



01 Background: calibration must transfer

Preference-aligned LLMs are often overconfident. Post-hoc temperature scaling (TS) leaves the model frozen and is domain-dependent.

Llama-3.1 TS

conf-ECE ↓ · Manuscript Tables 1–2

0.0224 ID → 0.2513 OOD

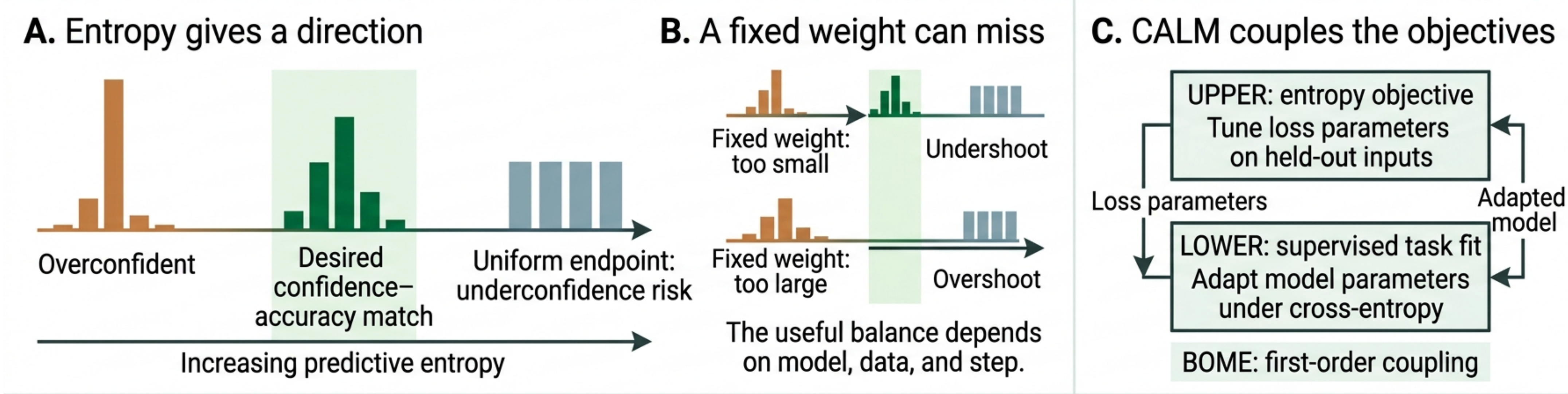
Calibration asks whether predicted confidence matches empirical correctness.

02 Motivation: why maximize entropy?

Entropy is a direction, not the destination.

In the motivating picture, an overconfident network is too sharp; the uniform, maximum-entropy endpoint is the opposite extreme. Moving between them can bring confidence closer to accuracy—or overshoot into underconfidence.

CALM: Bilevel Optimization for LLM Calibration



Entropy supplies the direction; bilevel coupling regulates the task trade-off.

Conceptual crossing intuition, not an empirical trajectory or a guarantee of zero calibration error.

03 Why bilevel instead of a fixed penalty?

$$\mathcal{L}_{\text{single}} = \mathcal{L}_{\text{tr}} - \beta H(f), \quad \beta \text{ fixed}$$

A fixed entropy weight can **undershoot** (residual overconfidence) or **overshoot** (underconfidence). The useful balance depends on the **model, data, and training step**. A fixed penalty does not explicitly identify where confidence matches accuracy.

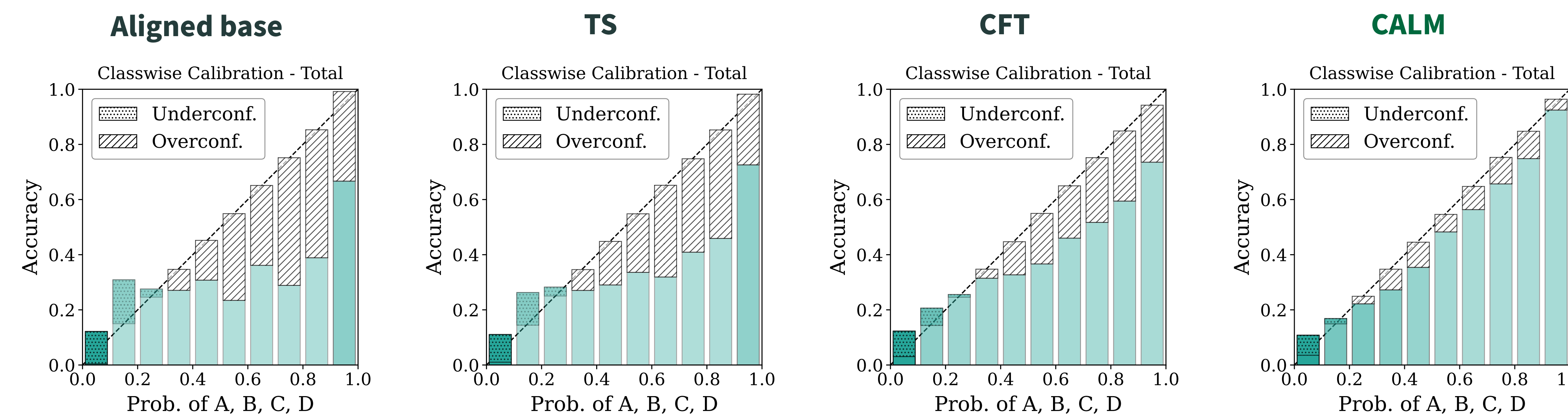
How far should optimization move along the entropy direction?

Bilevel loss design couples calibration pressure to task fitting.

The contribution is this constrained interaction; BOME is the enabling optimizer.

04 Evidence: reliability under domain shift

Mistral-7B, OOD. Bars below the diagonal indicate overconfidence (Fig. 1).



CALM moves the reliability profile closer to the diagonal.

05 CALM: calibration-oriented loss design

Learn the loss parameters and the model together.

$$f_{\alpha, \theta}(y \mid x) = \text{softmax}(d \odot z_{\theta}(x) + l)_y$$

$\alpha = (d, l)$: per-vocabulary scales and shifts; scalar TS ties $d_y = 1/T, l = 0$.

UPPER LEVEL

Calibration-oriented objective

$$\min_{\alpha} -\mathbb{E}_{x \in S_V} H[f_{\alpha, \theta(\alpha)}(\cdot \mid x)]$$

Held-out inputs; no additional calibration labels.

LOWER LEVEL

Supervised task fitting

$$\theta(\alpha) \in \arg \min_{\theta'} -\mathbb{E}_{(x, y) \in S_T} \log f_{\alpha, \theta'}(y \mid x)$$

Cross-entropy anchors the model to task discrimination.

Manuscript Eqs. (8)–(9): **upper** = entropy; **lower** = task loss.

06 BOME: the first-order enabler

Exact hypergradients are expensive at LLM scale. A BOME-style update uses **only first-order gradients**, avoiding Hessian inverses and differentiation through the full inner optimization.

$$q(\alpha, \theta) = \mathcal{L}_{\text{tr}}(\theta; \alpha) - \min_{\theta'} \mathcal{L}_{\text{tr}}(\theta'; \alpha)$$
$$d_k = \underbrace{g_{\text{ent}}}_{\text{entropy direction}} + \underbrace{\lambda_k g_q}_{\text{adaptive task correction}}$$

20 inner steps → surrogate residual → joint update

~2 hours / ~38 GB / one A100 (40 GB)

07 Evaluation: two kinds of transfer

Training → evaluation	What shifts?	Metric ↓
Alpaca → MCQA	Domain + format	conf-/cw-ECE
PopQA → TriviaQA	Domain only	Sem-ECE

4 aligned LLMs: Llama-3.1, Vicuna-7B, OLMo-2-7B, Mistral-7B.

Free-form protocol: 20 samples/question; semantic-cluster frequency is confidence (300 evaluation questions).

08 Ablations: coupling and vectorization

Mistral-7B, OOD. Calibration and accuracy must be read together.

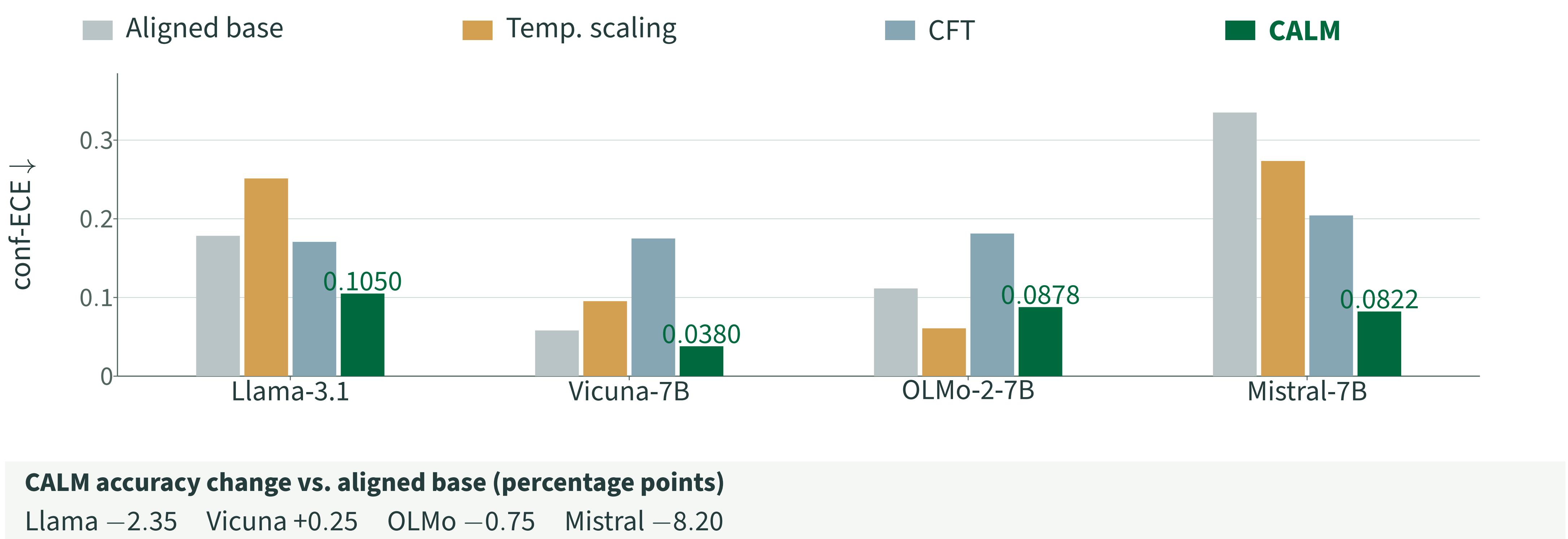
Method / removed component	conf-ECE ↓	Accuracy ↑
Aligned base	.3351	59.8%
Regularization / no bilevel constraint	.2100	44.2%
Iterate / no residual coupling	.0923	25.8%
Scalar CALM / no vector parameters	.3140	59.9%
CALM / full formulation	.0822	51.6%

Low ECE alone is not success.
Coupling avoids the severe accuracy collapse of Iterate.

Tables 1, 12. Regularization and Iterate are **ablations**, not external baselines.

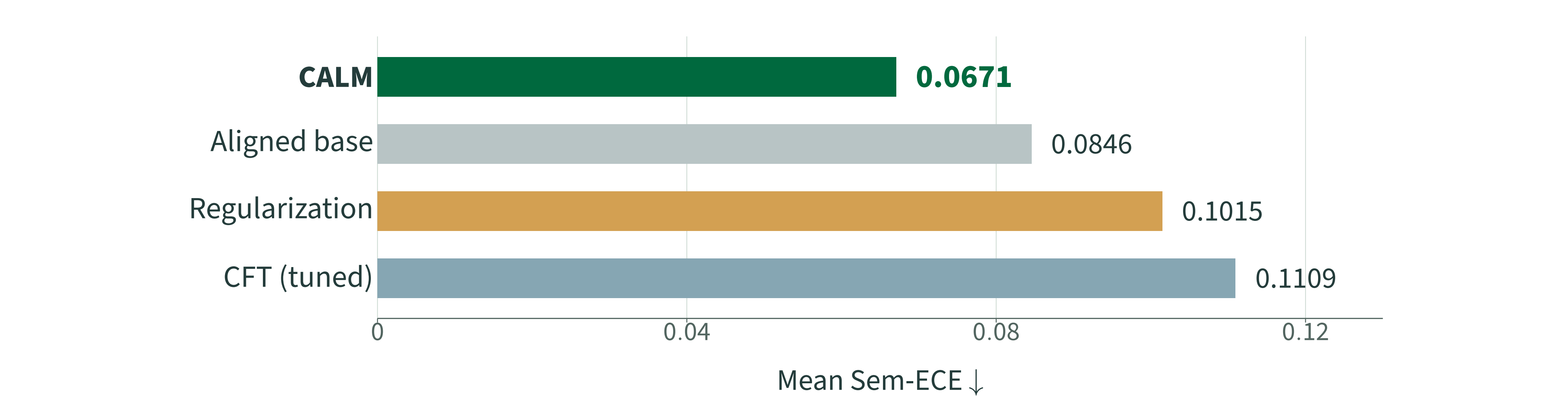
09 OOD MCQA: calibration transfers

3 / 4 models: lowest reported conf-ECE. Table 1



10 Open-ended QA: fixed format, new domain

PopQA → TriviaQA (Table 3)



Best mean across four models; 20.7% below the aligned base. LS collapsed (.6462 mean; omitted).

11 Robustness: seeds and language ability

3 seeds: means improve over base; variability is model-dependent (Table 11).

Model	CALM conf-ECE (mean ± sd)	Base
Llama-3.1	.110 ± .085	.1784
Vicuna-7B	.040 ± .002	.0581
OLMo-2-7B	.089 ± .002	.1115
Mistral-7B	.149 ± .121	.3351

Language ability: OOD mean accuracy change (pp)

	CALM	Regularization	Iterate
vs. aligned base	−2.1	−3.2	−21.3
Average of four models, each evaluated on five tasks; computed from Table 4.			

12 Takeaway: a formulation, not just a solver

Entropy supplies the direction.
Bilevel controls the trade-off.
BOME makes it tractable.

Calibration is empirical, not guaranteed; utility costs and training overhead remain.