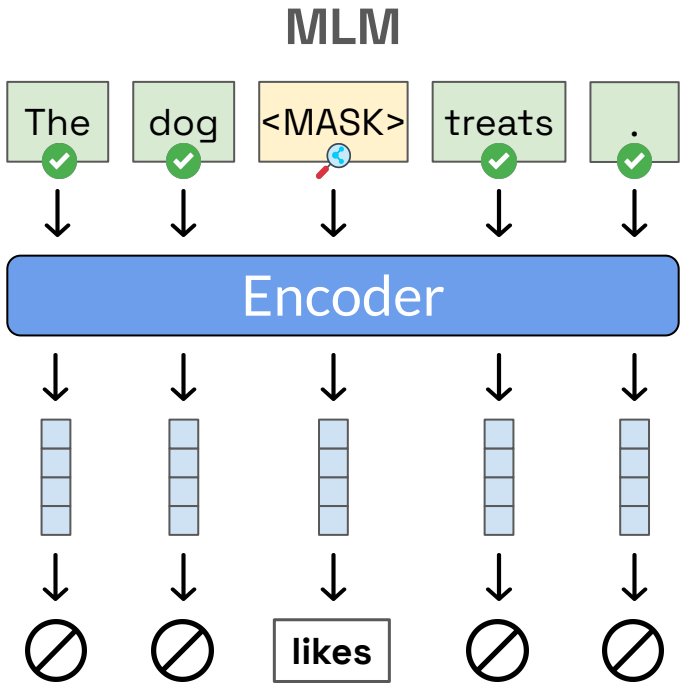
PRESENTATION AT EBAY - MAY 7, 2026

Encoders vs. Generative LLMs

In the modern era of NLP, two primary types of models dominate: encoder- and decoder-only generative models.



$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$



$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}})$$

Why Encoder Models?



With the recent explosion of generative AI models, researchers tend to forget encoders are still by far the most widely used deep learning text models in the industry!



Versatility
Encoder representations can be used for tasks like classification, retrieval, NER, extractive Q&A...



Efficiency
Encoder models are inherently more compact than decoder models, making them significantly faster for both fine-tuning and inference.

BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs





Nicolas Boizard



Théo Deschamps-Berger



Hippolyte Gisserot-B.

Question:

- ➔ What drives adaptation quality?
- ➔ Can adaptation scale without the original pre-training data?
- ➔ Can resulting encoders compose with the causal ecosystem?

arXiv

CentraleSupélec

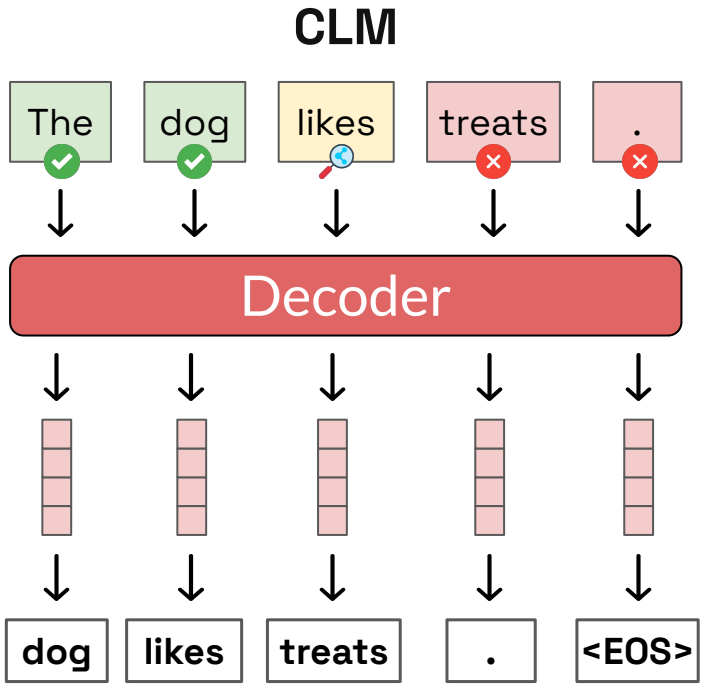
université PARIS-SACLAY

diabolocom

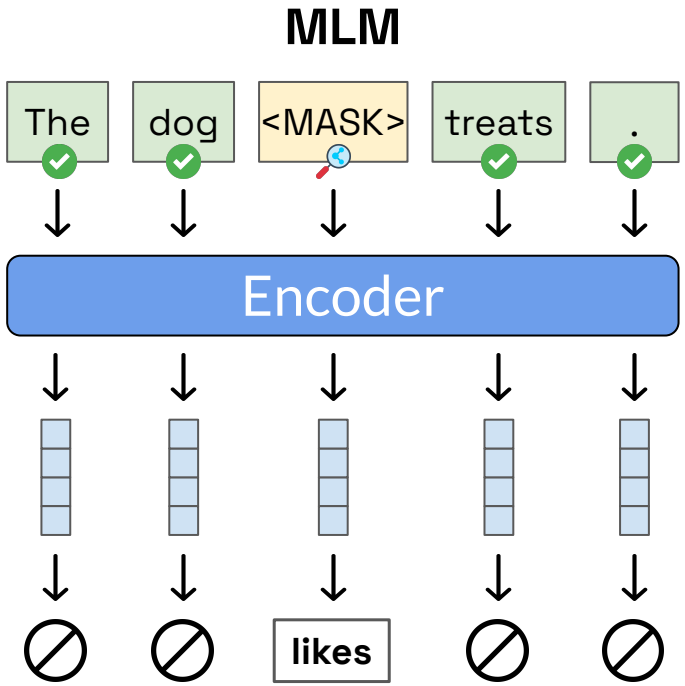
ARTEFACT RESEARCH CENTER

What drives adaptation quality?

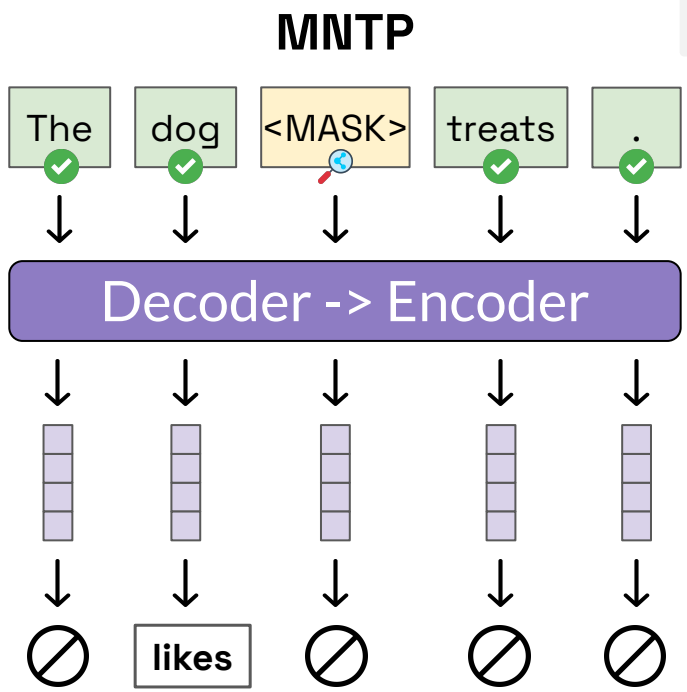
Many training objectives exist. Which one should I choose to adapt a decoder into an encoder?



$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log P(x_t \mid x_{<t})$$



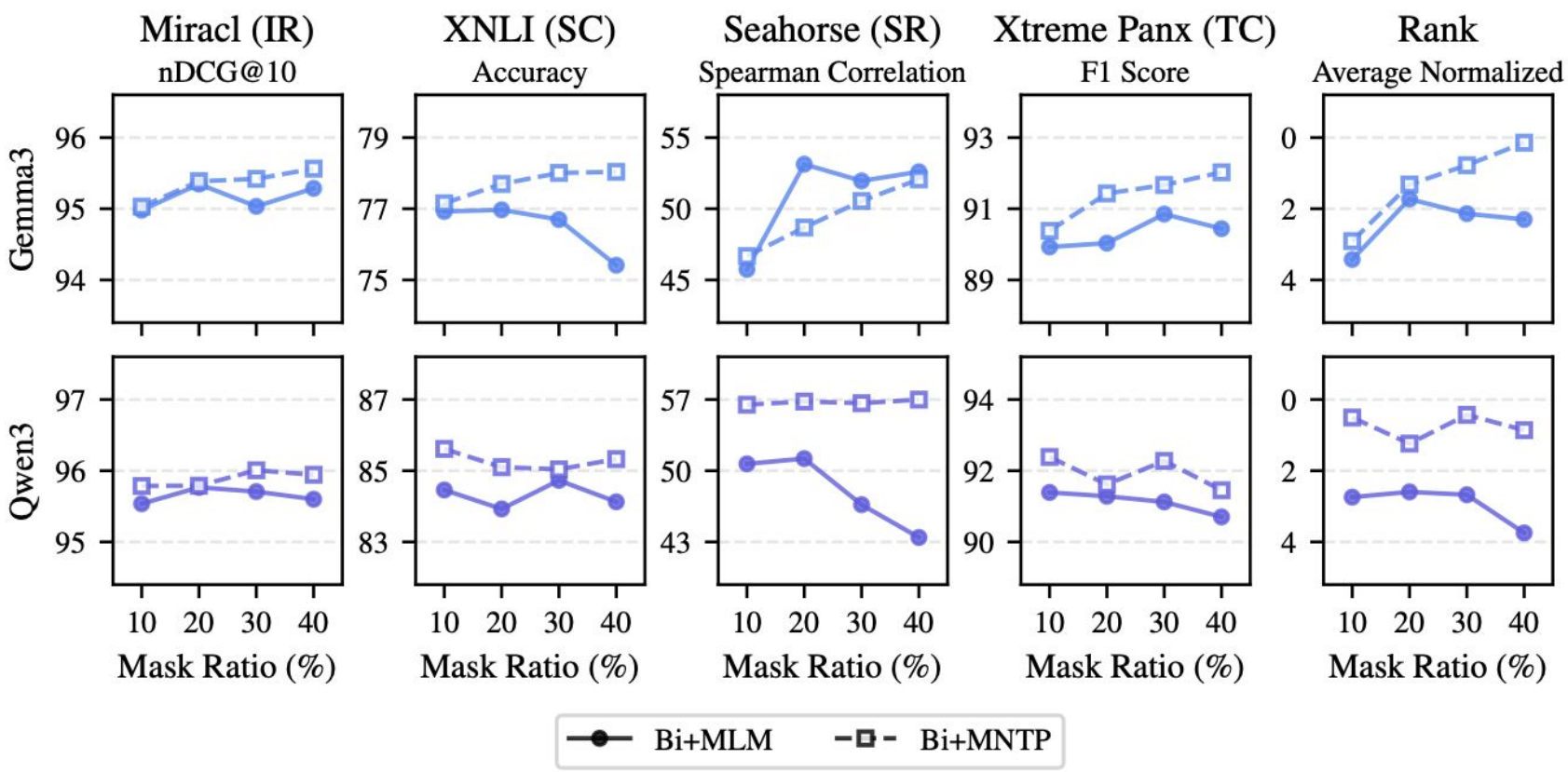
$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i \mid x_{\setminus \mathcal{M}})$$



$$\mathcal{L}_{\text{MNTP}}(\mathbf{x}) = -\sum_{i \in \mathcal{M}} \log p_{\theta, i-1}(x_i \mid \mathbf{x}_{\mathcal{M}})$$

What drives adaptation quality?

Existing methods conflate critical design choices like training objectives and attention mechanisms, leaving no consensus on what drives quality



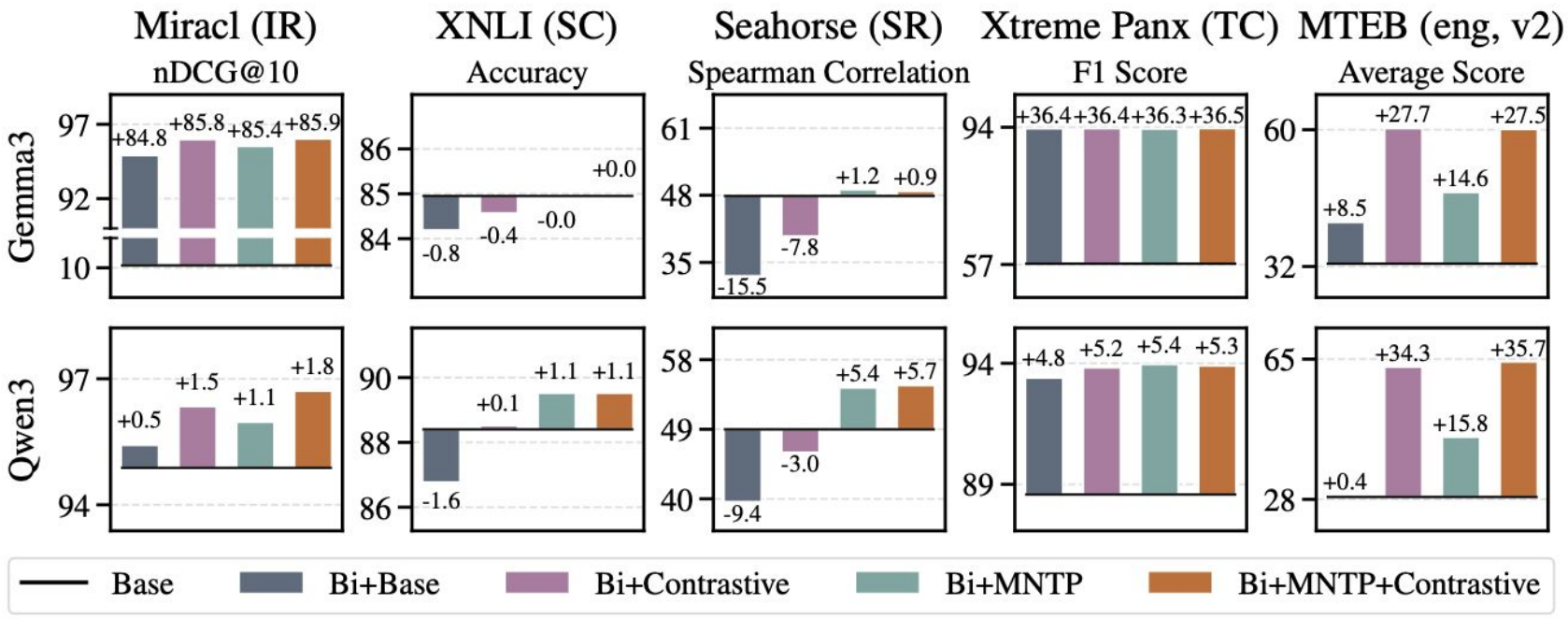
MLM vs MNTP

1. MNTP > MLM
2. No Optimal Ratio: depends on objective, model and task

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

What drives adaptation quality?

Recent work mostly skips the masking phase and runs contrastive training alone, we disentangle these choices here



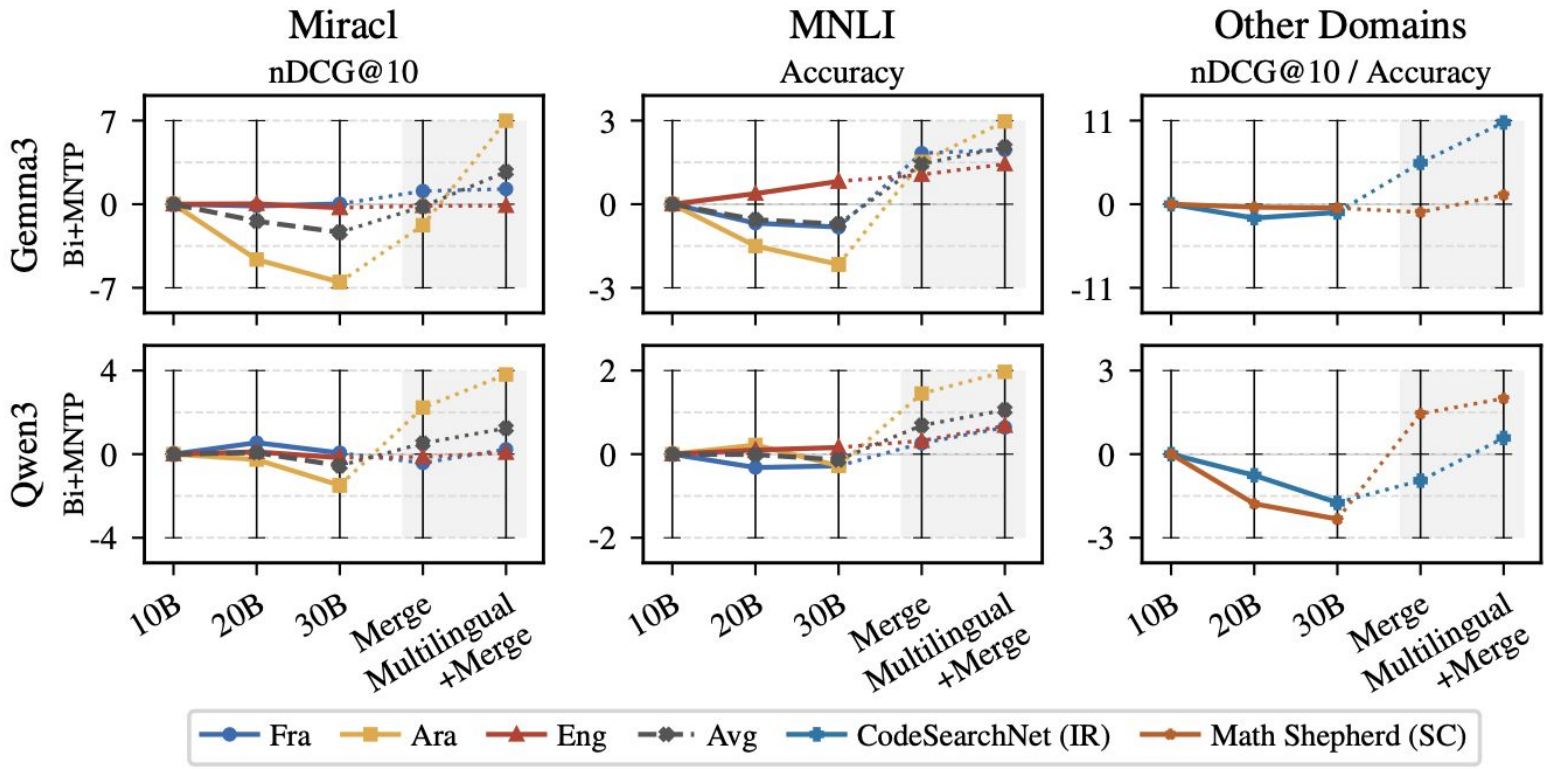
Contrastive vs. MNTP

1. Contrastive → General Embedding
2. MNTP → Tasks specific
3. Two-phase training (MNTP x Contrastive) gets the best of both

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

Can adaptation scale without the original pre-training data?

Question: How to scale the adaptation process while aiming to preserve the foundational knowledge of the base models?



Weight merging + light multi-domain mix

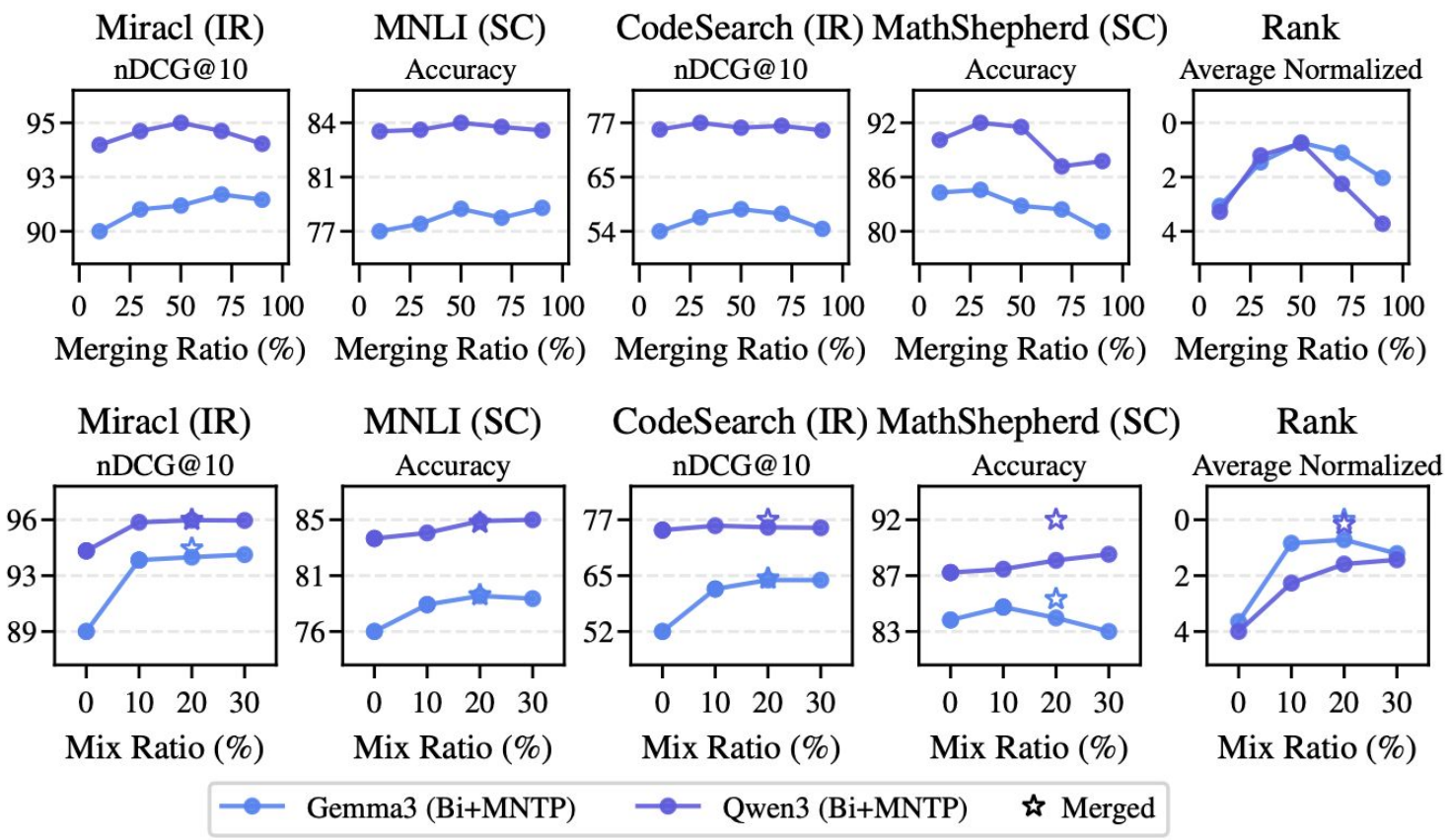
1. Weight merging preserves the base model's knowledge and newly acquired bidirectional capabilities.

2. A light multi-domain mix is sufficient to maximize knowledge preservation.

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

Can adaptation scale without the original pre-training data?

Question: How much of the base model should we keep and with what data mix?



Why merging works

Adapted and base models stay close in weight space: Cosine similarity: **0.78** (Gemma) / **0.97** (Qwen)

Best performance recipe

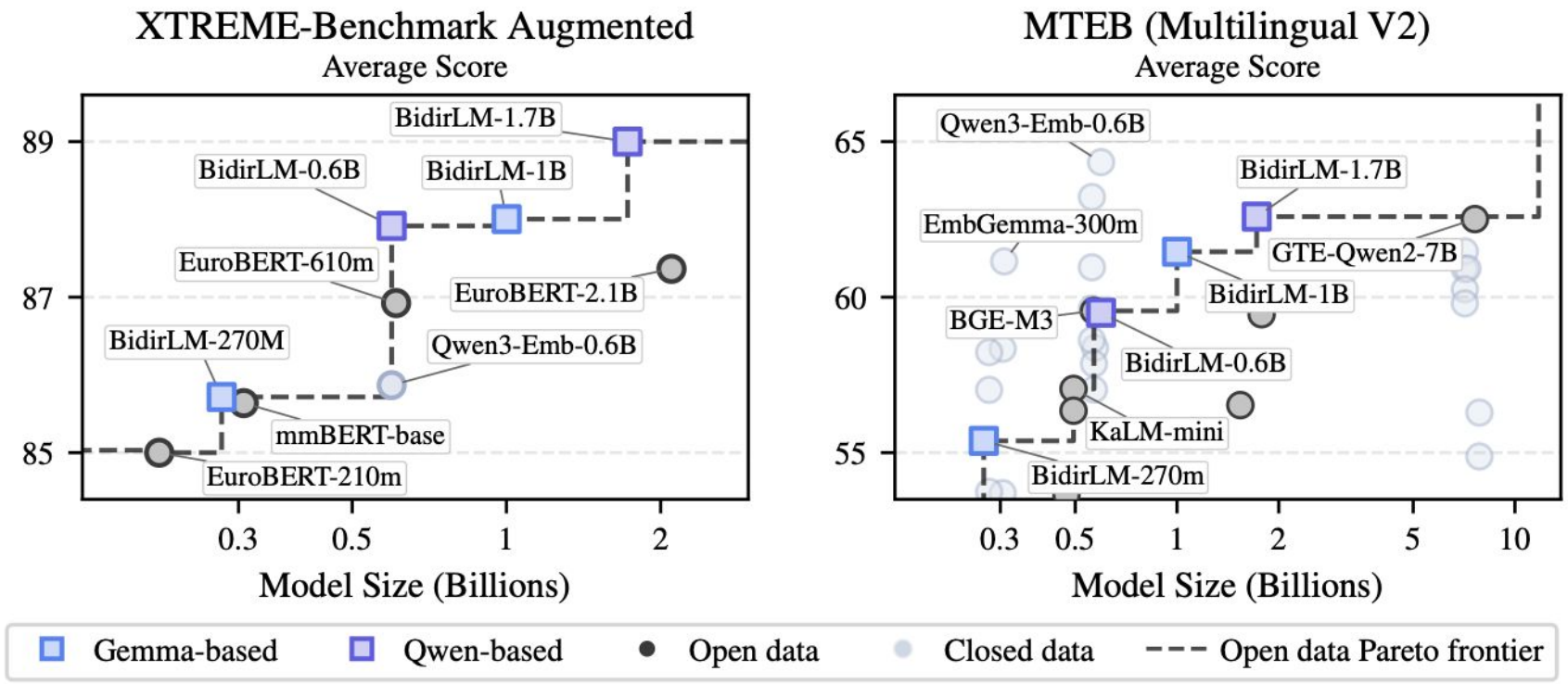
1. Data mix: 80% English / 20% split evenly across multilingual, math, code

2. Merging ratio: 50%

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

Can adaptation scale without the original pre-training data?

Question: How do we compare against other works?



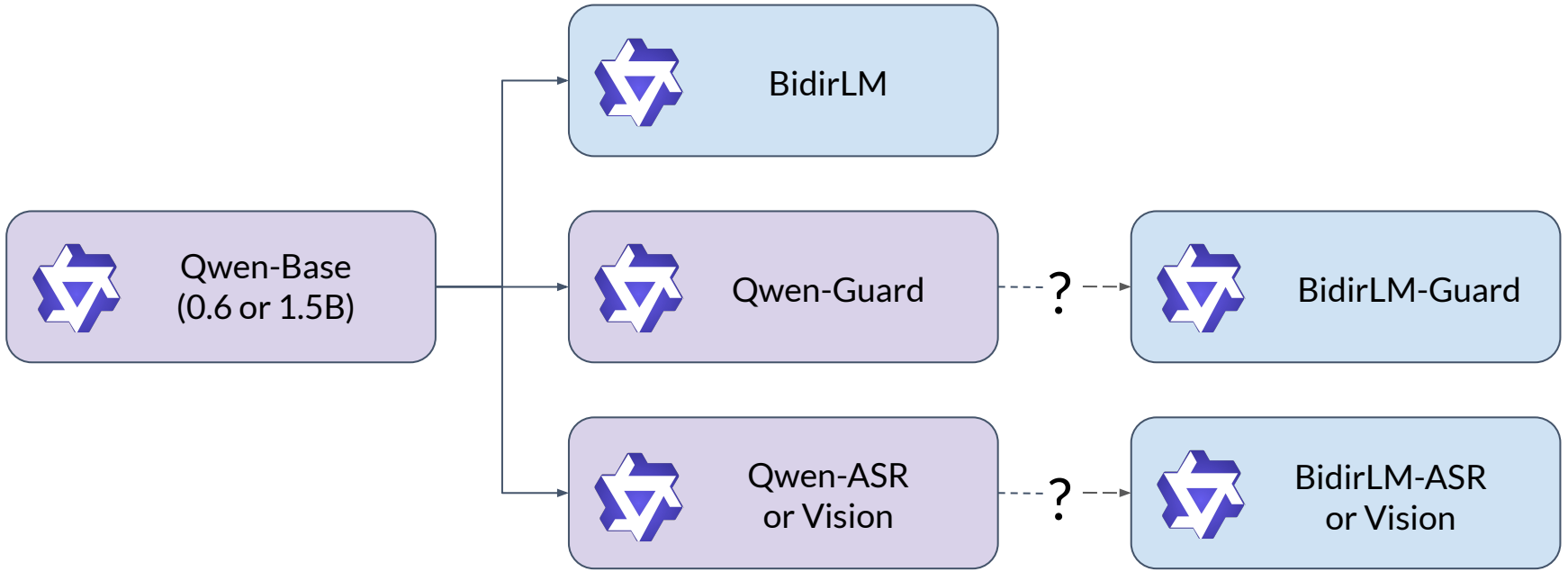


#1 on the open-data Pareto frontier

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

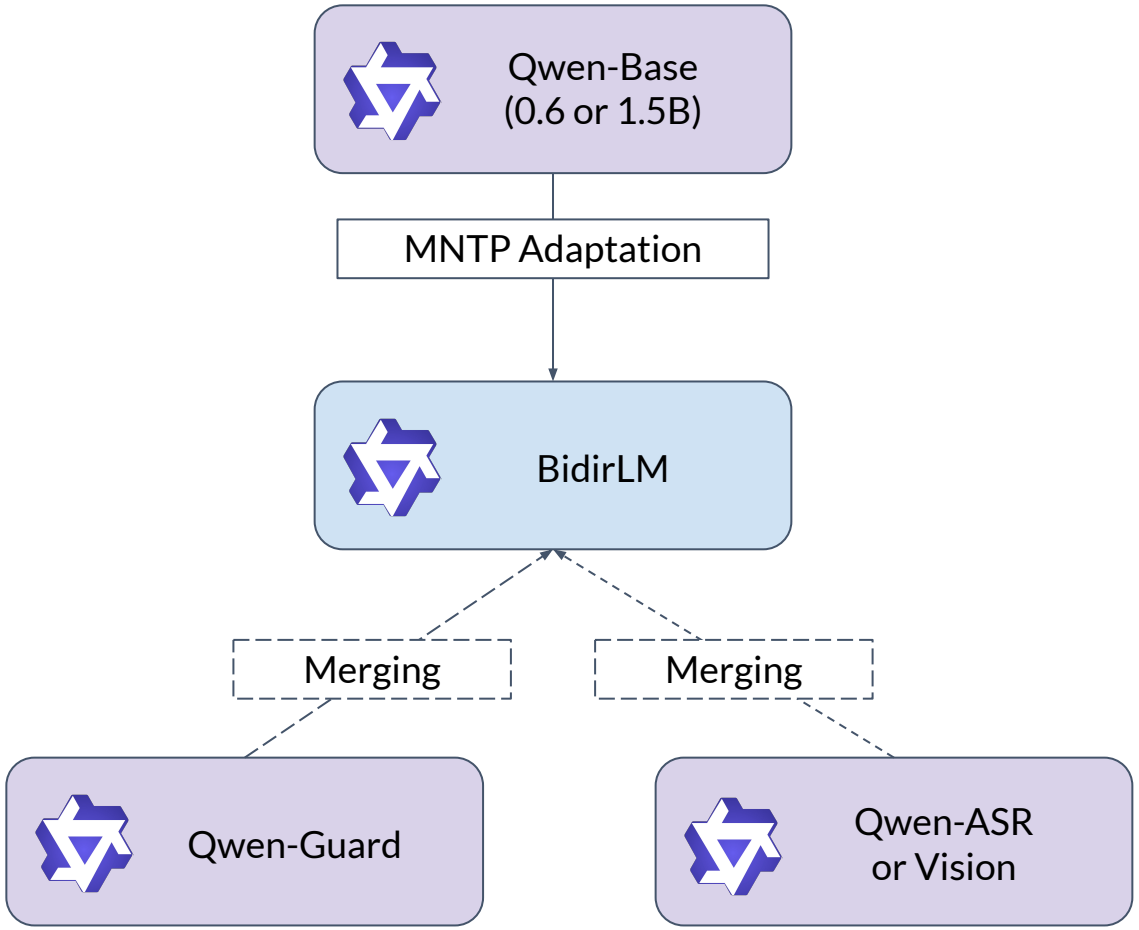
Integration with the Causal Ecosystem.

Question: Current adapted encoders rely on rigid pipelines that fail to compose with specialized causal models, can we transfer knowledge / capabilities from fine tuned models ?



N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

Efficiently Composing with Causal Ecosystem via Merging





Minimal Cost

Adaptation training is performed only once.

Adapt as You Go

Continually add new knowledge and modalities / keeping up with the latest fine-tuned models.

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

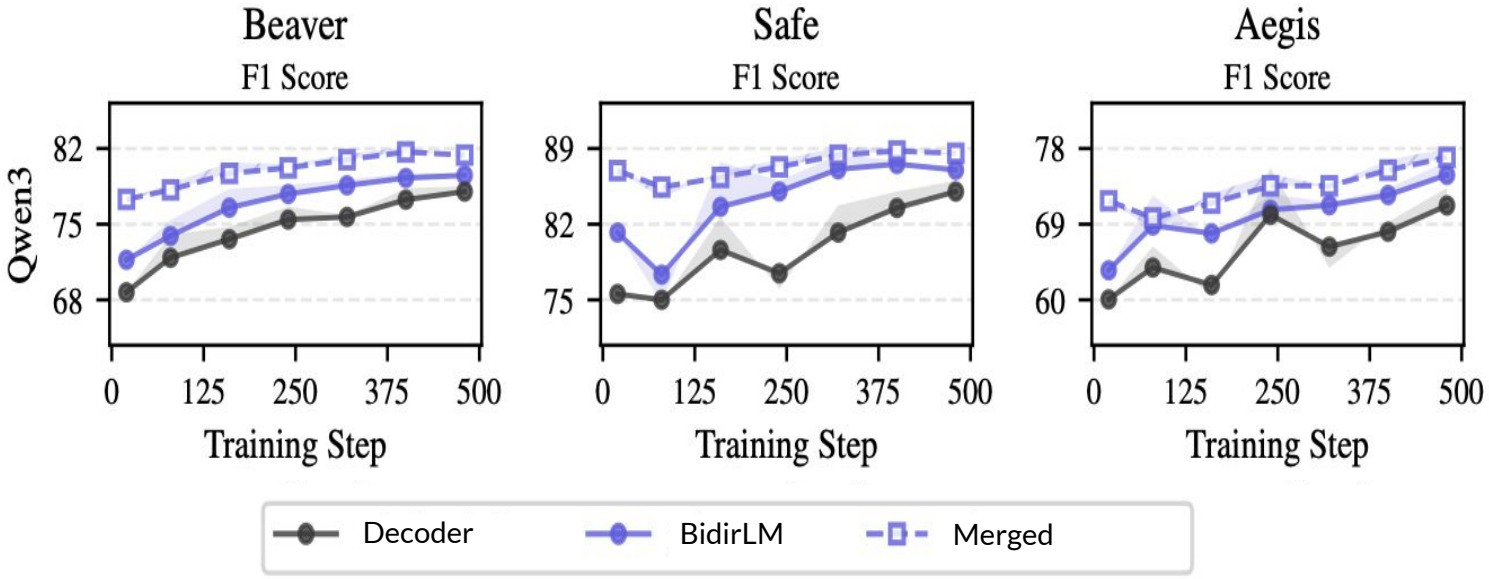
Domain transfer (Prompt Moderation)

Decoder:
- Qwen-Guard-0.6B

BidirLM:
- BidirLM (Qwen-0.6B)

Merged:
- 50% → BidirLM
- 50% → Qwen-Guard

Finetuning: 1 MI250x, 2min on Beaver Train Set



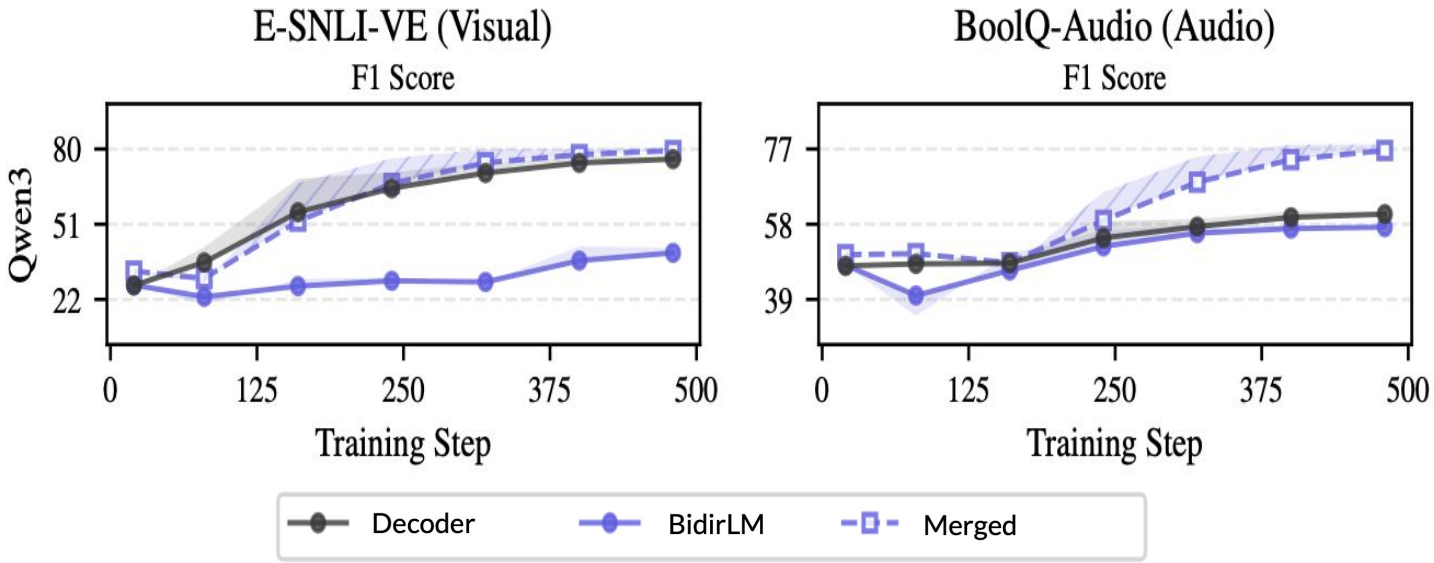
N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

Modality transfer

Encoders:
- BidirLM (Qwen based)

Specialists:
- Qwen-Vision
- Qwen-ASR (Audio only)

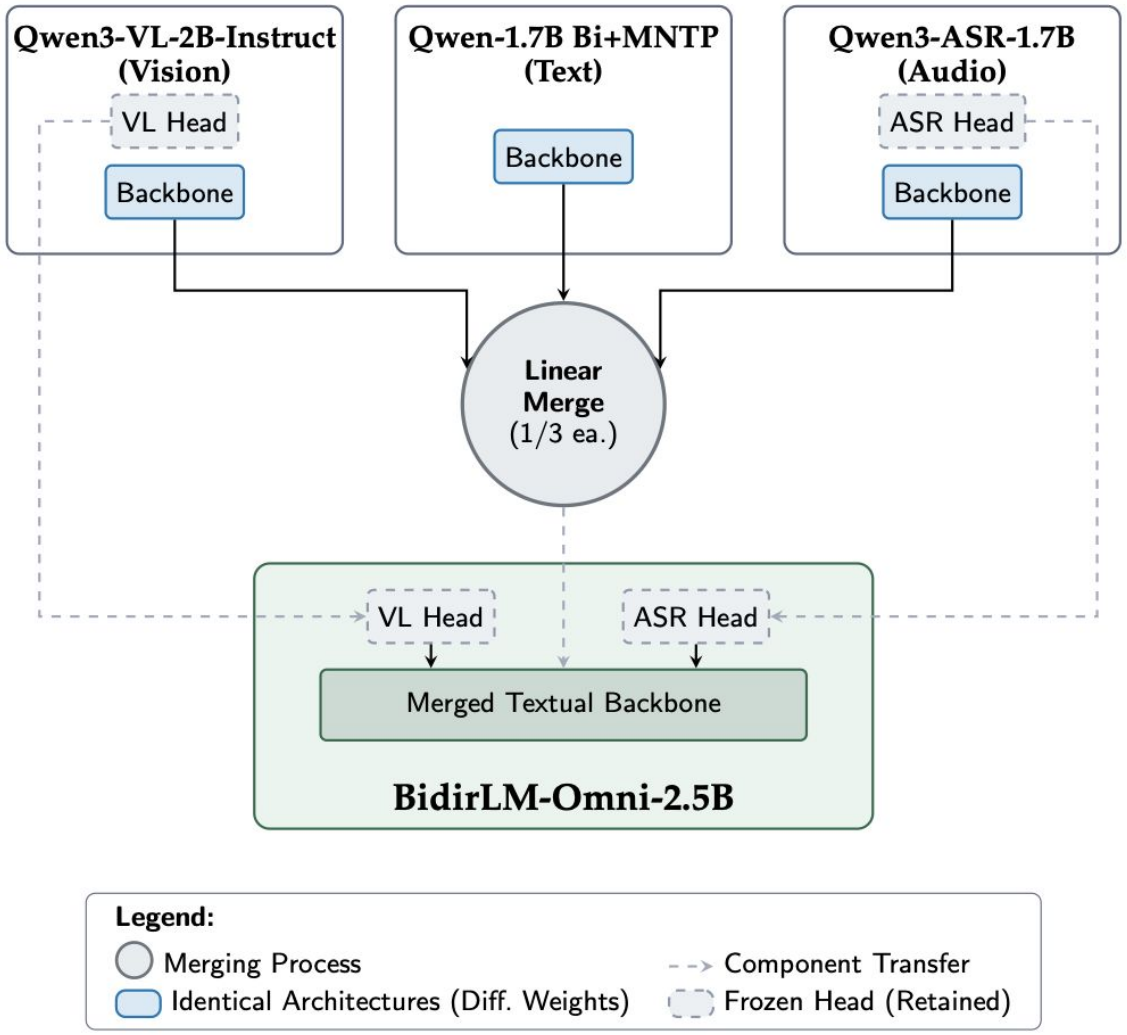
Merging: 50/50
Modality Head: Frozen
Finetuning: 1 MI250x, 2min
Attention: All bidirectional



Weight merging enables the efficient composition of domain and modality causal model specialists into an adapted encoder, **succeeding even in the absence of shared prior modalities.**

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

BidirLM-Omni-2.5B



Training

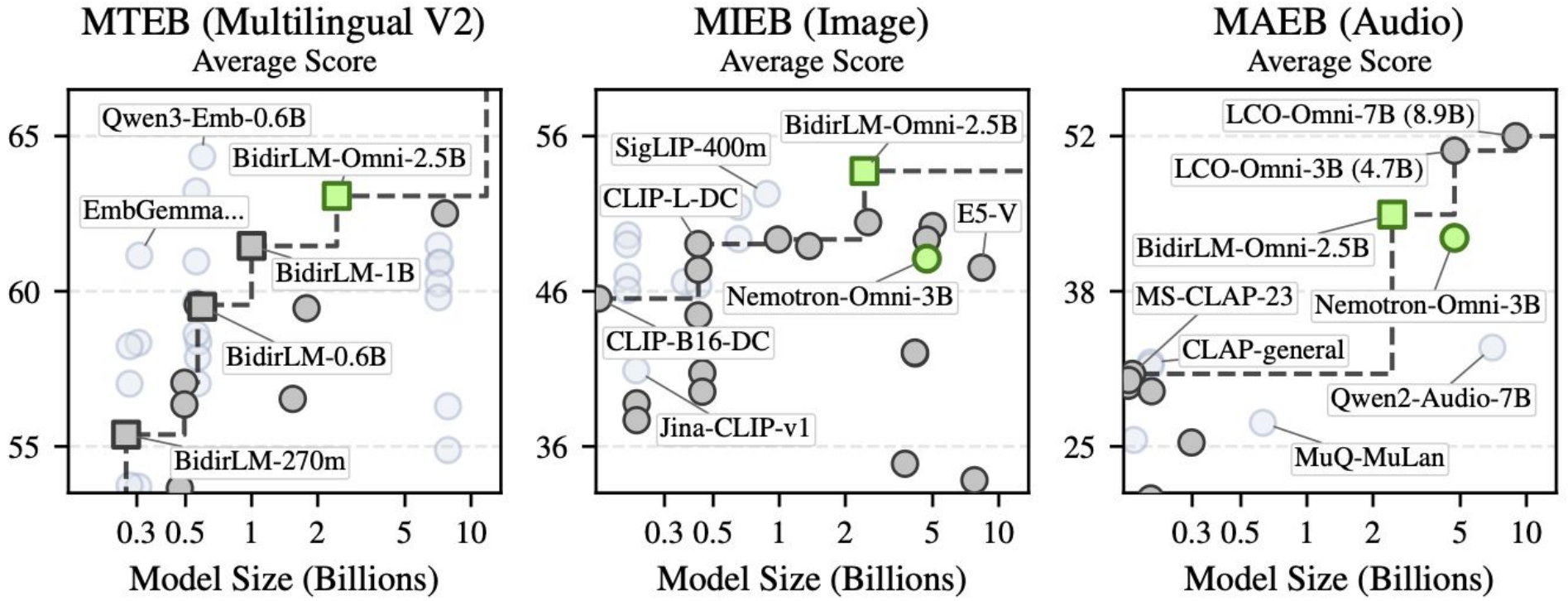
1.8M-pair contrastive corpus mixing:

- **65% text-text**
  - 1.2M multi-domain corpus with hard negatives
- **17.5% audio-text**
  - Laion-Audio-300M (200K, audio-description) and LibriSpeech ASR (100K, speech-transcription)
- **17.5% image-text**
  - Colpali (100K, document-query), NatCap, and MSCOCO (100K each, image-description)

Loss: InfoNCE
Merging: 1/3
Modality Head: Frozen
Budget: 8 MI250x (total 250 hours)

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

BidirLM-Omni-2.5B



MTEB

Rank (Bor…	Model	ens	Mean (T…	Mean (TaskT…	Bitext …	Classification	Clustering	Instruction R…	Multilabel Class…	Pair Classificat…	Reranking	Retrieval	STS
15	BidirLM-Omni-2.5B-Embedding		63.50	55.48	72.29	65.90	51.50	0.97	26.54	80.65	64.72	61.04	75.73

MAEB

Rank (Bor…	Model	ng D…	Max Tokens	Mean (T…	Mean (TaskT…	Any Any Re…	Audio Classifica…	Audio Clust…	Audio Multilabel…	Audio Pair Class…	Audio Rera…	Audio Zeroshot C…
2	BidirLM-Omni-2.5B-Embedding		32768	51.87	52.42	47.47	**59.77**	6.18	46.23	66.75	74.76	**65.77**

MIEB

Rank	Model	Mean (T…	Multilingual R…	Retrieval	Compositionality	Document Unders…	Image Classifica…	Image Clust…	Vision C…	Visual …	Visual …	Zero Shot Classi
2	BidirLM-Omni-2.5B-Embedding	55.76	56.97	29.31	46.00	65.11	73.75	66.18	55.39	67.87	45.70	51.28

N. Boizard et al., "BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs"

Thank you for your attention.
Time for questions!



BidirLM: From Text to Omnimodal Bidirectional Encoders by Adapting and Composing Causal LLMs

N.Boizard., et al.

arXiv

CentraleSupélec

université PARIS-SACLAY

diabolocom

cohere









Thank you for your attention!