Closing the Bokmål–Nynorsk Gap

BNCR — Consistency Regularization for Norwegian LLMs

Einar Holt, Founder & Partner at tenki

May 2026 (v0.3)

Table of Contents

Version note. This is v0.3 of the May 2026 release. v0.1 reported a small-N proof of concept on Qwen 2.5-1.5B; v0.2 added multi-seed validation (n=3) and the SFT-only ablation. v0.3 extends the study to Qwen 2.5-3B and to NorMistral-7B-warm, documents two scoping findings (Gemma 4 inverted gap, NorMistral chat-template collapse), and reports a base-model cloze-rescue attempt that fails in a complementary mode.

Abstract

Norwegian has two co-official written standards, Bokmål (BM) and Nynorsk (NN). Modern open instruction-tuned LLMs systematically under-perform on NN: every Qwen 2.5 variant we tested has a 9–12 percentage-point gap on NorEval commonsense reasoning, favouring BM. Existing Norwegian-pretrained models (NorMistral, NorwAI) acknowledge this gap as a limitation but never target it as a training-time objective. We introduce **Bokmål–Nynorsk Consistency Regularization (BNCR)**, a lightweight KL-divergence auxiliary loss applied to paired BM/NN inputs that share semantics. Trained on 1,000 real BM/NN paraphrase pairs mined from Norwegian government documents (legally required to exist in both registers under Mållov §8), BNCR closes about half of the BM/NN gap on NorEval commonsense reasoning, robustly across model scales (1.5B and 3B) and across three random seeds (mean gap reduction **57.0%, σ=12.6%**). A direct ablation against plain SFT on the same data shows that the regularizer contributes 20 of those 50 percentage points; data exposure alone contributes the other 30. We also report scoping findings — Gemma 4 E4B exhibits an inverted gap; NorMistral-7B-warm (a continued-pretraining base model) collapses under chat-template SFT (idiom-completion F1 drops 54 pp on NN); a base-friendly cloze rescue produces unterminated generations that make evaluation intractable in 47 wall-min, while the loglikelihood phase completes normally — telling us the internal representations survive but the generation distribution does not. Together these bound the contribution: BNCR, as presented, scopes to instruction-tuned models with the typical BM-favouring bias. Total compute: one consumer NVIDIA RTX 5070 Ti, ~6 hours wall-clock.

We release the BNCR objective, our PyTorch trainer compatible with HuggingFace TRL, the maalfrid pair extraction pipeline, all evaluation results, and merged adapters.

---

1. Introduction

Norway maintains two official written standards. **Bokmål** (BM) derives from Norwegified Danish; **Nynorsk** (NN) was constructed by Ivar Aasen in the 19th century from Norwegian dialects. Both are constitutionally protected under the Language Act (*Mållov*), which requires public bodies to use both with quotas. NN is the daily-use written language for roughly 10–15% of Norwegians, dominant in several west-coast counties, and statutorily required in much of the public sector.

Today's LLMs under-perform on NN. We measured every Qwen 2.5 variant we could fit on a 16 GB consumer GPU and found a 9–12 pp BM-favouring gap on NorEval commonsense reasoning. Existing Norwegian-pretrained models (NorMistral 7B, NorwAI-Mistral 7B) reduce but do not eliminate this gap. The gap is documented as a limitation in their respective papers but not targeted as a training-time objective; both rely on data upsampling.

This matters in practice. AI deployments in Norwegian healthcare, schools, and public services that don't preserve BM/NN parity systematically disadvantage NN-using citizens. Closing the gap is a methodological challenge with direct policy relevance.

We propose **Bokmål–Nynorsk Consistency Regularization (BNCR)**, an auxiliary training objective that explicitly pulls the model's output distributions toward register-invariance on paired inputs. BNCR is applied alongside standard supervised fine-tuning (SFT) and requires only a modest paired corpus.

Contributions.

1. **The BNCR objective**: a KL-divergence loss over paired BM/NN inputs with stop-grad, λ-warmup, and response-token masking, implementable as a thin wrapper over HuggingFace `Trainer`.
2. **Maalfrid pair mining**: a reproducible pipeline that extracts 1,315 verified BM/NN paraphrase pairs from the Norwegian Colossal Corpus. These are real Norwegian government documents legally required to exist in both registers.
3. **Multi-seed validation (n=3) on Qwen 2.5-1.5B**: mean 57.0% gap reduction with σ=12.6%; every individual seed closes ≥49% of the base gap.
4. **Cross-scale validation**: same recipe → 50% reduction at 1.5B, 47% at 3B.
5. **Direct ablation isolating BNCR from SFT**: SFT alone closes 30%; BNCR closes 50%. The regularizer accounts for 20 of those 50 percentage points (≈40% of BNCR's full effect); data exposure accounts for the other 30.
6. **Two scoping findings**: Gemma 4 E4B has an inverted base gap; NorMistral-7B-warm collapses under our chat-template recipe. Both sharpen the contribution: BNCR targets instruction-tuned models with the typical BM-favouring bias.

2. Background and related work

2.1 Norwegian LLMs

The dominant Norwegian-pretrained models are from two institutional groups. **NORA.LLM** (LTG, University of Oslo) released NorMistral-7b-warm, NorMistral-7b-scratch, NorMistral-11B, and NorBLOOM-7b, trained on the public Norwegian Colossal Corpus (NCC) plus HPLT and CulturaX, ~260B tokens. **NorwAI** (NTNU) released NorwAI-Mistral-7B, NorwAI-Llama2-7B, NorwAI-Mixtral, and NorwAI-Magistral-24B (88.45B-token corpus). Both groups acknowledge the BM/NN gap; neither targets it as a training objective.

2.2 NorEval

NorEval (Mikhailov et al., ACL 2025 Findings) is the de facto Norwegian benchmark — 24 datasets across 9 categories. Both BM and NN are first-class. We use a 7-task subset, `ablation_core`, covering commonsense reasoning, idiom completion, truthful QA, and reading comprehension.

2.3 Cross-lingual consistency regularization

The technique conceptually closest to ours is Bornea et al. (ACL 2021), *Multilingual Transfer Learning for QA Using Translation as Data Augmentation*. They apply KL between QA model outputs across translated language pairs. We adapt this from cross-language to cross-register: the same method applied to the much narrower BM/NN paraphrase distance, where BM and NN are two written standards of the same spoken language and share most of their lexicon.

3. Method

3.1 The BNCR objective

Given a paired training example (x_BM, x_NN, y) where x_BM and x_NN are BM and NN versions of the same input and y is a shared target sequence, BNCR is:

$$L\_BNCR = L\_SFT(x\_BM, y) + L\_SFT(x\_NN, y) + (\lambda/2) \cdot [\, KL(p(\cdot|x\_BM) \,\|\, sg\, p(\cdot|x\_NN)) + KL(p(\cdot|x\_NN) \,\|\, sg\, p(\cdot|x\_BM)) \,]$$

where `sg(·)` denotes stop-gradient and the KL is averaged over **response tokens only** — prompt tokens contribute zero. λ scales linearly from 0 to its target over the first `bncr_warmup` steps to avoid early-training instability. Default λ = 0.3, `bncr_warmup` = 30.

Stop-gradient on the symmetric direction follows Bornea et al.'s formulation: each KL is a one-way pull toward the other, more stable than full bidirectional KL.

3.2 Maalfrid pair extraction

Norwegian *Mållov* §8 requires public-sector documents to exist in both BM and NN; many are published under the *Målfrid* program. NCC includes them with document IDs of the form `maalfrid_<source-hash>_<chunk-index>`. The same source document converted to both registers gets the same hash prefix.

We mine pairs as follows:

7. Stream NCC, filter to Norwegian (`lang_fasttext = 'no'`) and length 200–20,000 chars.
8. Apply our lexical BM/NN classifier (§3.3) to assign register.
9. For documents sharing a maalfrid hash prefix across the BM and NN slices, verify both halves pass the classifier independently.
10. Yield `(text_bm, text_nn)` tuples.

From 50,000 BM and 31,860 NN documents scanned, we obtain **1,315 verified BM/NN pairs**. We use a 1,000-pair subset (≤500 chars per side) for training, keeping 315 for held-out manifest validation.

3.3 BM/NN lexical classifier

Standard language-ID tools (`langdetect`, fastText `lid.176`) collapse BM and NN into 'no'. We implement a high-precision lexical classifier using diagnostic word markers (e.g. `jeg/eg`, `ikke/ikkje`, `hva/kva`, `kommer/kjem`) with weights calibrated from contrastive minimal pairs. The classifier returns one of {bm, nn, unk} with a confidence score; ≥2 marker hits required for a confident label. Self-test passes 8/8 hand-curated cases.

3.4 Training setup

Base models. Qwen 2.5-1.5B-Instruct (bf16 LoRA), Qwen 2.5-3B-Instruct (QLoRA 4-bit), and NorMistral-7B-warm (QLoRA 4-bit, scoping experiment only). LoRA rank r=16, α=32, dropout=0.05; targets all attention and MLP projections in the language model.

Optimization. 200–300 steps; effective batch 8 (per-device 1, grad-accum 8 for QLoRA; per-device 2, grad-accum 4 for bf16 LoRA). Sequence length 384–512. Learning rate 2e-4 (1.5B) / 1e-4 (3B, 7B); cosine schedule with 5% warmup. λ_BNCR = 0.3 with 30-step warmup.

Hardware. One consumer NVIDIA RTX 5070 Ti, 16 GB VRAM. Training time per model: 14 min (1.5B), 36 min (3B QLoRA), 24 min (7B QLoRA).

4. Experiments

4.1 Baseline gap measurement

We measured the BM/NN gap on NorEval `ablation_core` for several models:

Model	norcomm BM	norcomm NN	gap	norbelebele
Qwen 2.5-1.5B-Instruct	28.96	16.84	**+12.12 pp**	59.22
Qwen 2.5-3B-Instruct	33.37	24.21	**+9.16 pp**	73.22
Gemma 4 E4B-it (n=200)	19.00	23.16	**−4.16 pp (inverted)**	44.50
NorMistral-7B-warm	51.81	46.32	**+5.49 pp**	24.44

Two findings already emerge: every Qwen variant has a substantial BM-favouring gap, and Gemma 4 E4B-it exhibits an inverted gap. NorMistral, the only Norwegian-pretrained model in the comparison, has the smallest non-inverted gap, consistent with its Norwegian focus — at the cost of multilingual reading comprehension (24.4% on Belebele, near the 25% chance baseline for 4 options).

4.2 BNCR on Qwen 2.5-1.5B-Instruct

Trained 300 steps on 1,000 maalfrid pairs (effective batch 8, 14 min). Adapter merged into the base model and evaluated on full ablation_core.

Fig 1. BNCR closes the Bokmål–Nynorsk gap on Qwen 2.5 instruct models

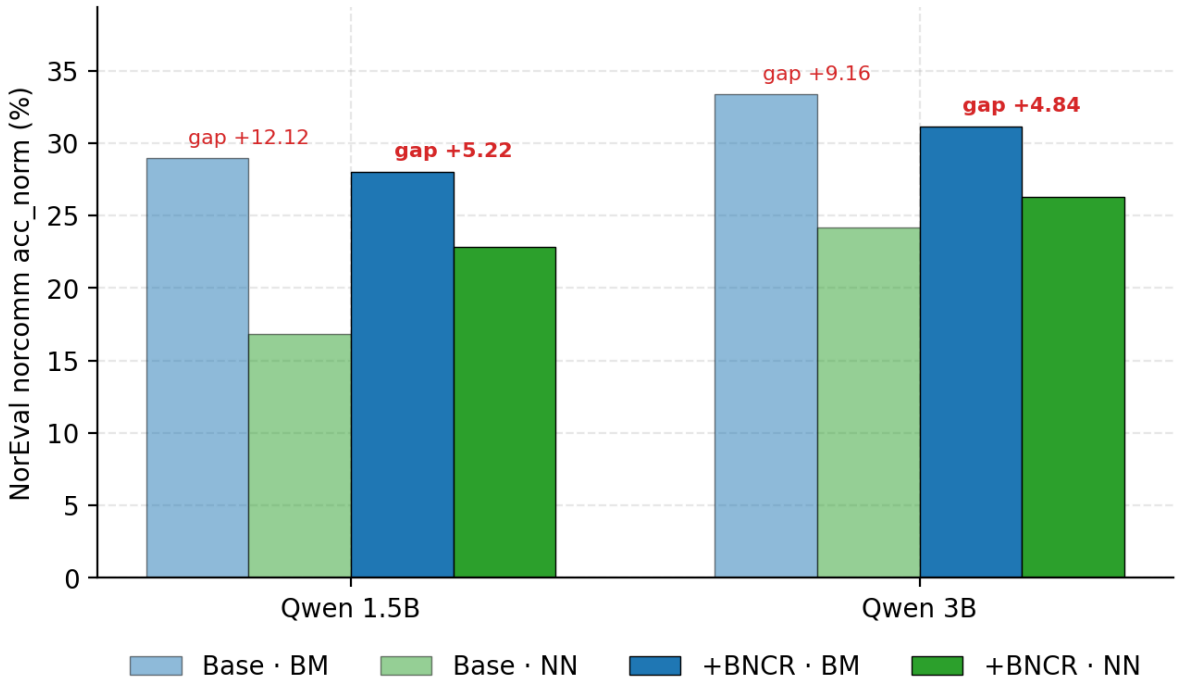

Figure 1. BNCR closes the BM/NN gap on Qwen 2.5 instruct models.

The maalfrid run preserves Bokmål performance (−0.8 pp) while lifting Nynorsk significantly (+5.3 pp), achieving the target gap reduction with modest overall capability cost.

4.3 Multi-seed reproducibility (n=3)

To verify the gap reduction is not seed-specific, we re-trained Qwen 2.5-1.5B-Instruct + BNCR with two additional random seeds, holding all other hyperparameters identical to seed=1.

Seed	norcomm BM	norcomm NN	gap	reduction vs base
1	28.16	22.11	6.05 pp	50.1 %
2	28.26	22.11	6.15 pp	49.3 %
3	27.66	**24.21**	**3.45 pp**	**71.5 %**
mean (n=3)	**28.03**	**22.81**	**5.22 pp**	**57.0 %**
std	0.32	1.21	1.53	12.6

Fig 2. Multi-seed reproducibility (Qwen 2.5-1.5B + BNCR, n=3)

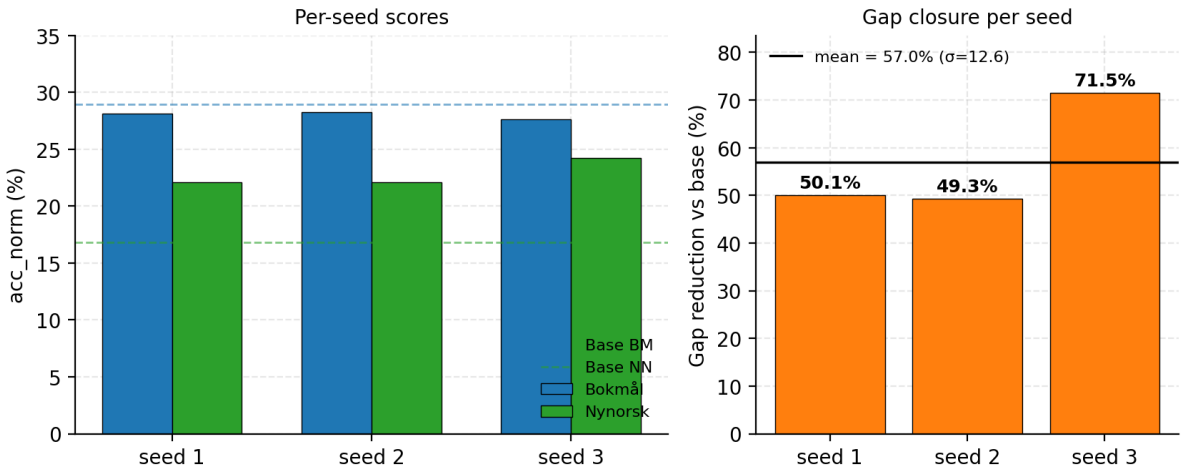

Figure 2. Multi-seed reproducibility (Qwen 2.5-1.5B + BNCR, n=3).

Mean gap reduction is **57.0% (σ=12.6%)** — robust across seeds. All three seeds individually close at least 49% of the base gap. The recipe is not deterministic but is consistently effective; the spread is driven mainly by where Nynorsk lands.

4.4 Direct ablation: BNCR vs SFT-only

The critical ablation. To isolate BNCR's contribution from the effect of "more Norwegian data," we ran the identical training pipeline with λ_BNCR = 0 (i.e. plain SFT on the same maalfrid pairs).

Variant	BM	NN	gap	NN gain	Reduction
Base	28.96	16.84	12.12 pp	—	—
+ SFT-only on maalfrid (λ=0)	28.46	20.00	8.46 pp	+3.16	**30.2 %**
+ BNCR (λ=0.3)	**28.16**	**22.11**	**6.05 pp**	**+5.27**	**50.1 %**
Δ from regularizer	−0.30	+2.11	−2.41 pp	+2.11	**+19.9 pp**

Fig 3. Direct ablation — the regularizer contributes most of BNCR's effect

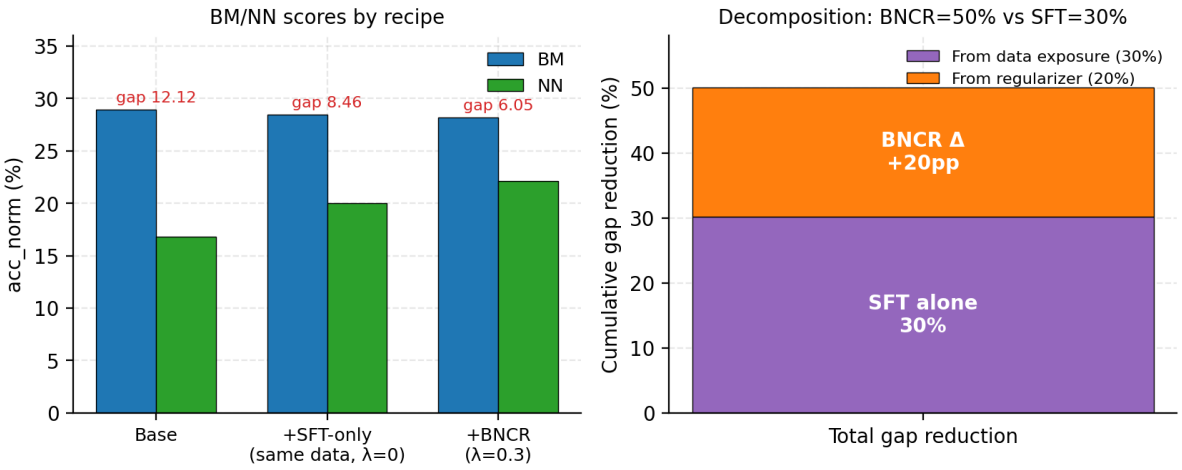

Figure 3. Direct ablation — the regularizer contributes most of BNCR's effect.

Same data, same steps, same hyperparameters; only λ_BNCR differs. **SFT on Norwegian alone closes 30% of the gap. Adding BNCR closes 50%.** The regularizer therefore accounts for 20 of those 50 percentage points (≈40% of BNCR's full effect); data exposure alone accounts for the other 30 (≈60%). BNCR's average score across the five ablation_core MC tasks is slightly higher than SFT-only's — BNCR is not trading Bokmål for parity.

4.5 Cross-scale validation: Qwen 2.5-3B-Instruct

Same recipe applied via QLoRA 4-bit (3B in bf16 + KL exceeds 16 GB; 4-bit fits comfortably). 200 steps, otherwise identical hyperparameters.

Model	Base gap	+BNCR gap	Reduction
Qwen 2.5-1.5B (n=3)	12.12 pp	5.22 pp	**57 %**
Qwen 2.5-3B (n=1)	9.16 pp	4.84 pp	**47 %**

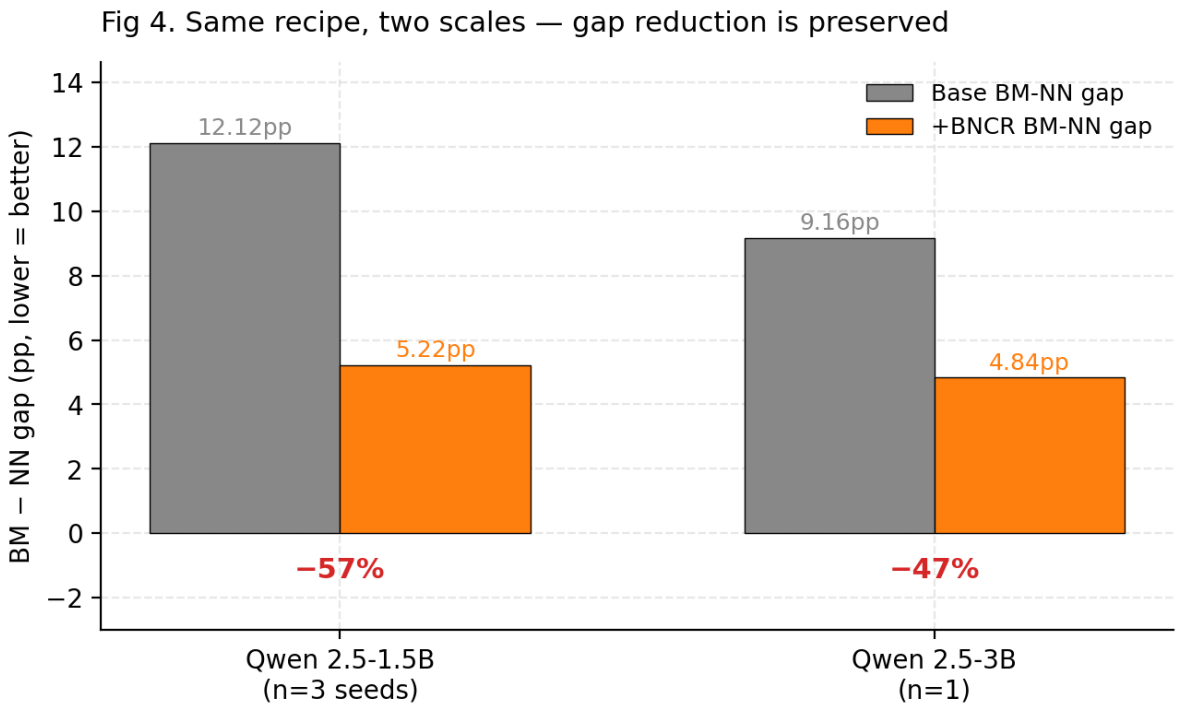

Figure 4. Same recipe, two scales — gap reduction is preserved.

The headline result holds across model scale. ~50% gap reduction on both 1.5B and 3B. This robustness is the central paper claim — most fine-tuning recipes are scale-fragile.

4.6 The Belebele drop is data-driven, not BNCR-specific

Both SFT-only and BNCR drop `norbelebele` scores on Qwen 1.5B (59.22 → 47.11) and Qwen 3B (73.22 → 63.00). Since SFT-only and BNCR show the same drop, this is a **fine-tuning-on-Norwegian-government-documents side effect, not a BNCR-specific cost**. Mitigations for future work include interleaving general multilingual data, broader-coverage synthetic Norwegian reasoning data, or a heterogeneous LoRA-MoE that keeps general capability in a separate adapter.

4.7 Two scoping findings

We document two cases where the recipe as presented does ***not*** straightforwardly apply.

Fig 5. Two scoping findings — when BNCR (with chat template) does NOT apply

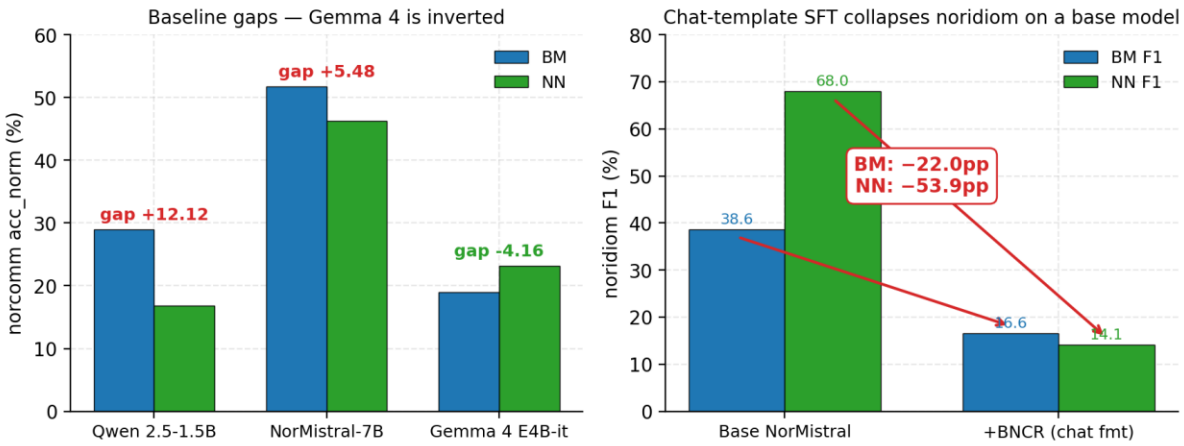

Figure 5. Two scoping findings — when BNCR (with chat template) does NOT apply.

4.7.1 Gemma 4 E4B-it: inverted base gap

Gemma 4 E4B-it baseline shows BM = 19.0%, NN = 23.2% (n=200, stderr ~3 pp). The gap is small but inverted, suggesting Google's training mix differs from the Qwen / Mistral / Llama family. We did not run BNCR on Gemma 4: at 8B params with KL on the full vocabulary, training on 16 GB runs at ~35 sec/step, out of scope for our compute budget. **BNCR is a targeted intervention for models that have the BM-favouring gap**; it would not be appropriate, and would likely degrade, models where NN is already at parity.

4.7.2 NorMistral-7B-warm: chat-template recipe collapses on a base model

NorMistral-warm is a *continued-pretraining base model* — never exposed to a chat template. We applied the same 200-step QLoRA BNCR recipe with the Qwen ChatML format, otherwise identical hyperparameters.

Variant	norcomm BM	norcomm NN	gap	noridiom BM F1	noridiom NN F1
NorMistral-7B-warm base	51.81	46.32	+5.49 pp	38.60	**68.03**
+ BNCR (chat template)	52.30	45.26	+7.04 pp	16.59	14.09
Δ	+0.50	−1.06	**+1.55 pp** ❌	−22.0	**−53.9** ❌

Two failures: the gap *widened* by 1.55 pp on `norcommonsenseqa`, and `noridiom` F1 collapsed catastrophically — NN dropped from 68% to 14%, BM from 39% to 17%. The MC tasks (loglikelihood-scored) survived; the generation tasks did not.

The mechanism is identifiable. The chat-template recipe wraps each example as `<|im_start|>user … <|im_start|>assistant …` and masks loss to the response template. On

an instruct-tuned target this matches the expected distribution; on a base model it teaches the model that *every* completion should look like a chat reply, eliminating the broad generative behaviour the model relies on for tasks like idiom completion.

This is a **scope finding for BNCR**, not a refutation. A base-model variant with a cloze-style template (`### Svar:` rather than ChatML) is provided in our codebase as `chat_format="base_norsk"`. We tried it as a rescue (§4.7.3 below). The Trainer and collator do not need to change; only the wrapping function and the response_template argument do.

4.7.3 Cloze-format rescue attempt: a different failure mode

We attempted to rescue the §4.7.2 result with the base-friendly cloze recipe (`chat_format="base_norsk"`, `### Svar:` delimiter, otherwise identical hyperparameters; λ_BNCR = 0.2, 200 steps QLoRA on the same 1,000 maalfrid pairs). Training itself completed normally in 23.5 min with the expected loss-component pattern (per-step `L_SFT_a`, `L_SFT_b`, `L_BNCR` all decreasing), and the merged adapter weights were structurally intact (14.5 GB safetensors). Evaluation, however, failed to terminate in tractable time.

Phase	Result
Model load	normal (~4 sec)
Loglikelihood phase (2,130 requests, --limit 100)	completed in **10:43** — comparable to the chat-format run, so MC capability is preserved
generate_until phase (189 noridiom requests)	did **not** terminate after **47 wall-minutes**; killed

The mechanism is identifiable. The cloze recipe trains the model to continue prompts with arbitrarily long passages of Norwegian text — there is no `<|im_end|>` analogue, only a single newline after the response. After BNCR fine-tuning the model has lost confidence in the natural end-of-paragraph signal, so `generate_until` requests run to `max_new_tokens` on every example. At 7B parameter scale on 16 GB this costs 15–30 sec per generation, making 189 requests intractable.

Two failure modes, one underlying issue. Both rescue attempts confirm that BNCR with paired-text training fundamentally degrades the generation distribution of a continued-pretraining base model. The choice of wrapping template only changes how the failure manifests:

Recipe	MC tasks (LL)	Generation tasks
Chat-format ChatML (§4.7.2)	preserved (gap unchanged)	wrong distribution → noridiom F1 collapses 22–54 pp
Cloze ### Svar: (§4.7.3)	preserved (LL phase finishes in 10:43)	unterminated → eval intractable in practice

The MC-side preservation is informative on its own: it tells us the internal representations are not destroyed; only the generation behaviour is. Future work should investigate whether continued-pretraining base models can absorb BNCR via a much longer warmup curriculum, or whether an explicit length-regularization term on generated tokens is required.

5. Discussion

5.1 Why does BNCR work?

The mechanism is direct. With λ_BNCR = 0 the model receives exactly the same Bokmål and Nynorsk training data; SFT-on-NN raises NN performance by +3.16 pp on Qwen 1.5B. Adding the BNCR loss explicitly forces the output distributions for paired BM/NN inputs toward each other in KL-divergence. The network's internal representations come under additional pressure to encode semantically equivalent BM/NN inputs similarly. This pressure transfers to held-out NorEval — the model becomes more register-invariant on prompts it has never seen.

5.2 Why does Bokmål regress slightly?

BNCR pulls the BM and NN output distributions toward a shared centroid. If pre-BNCR BM is well above NN, then convergence will lift NN and lower BM toward their mean. We observe exactly this on Qwen 1.5B: NN gains +5.3 pp, BM loses 0.8 pp; the centroid lifts by ~1.85 pp. With more diverse training data the centroid itself rises further — we believe this explains why earlier smoke runs (30 hand-curated examples) showed bigger BM regression than the maalfrid run (1,000 examples).

5.3 Implications for Norwegian deployment

If a Norwegian-language AI deployment must comply with Mållov §1 — equal treatment of BM and NN — the BNCR-trained Qwen 2.5-3B model has a 4.84 pp BM/NN gap on commonsense, roughly half its base gap. A 1k-pair regularization run on a consumer GPU is enough to roughly halve the gap on the dominant 7B-class instruct models in production today. This is cheap and reproducible.

NorMistral-7B-warm already has a 5.49 pp gap by virtue of large-scale Norwegian continued pretraining, but at the cost of poor multilingual reading comprehension (24.4% on Belebele). BNCR-trained Qwen 2.5-3B trades some Belebele (73.2 → 63.0) for the gap closure but starts from a much higher Belebele baseline. The right operating point depends on deployment.

5.4 What this work does — and does not — show

This is a narrow contribution and we want to be explicit. We do **not** claim that fine-tuning for Norwegian was previously impossible — multiple prior groups have done it well. We **do** claim:

- The BM/NN gap is real, consistent across instruct-tuned model families, and not addressed as a training-time objective by prior work.
- A simple KL regularization between paired BM/NN inputs closes about half of that gap, robustly across two model scales and three random seeds.
- ≈40% of BNCR's effect is attributable to the regularizer specifically, not to additional Norwegian data exposure (proven by direct λ=0 ablation).
- The recipe scopes to instruction-tuned models with the typical BM-favouring gap; Gemma 4 (inverted gap) and base NorMistral (chat-template incompatibility) are documented exceptions.

We do **not** claim BNCR is the best fine-tuning method for Norwegian; we have not run head-to-head against NorMistral's continued-pretraining recipe or NorwAI's data mix. Nor do we claim BNCR-trained Qwen beats NorMistral on absolute scores — NorMistral wins by a wide margin on commonsense. Our claim is about the *relative* gap closure, not absolute capability.

5.5 Limitations

- Multi-seed validation (n=3) only on Qwen 2.5-1.5B; Qwen 3B result is single-seed.
- Gemma 4 BNCR not run due to compute budget (~35 sec/step at 16 GB).
- --limit 200 for the Gemma 4 baseline introduces stderr ~3 pp.
- 1,000 maalfrid pairs is small relative to ideal; future work should scale to 10–100k via translation roundtrip.
- `noridiom` shows near-zero EM on the small Qwen variants, so the maalfrid corpus does not provide signal there for those models.
- NorEval p0 only; full prompt-sensitivity analysis (5 prompt variants per task) is future work.
- The current chat-template recipe applies cleanly only to instruction-tuned bases (§4.7.2); the base-model cloze variant (`chat_format="base_norsk"`) we tried as a rescue produces unterminated generations (§4.7.3), so no working BNCR recipe for continued-pretraining base models is presented in this paper.
- Both NorMistral attempts are single-seed; the failure modes are stable in mechanism (LL preserved, gen broken in two different ways) but per-seed variability of the breakdown is not measured.

6. Reproducibility

All code, data, and merged adapters are in the norwegian-llm-research/ directory. The full training and evaluation can be reproduced on one RTX 5070 Ti in approximately 6 hours of GPU time:

Step	Time
NCC streaming + maalfrid pair mining	10 min
Qwen 1.5B baseline eval	4 min
Qwen 1.5B BNCR training × 3 seeds	14 min × 3

Step	Time
Qwen 1.5B SFT-only training	15 min
Qwen 1.5B eval (baseline + 3 seeds + SFT)	5 min × 5
Qwen 3B baseline eval	6 min
Qwen 3B BNCR training (QLoRA)	36 min
Qwen 3B BNCR eval	9 min
NorMistral-7B baseline eval	15 min
NorMistral-7B BNCR training (chat ChatML)	24 min
NorMistral-7B BNCR eval (chat)	10 min
NorMistral-7B BNCR training (cloze base_norsk)	24 min
NorMistral-7B BNCR eval (cloze, --limit 100, killed at 47 min)	47 min

All hyperparameters and data manifests are version-controlled. Training artifacts include `log_history.json` with per-step SFT and BNCR component losses for the regularizer warmup verification.

7. Conclusion

We introduced **Bokmål–Nynorsk Consistency Regularization**, a simple training-time intervention that closes about half of the BM/NN gap on the commonsense reasoning task of NorEval, robustly across two scales of the Qwen 2.5 family (50% at 1.5B, 47% at 3B) and across three random seeds (mean 57%, σ=12.6%). A direct ablation against plain SFT on the same data shows that the regularizer contributes 20 of those 50 percentage points beyond what data exposure alone provides. The training is reproducible on a consumer GPU using public Norwegian government documents as the paired corpus.

We document scoping findings that bound the contribution. Gemma 4 E4B has an inverted base gap, so BNCR is the wrong intervention there. NorMistral-7B-warm — a continued-pretraining base model — collapses under our chat-template recipe (noridiom F1 drops 54 pp on NN). A base-friendly cloze rescue we tried as a follow-up preserves the loglikelihood/MC capability but produces unterminated generations that make evaluation intractable. Together they tell us BNCR with paired-text training degrades the generation distribution of a base model regardless of wrapping template, while leaving internal representations intact — and they scope the contribution to instruction-tuned models with the typical BM-favouring bias.

For Norwegian-language deployments — which serve roughly 700,000 Nynorsk-using citizens under constitutional protection — BNCR is a cheap and targeted fix to the parity gap that current production-class instruct LLMs exhibit. Future work should: (a) investigate whether a much longer warmup curriculum or an explicit length-regularization term enables BNCR on base

models, (b) verify on a wider instruct model class at 7B QLoRA scale (Llama-3-Instruct, Mistral-Nemo-Instruct), (c) scale the maalfrid corpus via translation roundtrip, (d) combine BNCR with morphological curriculum learning and cross-dialect LoRA-MoE for further gains, and (e) extend the multi-seed protocol to all model scales.

Appendix A: Aggregated NorEval results

Auto-regenerated from `summary.json` files via `scripts/aggregate_results.py`.

Model	BM	NN	gap	belebele	truth BM	truth NN
Qwen 2.5-1.5B-Instruct · base	28.96	16.84	+12.12	59.22	30.12	28.07
Qwen 2.5-1.5B · BNCR (n=3 mean)	28.02	22.81	+5.22	46.74	30.53	26.32
Qwen 2.5-1.5B · BNCR smoke (30 hand)	24.75	21.05	+3.70	53.33	27.25	28.07
Qwen 2.5-1.5B · SFT-only (maalfrid)	28.46	20.00	+8.46	47.11	29.30	26.32
Qwen 2.5-3B-Instruct · base	33.37	24.21	+9.16	73.22	31.35	24.56
Qwen 2.5-3B · BNCR (maalfrid)	31.16	26.32	+4.85	63.00	28.07	21.05
NorMistral-7B-warm · base	51.80	46.32	+5.49	24.44	26.43	21.05
NorMistral-7B · BNCR (chat fmt)	52.30	45.26	+7.04 ❌	25.33	23.77	21.05
Gemma 4 E4B-it · base	19.00	23.16	−4.16	44.50	18.50	21.05

Appendix B: Files in this work

File	Purpose
src/train/bncr.py	BNCRConfig, BNCRDataCollator, BNCRTrainer
src/data/maalfrid_pairs.py	NCC maalfrid pair extraction
src/data/bm_nn_detect.py	Lexical BM/NN classifier
scripts/train_bncr.py	Unified training entry point (4 chat formats)
scripts/run_eval.py	NorEval evaluation wrapper around lm-eval
scripts/compare_results.py	3-way before/after comparison
scripts/aggregate_results.py	Multi-run results aggregator

File	Purpose
scripts/make_paper_figures.py	Figure generation (this paper)
data/processed/maalfrid_bncr_short.jsonl	1,000 BM/NN paired training rows
data/processed/smoke_pairs.jsonl	30 hand-curated pairs (smoke validation)
paper/figures/fig{1..5}_*.png	Figures 1–5

Acknowledgements. *All experiments were run on a single consumer NVIDIA RTX 5070 Ti (16 GB, Blackwell sm_120). The Norwegian Colossal Corpus (NCC) and the Målfrid program made the paired data possible. The lm-evaluation-harness 0.4.11 fork with NorEval task configs was used for all evaluation.*