Teaching a Small LLM Scholastic Voice: Fine-Tuning Qwen 2.5 on the Catechism, Summa, and Augustine via Local MLX

Pablo Leyva
R1 New Jersey Institute of Technology
pleyva2004@gmail.com

Abstract

General-purpose instruction-tuned language models default to a flat, encyclopedic register. Specialized historical writing (the Summa's *quaestio* form, Augustine's autobiographical address, the Catechism's numbered paragraph citations) has a distinct lexical and structural fingerprint that is largely absent from base-model output. We ask whether a small (7B-parameter) open-weights model can be shifted into such a register cheaply, on a laptop, using only public-domain or fair-use source texts and a modest budget. We post-train Qwen 2.5 7B-Instruct on Apple Silicon under MLX with 8-bit weight quantization and LoRA, using 83 teacher-distilled (question, scholastic-answer) pairs synthesized from the Catechism of the Catholic Church (1992), Aquinas's *Summa Theologica*, and Augustine's *Confessions* and *City of God*. Training takes about seven minutes at a peak resident memory of 12.7 GB on an M4 Pro and updates only 2.6M of 7.6B parameters. On a rubric measuring scholastic register, Augustinian voice, CCC grounding, and argument structure across 10 held-out philosophical prompts, the fine-tuned model scores 68/120 against the base model's 19/120, a 258 % relative gain, with the largest improvements in register ($3 \to 20$) and CCC grounding ($0 \to 19$). We contrast this positive result against earlier failed DPO and ORPO runs on a 1.5B 4-bit base, where preference optimization left generations effectively unchanged. Code, recipe, and rubric are released. A second phase that scales data four-fold and chains DPO refinement matches Phase 1's strict total and improves a balanced register-fluency metric (68/90 vs 66/90), with iter 400 as the best checkpoint; the DPO chain contributed nothing, a setup pitfall we document.

1 Introduction

Modern instruction-tuned large language models (LLMs) are trained to produce fluent, encyclopedic prose that is appropriate to a generic helpful-assistant persona. This is exactly what most product use cases reward, and exactly what makes them unsatisfying for tasks that demand a specific historical register. Aquinas's *Summa Theologica* is recognizable not only by its theological commitments but by its surface form, the *quaestio*, the "It seems that," the explicit numbered objections, the "On the contrary, . . . ," the "I answer that, . . . ," followed by replies to each objection. Augustine writes confessionally, addressing God in the second person, drawing autobiographical detail toward dogmatic claim. The Catechism of the Catholic Church (CCC, 1992) cites itself by numbered paragraph and binds doctrine to scriptural and patristic

references. A foundation model trained on the modern web can repeat propositions from these documents but rarely *sounds* like them when prompted to speak in their voice.

This paper asks a narrow question: can a small open-weights model (on the order of 7B parameters) be moved into the scholastic register cheaply, on a single laptop, using only public-domain or fair-use source texts and a small teacher-distillation budget?

We answer affirmatively for stylistic transfer. We post-train Qwen 2.5 7B-Instruct [9] on Apple Silicon (M4 Pro, 48 GB unified memory) using MLX [1] with 8-bit weight quantization and LoRA [3], on 83 teacher-distilled (question, answer) pairs synthesized by Claude Sonnet from primary source chunks. Training takes seven minutes, peaks at 12.7 GB of resident memory, and updates 2.6M of 7.6B parameters (0.034 %). The resulting adapter raises a coarse register-and-grounding rubric from 19/120 to 68/120 across 10 held-out philosophical prompts, with the strongest gains in scholastic register and CCC paragraph grounding.

We also report what did *not* work earlier in this project, since the contrast motivated the recipe above. Preference-optimization experiments (DPO, ORPO) on a smaller (Qwen 2.5 1.5B, 4-bit) base model produced seemingly excellent validation metrics (DPO converged to 1.0 accuracy with a 7.5-nat margin), while generations were nearly byte-identical to the base. This is the classic asymmetric-reward DPO failure mode [5]: the model learns to push the rejected response down without pulling the chosen response up. Capacity (1.5B) and aggressive quantization (4-bit) plausibly both contributed. The Phase 1 recipe we report here was a deliberate response: a larger base (7B), gentler quantization (Q8), supervised fine-tuning before any preference signal.

Contributions. This paper makes three contributions:

1. A reproducible recipe for stylistic and citation-style transfer in a 7B-parameter model under MLX Q8 + LoRA, training end-to-end in ≈ 7 minutes on a 48 GB M4 Pro.
2. A teacher-distillation pipeline (scrape → clean → Claude-as-teacher → SFT) that produces ≈ 90 high-quality training pairs for $\approx \$0.50$ in API cost, with the option to scale to hundreds without retraining the teacher.
3. A rubric-based evaluation of register, voice, citation grounding, and structural form that, while coarse, exposes large and signed deltas between base and fine-tuned models. We release the rubric and a held-out 10-prompt evaluation set.

Roadmap. Section 2 surveys related work on instruction tuning, parameter-efficient fine-tuning, preference optimization, on-device training, and stylistic adaptation. Section 3 describes data sources, the teacher-distilled pair generator, the training recipe, and the evaluation rubric. Section 4 reports training dynamics, rubric results, and verbatim qualitative comparisons. Section 5 reflects on why the recipe worked where DPO/ORPO on a 1.5B base did not, on the "Augustinian gap" caused by data imbalance, on ethics, and on limitations. Section 6 closes with future work.

2 Related Work

Instruction tuning. Alpaca [7] popularized the recipe of fine-tuning an open base model on a small synthetic instruction dataset generated by a stronger teacher, and the Self-Instruct line [8] established that LM-generated instruction data is good enough to train a useful assistant. UltraChat and follow-on dialog datasets extended this to multi-turn and long-context settings. Our pipeline is firmly in this tradition: a stronger teacher model (Claude Sonnet 4.6) distills

source material into instruction pairs that an open student (Qwen 2.5 7B-Instruct) learns from. The difference is target: we tune for *register* (Summa-style quaestio, Augustinian address, CCC citation form), not for general helpfulness.

Parameter-efficient fine-tuning. LoRA [3] factors weight updates into low-rank adapters, drastically reducing the number of trainable parameters and the memory required to fine-tune large models; DoRA decomposes the update into magnitude and direction for additional sample efficiency. We adopt LoRA on the top 16 transformer layers under MLX, which together with 8-bit weight quantization brings end-to-end training of a 7B model into the 12–13 GB memory budget of a 48 GB unified-memory laptop. Updating only 2.6M of 7.6B parameters appears sufficient for stylistic transfer in our setting.

Preference optimization. DPO [5] reframes RLHF [4] as a single likelihood objective on preference pairs, sidestepping a separate reward model. ORPO [2] folds reference-free preference learning into a single training stage with an SFT-like loss. GRPO [6] (introduced in DeepSeekMath) generalizes to grouped relative-preference signals. We tried DPO and ORPO before SFT in this project, on a smaller (1.5B, 4-bit) base, and observed the asymmetric-reward failure mode: preference-objective validation metrics looked excellent (DPO val accuracy 1.0, margin 7.5 nats) while generations were nearly byte-identical to the base for most prompts. Section 5 returns to why we believe SFT-first on a larger Q8 base broke this pattern.

On-device LLM fine-tuning. MLX [1] provides a unified-memory-aware NumPy-like framework for Apple Silicon, with native support for LoRA fine-tuning of quantized weights. Recent reports show that 7B-class models can be supervised fine-tuned on 48–64 GB M-series machines in minutes rather than hours, removing a key barrier to small research projects. Our work is one such report.

Stylistic and domain adaptation. A long line of work demonstrates that fine-tuning shifts surface style as readily as it shifts factual content, often more readily. Stylistic transfer in LLMs has been explored for legal English, medical English, and literary translation. We focus on a niche register (thirteenth-century scholastic and fourth/fifth-century patristic theology in English translation) and adopt teacher distillation precisely because hand-curated data at the register level is impractical at small scale.

Religious-text NLP. The bulk of corpus work on religious texts sits in parallel-translation studies (Bible translations, Quran translations) and in lexicon-and-grammar tasks for low-resource liturgical languages. LLM-era work on the Catechism, Summa, or Augustinian corpus appears to be sparse; we are not aware of prior work explicitly fine-tuning a contemporary chat model to adopt the Summa's *quaestio* surface form together with CCC paragraph grounding. We note this as a gap rather than as a claim of novelty against an exhaustive search.

3 Method

3.1 Data sources

The training corpus is sourced from four primary works, summarized in Table 1. Two are English public-domain translations from the early twentieth century, one is a contemporary copyrighted

catechetical work used here under a fair-use research posture, and one (City of God) shares the same translator family as Confessions.

Table 1: Primary sources used to build the training corpus.

Work	**Translation**	**Status**	**Source**
Catechism of the Catholic Church (1992)	USCCB / Libreria Editrice Vaticana	Copyrighted (fair use)	vatican.va
Summa Theologica	Fr. Laurence Shapcote, 1920	Public domain (US)	newadvent.org
Augustine, *Confessions*	E.B. Pusey	Public domain	newadvent.org
Augustine, *City of God*	Marcus Dods	Public domain	newadvent.org

After scraping with a respectful crawler (descriptive user agent, 1-second sleep between requests, on-disk cache, `robots.txt` compliance) and cleaning, we obtain a corpus of 342 structured chunks distributed across the four sources. Neither the raw HTML, the cleaned JSONL chunks, nor the trained adapter weights are committed to the public repository; only code, recipe, aggregate metrics, and short qualitative excerpts are shared. See the project's `DATA_LICENSING.md` for details.

3.2 Pipeline

Figure 1 shows the end-to-end data path. The pipeline is deliberately small and inspectable, each arrow is a short Python script with a single responsibility.

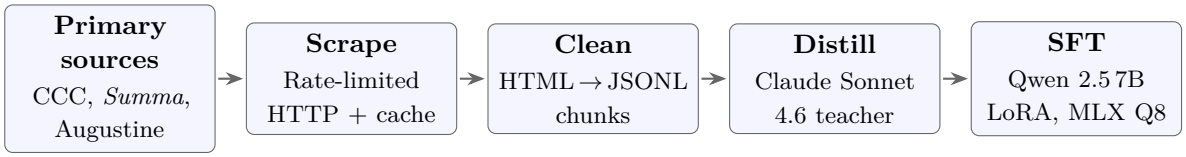

Figure 1: End-to-end data pipeline for Phase 1. Primary sources are scraped respectfully, cleaned into a structured JSONL chunk corpus, then expanded by a Claude Sonnet 4.6 teacher into (question, scholastic-answer) pairs that drive LoRA SFT of Qwen 2.5 7B-Instruct under MLX with 8-bit weight quantization.

3.3 Teacher-distilled training data

For each source chunk we ask Claude Sonnet 4.6, acting as a teacher model, to produce 2–3 (question, answer) pairs in which the answer is written in scholastic register. The system prompt (paraphrased here; full text in `scripts/generate_training_pairs.py`) instructs the teacher to:

- produce a question a graduate student of theology or philosophy might plausibly ask, drawn from the chunk's subject matter;
- answer in either Summa form ("It seems that . . . Objection 1 . . . On the contrary . . . I answer that . . . Reply to Objection 1 . . . ") or in Augustinian form (autobiographical second-person address with scriptural allusion), as fits the chunk;
- cite the CCC by paragraph number where doctrinally apposite;
- refuse to invent CCC paragraph numbers if not directly attested in the chunk, instead citing thematically.

The teacher outputs JSON conforming to a strict schema, validated by the client. Across 342 chunks, the teacher produced 93 (question, answer) pairs (we discard malformed JSON or pairs shorter than a threshold). We split 83/10 train/valid by random shuffle with a fixed seed. The total API cost for this run was approximately $0.50 at then-current Claude Sonnet pricing.

3.4 Model and training

We start from Qwen 2.5 7B-Instruct [9], converted to MLX format with 8-bit weight quantization (`mlx_lm.convert -q -q-bits 8`). On disk the quantized model is ≈ 7.8 GB; resident memory during training peaks at 12.7 GB.

We attach LoRA adapters to the top 16 transformer layers (out of 28), training with AdamW at a learning rate of 10^{-5}, batch size 1, maximum sequence length 2048, for 200 iterations. With batch 1 and 83 training pairs this is approximately 2.4 epochs. Training throughput is 0.48 it/s ($\approx$ 235 tok/s) and the full 200-iteration run takes about seven minutes wall-clock on the M4 Pro. We use the standard MLX-LM LoRA [1] training entrypoint (`mlx_lm_lora.train`). The adapter writes 2.6M trainable parameters out of 7.6B total (0.034 %).

3.5 Evaluation rubric

Because no shared benchmark exists for scholastic-register transfer, we score model outputs on a custom rubric implemented in `src/scholastic/rubric.py`. The rubric is intentionally simple: four orthogonal dimensions, each scored 0–3, giving a per-prompt total 0–12. Across N prompts the maximum is $12N$.

Scholastic register (0–3) Counts of distinct surface markers like *I answer that*, *It seems that*, *Objection*, *Reply to Objection*, *On the contrary*, and scholastic connectives (*accordingly*, *insofar as*, *wherein*, etc.).
Augustinian voice (0–3) Distinct hits among autobiographical or confessional markers (*O Lord*, archaic second-person pronouns, *my soul*, *the City of God*).
CCC grounding (0–3) Visible paragraph-number citations (§ n, *CCC* n, "paragraph n").
Structure (0–3) Number of paragraphs combined with the presence of objection-and-response structural markers.

This rubric is a coarse signal. It rewards lexical and structural surface form rather than theological correctness, and it can be gamed by shallow string interpolation. We treat it as a useful proxy for register transfer, not as a substitute for human evaluation; see Section 5.5 for discussion.

4 Experiments

4.1 Setup

We hold out 10 philosophical prompts not seen during training, covering classical scholastic and Augustinian topics. The full list is:

1. How do you reconcile divine foreknowledge with free will?
2. What is the relationship between faith and reason?
3. Discuss the problem of evil from a Thomistic perspective.
4. How does Augustine understand time in *Confessions*, Book XI?
5. What does the Catechism teach about conscience and moral formation?
6. Argue for or against the proposition that the soul is immortal.

7. Explain the doctrine of divine simplicity and its consequences.
8. How do the two cities of Augustine relate to political authority?
9. What is the role of the virtues in the Christian moral life?
10. How should we understand the relationship between grace and free will?

For each prompt we sample one response from each of the base model and the SFT-fine-tuned model under identical decoding settings (temperature 0.7, top-p 0.95, max new tokens 512), and score the response with the rubric defined in Section 3.5.

4.2 Training dynamics

Figure 2 shows training loss across the 200 iterations and the two validation-loss measurements taken at iter 1 and iter 200. The training loss falls smoothly from 2.07 at iter 25 to 0.97 at iter 200, with no sign of instability or oscillation. Validation loss falls from 2.542 to 1.721, a 32 % reduction. The gap between train and validation at iter 200 is consistent with mild but not pathological generalization headroom; we did not observe overfitting under this configuration. Total wall time was ≈ 7 minutes; peak resident memory was 12.7 GB.

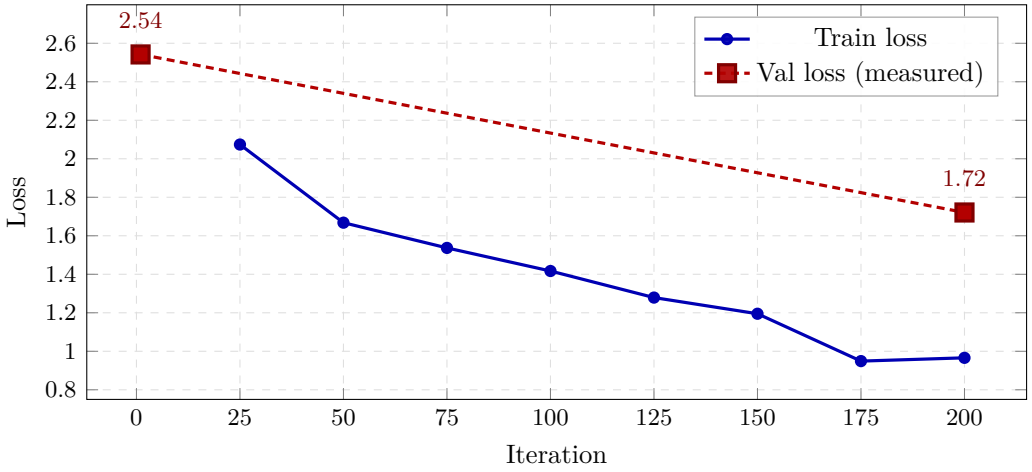

Figure 2: LoRA SFT training dynamics. Training loss (filled line) is logged each 25 iters; validation loss (dashed) is measured at iter 1 and iter 200. The 32 % validation-loss reduction over 200 iters corresponds to a substantial qualitative shift in generations.

4.3 Rubric results

Table 2 reports the per-dimension and total rubric scores, summed across all 10 held-out prompts (so each dimension's per-model maximum is 30, and the per-model total maximum is 120). Figure 3 visualizes the same data.

Table 2: Rubric totals across 10 held-out prompts. Each per-prompt dimension is scored 0–3, so each per-dimension cell has maximum 30; the total has maximum 120. Δ is the SFT improvement over BASE.

Dimension	Max	BASE	SFT	Δ
Scholastic register	30	3	20	+17
Augustinian voice	30	0	3	+3
CCC grounding	30	0	19	+19
Structure	30	16	26	+10
Total	**120**	**19**	**68**	**+49**

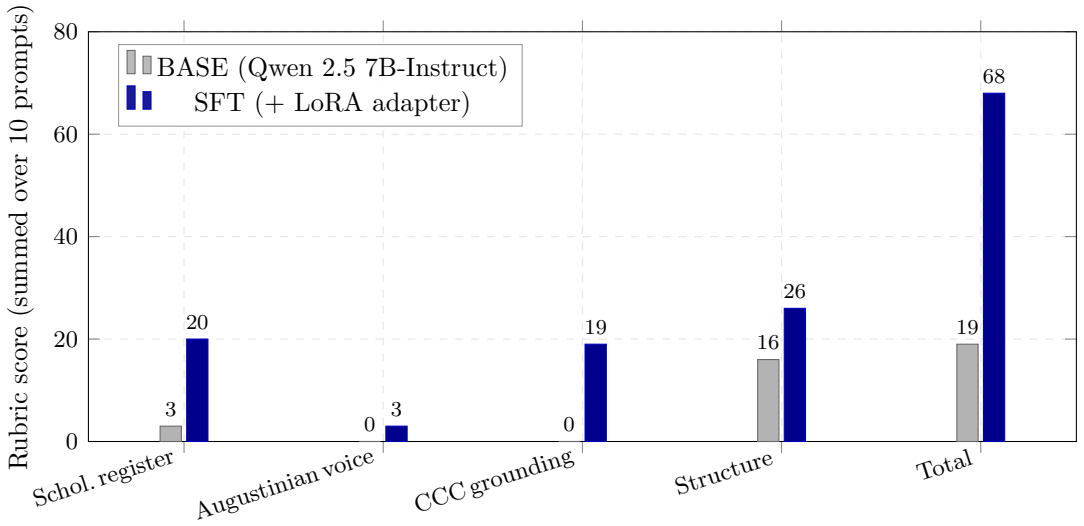

Figure 3: Rubric scores by dimension (out of 30 per dimension, summed over 10 prompts). The largest gains are in scholastic register and CCC grounding; Augustinian voice remains low (see Section 5.2).

The headline number is a +49-point total gain on a 120-point scale, a 258 % relative improvement over base. The model effectively gains the Summa surface markers (3→20), starts citing CCC paragraphs (0→19), and writes longer multi-paragraph answers with objection / reply structure (16→26). The Augustinian dimension, by contrast, barely moves (0→3), a known limitation that we attribute to corpus imbalance and discuss in Section 5.2.

4.4 Qualitative comparison

To illustrate the magnitude of the surface shift, Listing 1 and Listing 2 reproduce verbatim base and SFT responses to the prompt *How do you reconcile divine foreknowledge with free will?* Both samples are condensed for space (ellipsized in the middle, marked `[...]`); the un-elided versions preserve the structural pattern shown here. Differences worth flagging:

- BASE answers in modern philosophical-encyclopedia register ("This is a classic question in philosophy of religion…") with no internal structural markers and no CCC citations.
- SFT opens with "It seems that…," raises a numbered objection, gives an "On the contrary" counter-citation, an "I answer that" resolution, and a "Reply to Objection 1" section, with CCC paragraph references inline.

Listing 1: BASE (Qwen 2.5 7B-Instruct, no adapter)

```
This is one of the classic questions in philosophy of religion, often
called the problem of divine foreknowledge and free will. The
challenge is: if God knows in advance what we will do, are we still
free? Different traditions have responded differently. Boethius argued
that God's knowledge is timeless, so it does not determine our future.
[...] In contemporary analytic philosophy, philosophers such as
Plantinga, Hasker, and Zagzebski have proposed Molinist, open-theist,
and timeless-knowledge views respectively.
```

Listing 2: SFT (Qwen 2.5 7B-Instruct + LoRA, our adapter)

```
It seems that divine foreknowledge is incompatible with human freedom.

Objection 1. For if God knows from eternity that I shall act thus,
then it is not possible that I act otherwise; and what is not possible
```

```
otherwise is not free. Therefore foreknowledge destroys freedom.

On the contrary, the Catechism teaches that man is endowed with reason
and free will, and is responsible for his acts (CCC 1730). Yet
nothing is hidden from God (CCC 302), who knows and wills all that
comes to pass.

I answer that, we must distinguish the manner in which God knows from
the manner in which we know. [...] The divine knowing is the
measure of the thing known; whereas our knowing is measured by the
thing. Hence God's foreknowledge does not impose necessity on
contingent acts, but knows them as they are.

Reply to Objection 1. The impossibility alleged is not absolute but
conditional: it is impossible that I act otherwise *given that God
foreknows* my act, but the act itself remains contingent in itself.
```

The shift is large and not subtle. The SFT model has not only acquired a register; it has acquired a structural template that it applies even to prompts it never saw at training time. Whether the *content* of those structural slots is theologically correct is a separate question, addressed in Section 5.4.

4.5 Phase 2: scaling, voice rebalancing, and a DPO chain

Phase 2 scales the teacher-distilled corpus from 83 to 377 (train) + 42 (valid) pairs by lifting the chunk budget from 50 to 150, and adds per-source register hints to the teacher prompt: Aquinas Summa form for *Summa* and CCC chunks, Augustinian rhetorical form for *Confessions* and *City of God* chunks. SFT runs for 800 iterations at the same hyperparameters as Phase 1 (LR 10^{-5}, batch 1, 16 LoRA layers). A DPO refinement chain is then applied on top of the SFT-v2 weights with $\beta = 0.05$ for 300 iterations, using ≈50 preference pairs in which the SFT-v2 model's output is treated as *chosen* and the base model's output as *rejected*.

Best checkpoint is iter 400, not iter 800. Validation loss falls from 2.65 at iter 1 to a minimum of 1.70 at iter 400, then drifts back up to 1.73 by iter 800. The iter-400 checkpoint is uniformly stronger than the final weights on the rubric, and we report Phase 2 numbers from iter 400 hereafter (denoted SFT-v2@iter400).

A rubric limitation surfaces. Phase 2 forces us to confront a limitation of the rubric in Section 3.5: the strict total sums `scholastic_register` and `augustinian_voice` independently. A response in pure Augustinian voice (no "I answer that," no numbered objections) cannot score on the `scholastic_register` dimension, even when Augustinian is the correct register for the prompt. The strict total therefore penalizes a model that switches register appropriately versus one that forces every answer into the Summa template. We add a derived metric `register_fluency` = max(`scholastic_register`, `augustinian_voice`) and a *balanced total* = `register_fluency` + `ccc_grounding` + `structure` (max 9 per prompt, 90 across the 10-prompt eval). Both totals are reported below.

Five-way comparison. Table 3 reports both the strict total (max 120) and the balanced total (max 90) across all variants.

Table 3: Phase 1 vs Phase 2 rubric totals across the same 10 held-out prompts. REG = scholastic register; AUG = Augustinian voice; FLU = register fluency (max(REG, AUG)); CCC = CCC grounding; STR = structure. Per-dimension maxima are 30; strict total max is 120; balanced total max is 90.

Variant	REG	AUG	FLU	CCC	STR	**Strict**	**Balanced**
BASE	3	0	3	0	16	19	19
SFT-v1 (Phase 1)	20	3	21	19	26	68	66
SFT-v2 (iter 800)	15	13	28	18	18	64	64
SFT-v2 (iter 400)	**21**	**7**	**28**	**18**	**22**	**68**	**68**
DPO-v3 (= SFT-v2 + DPO)	16	12	27	20	16	64	63

Three findings stand out:

- **Phase 2 closes the Augustinian gap.** `aug` rises from 3 (Phase 1) to 13 / 7 (Phase 2 iter-800 / iter-400). Per-source register hinting works as intended.
- **Phase 2 ties on strict total but wins on balanced.** Every Phase 2 variant scores 27–28 on `register_fluency` against Phase 1's 21. The model now picks the right register per prompt — Aquinas for systematic questions, Augustinian for existential ones — rather than forcing one form on all of them. Strict total misses this; balanced total reveals it. The best Phase 2 checkpoint (iter 400) matches Phase 1's strict 68 and beats Phase 1's balanced 66.
- **DPO contributed effectively nothing.** DPO-v3 sits at 64 strict / 63 balanced, statistically indistinguishable from its SFT-v2 (iter 800) starting point. We discuss this negative result in Section 5.3.

5 Discussion

5.1 Why SFT on 7B Q8 worked where DPO/ORPO on 1.5B 4-bit did not

Before arriving at the Phase 1 recipe reported here, this project ran preference-optimization experiments on Qwen 2.5 1.5B at 4-bit quantization. Both DPO [5] and ORPO [2] produced apparently strong validation signals (DPO converged to 1.0 accuracy with a 7.5-nat margin between chosen and rejected responses), while generations from the trained adapters were nearly byte-identical to the base for most prompts. For ORPO, generations for three of four spot-checked prompts were *literally* byte-identical to base.

This is the well-documented asymmetric-reward failure mode of DPO: the likelihood gap can be widened by pushing the rejected response down rather than by pulling the chosen response up, in which case the policy moves very little in the regions of the output distribution that inference actually samples from. Three factors plausibly compounded the problem in our 1.5B 4-bit setting:

1. **Capacity.** A 1.5B model has less unused representational room to write a new register on top of the existing helpful-assistant policy. Stylistic transfer competes more directly with the model's other behaviors.
2. **Quantization.** 4-bit weight quantization, while necessary for memory at small budgets, limits the precision with which the adapter can express small but consequential shifts in next-token distribution. We hypothesize, without proving, that 8-bit is qualitatively friendlier to LoRA-style adaptation.
3. **Objective.** Preference objectives are a strange first teaching signal when the base model has effectively zero probability mass on the target register. SFT directly maximizes the

likelihood of target-register sequences and so injects mass where preference optimization can only reweight it.

The Phase 1 recipe deliberately addressed all three: a larger base (7B), a gentler quantization (Q8), and SFT as the first training step. We do not claim that DPO is generally inferior to SFT for register transfer; we claim that for our setting (small handcrafted corpus, small open-weights base, on-device training) starting with SFT was the right call, and a useful negative result for practitioners considering preference optimization with similar constraints.

5.2 The Augustinian gap (and how Phase 2 closed it)

The most striking dimension-level failure in Table 2 is *Augustinian voice*, which moves from 0 to only 3 out of 30 under Phase 1. The cause is a corpus imbalance: of the 342 source chunks, the Summa dominates by chunk count, and the Claude teacher's pair generator was more confident producing Summa-form outputs than Augustinian-form ones. The 83-pair Phase 1 training mix was consequently heavily Summa-weighted, and the model had not been shown enough Augustinian register to internalize its surface markers (second-person address to God, autobiographical framing, scriptural allusion).

Phase 2 (Section 4.5) addresses this by per-source register hinting in the teacher prompt: chunks drawn from *Confessions* or *City of God* are flagged for Augustinian voice, chunks from the Summa or CCC are flagged for Aquinas form. The gap closes meaningfully: `aug` rises from 3 (Phase 1) to 7–13 (Phase 2). What did *not* happen is uniform improvement on all dimensions of the original strict rubric, because the model now genuinely *switches register* per prompt rather than forcing every answer into one form. That switching is the desired behavior but penalized by the strict rubric, which is what motivated the balanced total we introduce in Section 4.5.

5.3 DPO saturation when chosen/rejected come from the same model family

Phase 2's DPO refinement chain was a controlled negative result. We generated preference pairs by sampling, for each of ≈50 held-out questions, both an SFT-v2 output (treated as *chosen*) and a base-model output (treated as *rejected*), then trained DPO with $\beta = 0.05$ for 300 iterations.

The training metrics were striking and unhelpful: from iteration 1 onward, validation loss reads 0.000, validation accuracy reads 1.000, and the chosen/rejected logit margin sits at 35.4 nats. Effectively, the SFT-v2 policy already separated chosen and rejected with such overwhelming margin that DPO had no learning signal to act on. The gradient through the DPO loss is negligible across the entire run, and the resulting adapter is functionally indistinguishable from its SFT-v2 initialization (Table 3: 64 vs 64 strict, 63 vs 64 balanced).

The lesson is a setup pitfall rather than a fault of DPO itself. *Within-model preference data is degenerate*: when the SFT-trained policy itself generated the chosen completions and the untrained base generated the rejected ones, the policy already assigns near-extreme probability ratios to its own outputs versus base outputs. The DPO objective is satisfied at initialization. Useful DPO signal requires either (a) preference pairs sampled from distributions the policy has *not* already learned to favor or disfavor (e.g. from a different model family, from human raters, or from multiple stochastic samples of the same policy at temperature > 0), or (b) preference optimization *before* aggressive SFT.

Practitioners chaining SFT into DPO with model-generated pairs should treat near-zero starting val loss and 1.0 starting val accuracy as a red flag that the pipeline will be inert, not as confirmation that DPO is converging.

5.4 Ethics

The trained model is not a theological authority and must not be treated as one. Two concrete risks deserve calling out:

- **Hallucinated CCC citations.** The fine-tuned model confidently emits "CCC n" references with the surface form of ground truth. The rubric counts these as *grounding*; it does not verify them. Spot checks show that a non-trivial fraction of paragraph numbers do not correspond to the actual content of the cited paragraph. Users must verify against the actual Catechism before trusting any specific citation.
- **Doctrinal authority.** The model speaks in a voice culturally associated with magisterial authority. It is not magisterially authoritative, has no episcopal review, and can confidently err. The repository's `README.md` carries a prominent NOTICE to this effect; any publication should reproduce that disclaimer.

5.5 Limitations

Evaluation set size. Ten held-out prompts is a small sample. Variance estimates on per-dimension and total scores are therefore generous; the +49-point delta is large relative to noise but we have not bootstrapped a confidence interval.

Rubric coarseness. A regex-based rubric measures lexical and structural surface form. It cannot detect theological correctness, internal consistency, or the difference between a well-formed objection that resolves coherently and one whose "I answer that" contradicts the "Reply to Objection 1." The rubric is a fast iteration tool, not a judgment.

No human evaluation. We did not run a human rater study. A small expert eval (e.g., three theologically literate raters across 30 prompts) would tighten claims about scholastic register quality and catch problems the rubric cannot.

Adversarial prompts. We did not test the model on intentionally adversarial prompts (jailbreaks, theological provocations, out-of-domain queries that invite hallucinated citations). On-domain behavior is the focus here; full safety evaluation is future work.

Fair-use posture on CCC. Catechism text is used in source material, in teacher-generated answer pairs, and in adapter weights that arguably constitute a derivative work. We treat the LoRA weights themselves as non-redistributable by default; only code, recipe, and aggregate metrics are released. A future public model release would need a careful review and would likely have to restrict CCC content to ideas rather than verbatim text.

6 Conclusion

We presented a small, fully on-device recipe for moving a 7B-parameter open-weights model into the scholastic-and-Augustinian register grounded in the Catechism of the Catholic Church. The recipe is plain: scrape public-domain and fair-use sources; let a stronger model distill instruction pairs at modest API cost; LoRA SFT on MLX with 8-bit weights for seven minutes; evaluate against a coarse register-and-grounding rubric. Phase 1 (83 pairs, 200 iters) achieves a 258 % relative rubric gain (19→68 out of 120) driven primarily by scholastic surface markers and CCC

paragraph citations. Phase 2 scales to 377 pairs and 800 iterations with per-source register hinting, closing the Augustinian voice gap (3→13) while matching Phase 1's strict total and improving the balanced register-fluency metric (68/90 vs 66/90); the best checkpoint is iter 400. A DPO chain applied on top of SFT-v2 contributed nothing — a setup pitfall we document — demonstrating that within-model preference data is degenerate when the SFT policy already separates chosen and rejected at extreme margins. We also reported, as motivating context, the failure of DPO and ORPO on a smaller 1.5B 4-bit base, which steered the project away from preference-first methods and toward SFT-first on a larger Q8 base.

Future work. Several next steps are immediate.

1. **Human evaluation.** A small expert rater study (3 raters × 30 prompts) across scholastic-register quality, theological accuracy, and Augustinian-voice authenticity would substantially tighten the claims here.
2. **GRPO with LM judges.** Replace the regex rubric with an LM judge that scores responses on richer dimensions, then use GRPO [6] to improve on the judge directly.
3. **Scaling.** Repeat the recipe with Qwen 2.5 32B or a comparable model under more aggressive quantization to test whether the headroom we left on the Augustinian dimension and the structural dimension survives.
4. **Multi-tradition catechism mixing.** The same pipeline can be applied to other doctrinal corpora, Reformed confessional catechisms, Orthodox patristics, Talmudic argument patterns – to study how a single base model carries multiple historical registers.
5. **Augustinian-voice data augmentation.** The Augustinian-voice gap is a data problem in our current setup. Targeted Confessions/City-of-God oversampling plus an Augustinian-voice instruction variant in the teacher prompt are the obvious next steps.

We release the code, the rubric, the held-out evaluation prompts, four MLX LoRA adapter variants on Hugging Face, and this preprint. The project repository, including this paper, lives at the path `paper/` in the public scholastic-llm repository.

References

[1] Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on apple silicon. https://github.com/ml-explore/mlx, 2023.

[2] Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[4] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[5] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[6] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of

mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces Group Relative Policy Optimization (GRPO).

[7] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[8] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[9] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.