Specializing a Small Open Model to Surpass a Frontier LLM on Norwegian Drug-Reimbursement Reasoning: A Proof of Concept

Made by Andreas Grønbeck | Co-Founder at tenki

2026-05-28

Abstract

Frontier large language models (LLMs) achieve strong general performance, yet their knowledge of narrow, jurisdiction-specific regulatory domains is bounded by training-data coverage and knowledge cutoffs. This paper reports a proof-of-concept experiment in which a small open-weight model, Qwen3-8B, was specialized for a single Norwegian administrative-law domain: the blue-prescription scheme (blåreseptordningen) that governs drug reimbursement. Using 4-bit quantized low-rank adaptation (QLoRA) on 197 curated instruction examples, the model was trained for 400 optimizer steps on a single consumer GPU (NVIDIA RTX 3060 Ti, 8 GB VRAM). The resulting specialized model is referred to as Nor-Med. On a 42-task domain benchmark, Nor-Med scored 79.5 percent, compared with 68.4 percent for Claude Sonnet 4.6 answering the same questions in a cold session, an absolute improvement of 11.1 percentage points. A retrieval-augmented variant of the same model scored 77.1 percent, indicating that retrieval degraded rather than improved performance once the relevant facts were encoded in the adapter weights. The specialized model runs locally at a fraction of the per-query cost of a hosted frontier model. The result does not imply that the smaller model is more capable in general. It demonstrates that targeted specialization can close, and locally reverse, a capability gap on a specific task at very low cost. Limitations, including the self-authored benchmark and the weakness of keyword-overlap evaluation, are discussed in detail.

Keywords

domain specialization; QLoRA; parameter-efficient fine-tuning; Norwegian language models; healthcare administration; model sovereignty; retrieval-augmented generation

1. Introduction

General-purpose large language models are trained to perform acceptably across a very wide distribution of tasks. This breadth is purchased at a cost: for any sufficiently narrow domain, a general model allocates only a small fraction of its capacity and training signal to the facts and conventions of that domain. When the domain is also subject to frequent regulatory change, two further problems compound. First, the model's knowledge is fixed at its training cutoff and cannot reflect subsequent amendments. Second, jurisdiction-specific rules are often counter-intuitive and poorly represented in the predominantly English, internationally sourced pretraining corpus.

Norwegian drug-reimbursement law provides a concrete instance of this problem. The blue-prescription scheme determines which medicines the state reimburses, under which diagnostic conditions, through which application route, and at what patient co-payment. The rules are codified in the blue-prescription regulation (Lovdata, 2007/2025) and administered by the Norwegian Health Economics Administration, Helfo (Helfo, 2026). The scheme is amended regularly; several provisions changed in 2025 and 2026.

This paper asks a narrow, falsifiable question: can a small open-weight model, specialized cheaply on a single consumer GPU, outperform a frontier model on this domain? The contribution is threefold. First, a 42-task benchmark for the domain is constructed and a frontier baseline is measured under a

cold-session protocol. Second, a specialization pipeline based on QLoRA is described and applied to Qwen3-8B. Third, the specialized model is evaluated against the baseline, and the effect of retrieval augmentation is measured. The remainder of the paper is organized as follows. Section 2 reviews the domain and the relevant methods. Section 3 describes the benchmark, the training procedure, and the evaluation protocol. Section 4 reports results. Section 5 discusses interpretation, limitations, and implications for technological sovereignty. Section 6 concludes.

2. Background and Related Work

2.1 The blue-prescription scheme

The blue-prescription regulation establishes three principal reimbursement routes. Section 2 governs pre-approved reimbursement, under which a physician may prescribe directly when the diagnosis and any additional conditions are met. Section 3 governs reimbursement by individual application, under which a physician applies to Helfo on the patient's behalf for medicines not on the pre-approval list. Section 4 governs medicines for communicable diseases, with entitlement extended to all persons present in Norway irrespective of national-insurance membership (Helfo, 2026; Lovdata, 2007/2025).

Several provisions relevant to this study changed after the knowledge cutoff of most contemporary frontier models. The former Section 3a was replaced by a revised Section 3 effective 17 June 2025, and from the same date all new individual-reimbursement decisions are time-limited to a maximum of two years (Helfo, 2026). Public financing of the antiviral nirmatrelvir/ritonavir (Paxlovid) ended on 1 April 2025. From 1 January 2026 the patient co-payment on blue prescriptions is 60 percent of the price, capped at 400 Norwegian kroner per dispensation, replacing the earlier rate. The Norwegian Medicines Agency was renamed the Norwegian Directorate for Medical Products in 2024. These facts are representative of information that a model with an earlier cutoff cannot possess.

2.2 Parameter-efficient fine-tuning

Low-rank adaptation (LoRA) freezes the weights of a pretrained model and learns a small number of low-rank update matrices, reducing the number of trainable parameters by several orders of magnitude (Hu et al., 2021). QLoRA extends this approach by quantizing the frozen base model to 4 bits using the NormalFloat (NF4) data type, with double quantization and paged optimizers, enabling fine-tuning of multi-billion-parameter models on a single consumer GPU without measurable loss of adaptation quality (Dettmers et al., 2023). These methods make domain specialization economically accessible to individuals and small organizations.

2.3 Specialization versus scale

A substantial body of practice holds that retrieval-augmented generation (RAG) is the preferred method for injecting domain or recency knowledge into a general model without fine-tuning. RAG retrieves relevant passages at inference time and conditions generation on them. An alternative is to encode the domain directly into adapter weights. The two approaches are not mutually exclusive, and their relative merit is task-dependent. This study measures both for the same specialized model and finds, for this domain, that they interact unfavorably. The finding is consistent with reports that noisy

retrieval can override correct parametric knowledge.

3. Methods

3.1 Benchmark construction

A benchmark of 42 tasks was authored within the TenkiBench framework under the category norwegian-healthcare/blaa-resept. Each task specifies a natural-language user prompt, a gold answer, and a machine-checkable evaluation rule. Tasks are labeled by difficulty: 14 easy, 21 medium, and 7 hard. Two evaluation methods are used. The regex_all method awards partial credit equal to the fraction of a set of regular-expression patterns that match the model output; it is used for tasks with verifiable factual content such as dates, paragraph references, rates, and diagnostic codes. The llm_judge method awards weighted partial credit based on the fraction of salient terms from each rubric criterion that appear in the output; it is used for open-ended procedural questions. The limitations of the latter method are addressed in Section 5.

The benchmark was designed to include tasks that stress knowledge unlikely to be held by a general model: post-cutoff regulatory facts, precise procedural details, and counter-intuitive rules where the correct answer contradicts a plausible default.

3.2 Baseline protocol

To obtain an unbiased frontier baseline, the 42 user prompts were presented to Claude Sonnet 4.6 in a separate session with no access to the gold answers, the evaluation rules, or any discussion of the task design. The resulting answers were scored with the same evaluation code used for the specialized model. This cold-session protocol avoids the inflation that occurs when the same agent that designed the benchmark also produces the answers; an earlier within-session attempt produced a misleadingly high 94 percent and was discarded.

3.3 Model and training

The base model was Qwen3-8B, an 8-billion-parameter open-weight model released in 2025 (Qwen Team, 2025). It was loaded in 4-bit NF4 precision with double quantization. LoRA adapters of rank 32 (alpha 64) were attached to all seven attention and feed-forward projection modules, yielding 87.3 million trainable parameters, or 1.05 percent of the total. Training used the paged 8-bit AdamW optimizer at a learning rate of 2 x 10^-4 with cosine decay, a per-device batch size of 1, gradient accumulation of 8, and a maximum sequence length of 512 tokens. The model was trained for 400 optimizer steps. The final training loss was 0.20. The resulting adapter, applied to the Qwen3-8B base, constitutes the model referred to throughout this paper as Nor-Med.

The training set comprised 197 instruction examples in chat format. Examples paired domain questions with concise, factually correct Norwegian answers covering the reimbursement routes, the post-cutoff regulatory changes, diagnostic conditions, and procedural rules. A subset of examples included relevant source passages in the system prompt to support retrieval-style conditioning.

3.4 Hardware and cost

All training and inference were performed on a single NVIDIA RTX 3060 Ti with 8 GB of video memory in a consumer desktop. Training required approximately 2.8 hours after memory configuration was tuned; an initial configuration that padded all sequences to maximum length and kept optimizer states resident exhausted video memory and was corrected by dynamic padding, a paged optimizer, and a reduced sequence length. Inference for the full 42-task benchmark completed in approximately 11 minutes without retrieval and approximately 35 minutes with retrieval.

3.5 Retrieval configuration

The retrieval-augmented variant used a keyword-based BM25 index over five domain source documents, retrieving the top five passages per query and injecting up to 2000 characters into the system prompt. The same trained adapter was used for both the retrieval and non-retrieval conditions; only the inference-time context differed.

4. Results

4.1 Aggregate performance

Table 1 summarizes aggregate scores. The specialized model without retrieval scored 79.5 percent, exceeding the Claude Sonnet 4.6 cold-session baseline of 68.4 percent by 11.1 percentage points. The retrieval-augmented variant scored 77.1 percent, exceeding the baseline by 8.7 percentage points but underperforming the non-retrieval variant by 2.4 points.

Table 1. Aggregate benchmark scores (42 tasks).

System	Score	vs. baseline
Claude Sonnet 4.6 (cold session)	68.4%	reference
Nor-Med (Qwen3-8B QLoRA), no retrieval	79.5%	+11.1 pp
Nor-Med (Qwen3-8B QLoRA), with retrieval	77.1%	+8.7 pp

4.2 Performance by difficulty

The non-retrieval model improved on the baseline for both easy and medium tasks, reaching 91.7 percent and 81.7 percent respectively. On hard tasks, which are evaluated almost entirely by the keyword-overlap method, all three systems scored below 60 percent; the retrieval variant scored highest at 56.0 percent, indicating that retrieval helped most on open-ended procedural items. The hard-task scores are noisy and should be interpreted with the caution discussed in Section 5.3.

Table 2. Scores by difficulty.

Difficulty (n)	Claude	Nor-Med (no RAG)	Nor-Med (RAG)
easy (14)	78.6%	91.7%	78.6%
medium (21)	69.4%	81.7%	83.0%
hard (7)	45.3%	48.2%	56.0%

4.3 The specialization gap

The baseline model failed systematically on post-cutoff and counter-intuitive items, while the specialized model answered them correctly. Representative cases are shown in Table 3. On the Paxlovid financing-end date, the baseline gave November 2023; the correct answer is 1 April 2025. On the replacement of Section 3a, the baseline gave February 2024; the correct date is 17 June 2025. On the patient co-payment, the baseline gave the superseded 2025 rate; the correct 2026 rate is 60 percent capped at 400 kroner.

Table 3. Selected items where the baseline failed and the specialized model succeeded.

Item	Baseline answer	Correct answer
Paxlovid financing end	November 2023	1 April 2025
Section 3a replaced	February 2024	17 June 2025
Max decision duration	three years	two years
2026 co-payment	39% / 790 kr	60% / 400 kr
Cancer-pain opioids	family doctor may apply	specialist required

4.4 Effect of retrieval

Retrieval improved some procedural items but degraded several factual items that the non-retrieval model answered correctly from parametric knowledge. On the co-payment item, retrieval injected a passage referencing the superseded rate, and the model reproduced it. On the prescriber-eligibility item, retrieval surfaced unrelated passages and the model answered a different question. The net effect across the benchmark was negative, as reflected in the 2.4-point gap between the two variants. A full task-by-task comparison is provided in Appendix A.

4.5 Cost

The specialized model runs locally on consumer hardware, incurring only electricity cost at inference. A hosted deployment of an 8-billion-parameter open model is priced on the order of one to two tenths of a US dollar per million tokens by commodity providers, against approximately three US dollars per million input tokens and fifteen per million output tokens for the frontier baseline. On an API-to-API basis this corresponds to a cost reduction of roughly one to two orders of magnitude; for local execution the marginal per-query cost approaches zero.

5. Discussion

5.1 Interpretation

The result supports a specific and limited claim: for a narrow, well-defined domain, a small open model specialized at low cost can exceed a frontier model on a task-aligned benchmark. The mechanism is not mysterious. The benchmark rewards knowledge of Norwegian regulatory facts, several of which postdate the baseline model's training cutoff or are underrepresented in its corpus. The specialized model was taught those facts directly. The frontier model, despite far greater general

capability, cannot answer what it was never trained on and cannot know what changed after its cutoff.

This is the central practical lesson. Value in applied language modeling often derives less from raw model capability than from alignment between the model's knowledge and the specific task. A smaller model whose knowledge matches the task can outperform a larger model whose knowledge does not.

5.2 Why retrieval degraded performance

The finding that retrieval reduced accuracy merits attention because it contradicts a common default. Once the relevant facts were encoded in the adapter weights, keyword-based retrieval added noise rather than signal. Retrieved passages sometimes contained superseded figures or addressed adjacent topics, and the model, conditioned to attend to its context, reproduced the retrieved content in preference to its parametric knowledge. This does not indict retrieval in general. It indicates that retrieval and parametric specialization must be designed together, and that retrieval quality, not merely its presence, determines its value.

5.3 Limitations

Several limitations bound the strength of the claim. First, the benchmark was authored by the same party that produced the training data. Although the training and evaluation items are distinct, both draw on the same source documents, and the alignment between them is higher than would obtain for an independently constructed benchmark. The result should therefore be read as a proof of concept, not as a general measurement of relative capability. Second, the llm_judge evaluation method rewards lexical overlap with rubric terms rather than semantic correctness; it can both under-credit correct paraphrases and over-credit keyword-matching text. The hard-task scores in particular should be treated as noisy. Third, the baseline reflects a single cold session with one frontier model and was not repeated across samples or temperatures. Fourth, the model has not been evaluated on real patient cases, adversarial inputs, or safety-critical failure modes, and is not fit for clinical or administrative deployment without substantial further work. Fifth, the experiment does not measure factual accuracy beyond what the evaluation patterns capture; a model could match a pattern while surrounding it with an incorrect explanation.

5.4 Implications for technological sovereignty

The broader implication concerns how organizations acquire value from artificial intelligence. Purchasing access to frontier models, without investment in adapting them to specific tasks, treats capability as a commodity that will deliver value on contact. This experiment suggests the opposite ordering. Value follows from alignment between a model and a task, and that alignment must be built. For a small country with its own language, legal system, and administrative conventions, the capacity to specialize open models locally is also a matter of sovereignty: it reduces dependence on external providers for domain-critical knowledge work and keeps both the data and the resulting capability under domestic control (see also Kode24, 2024).

6. Conclusion

A small open model, specialized for a single Norwegian regulatory domain using 4-bit QLoRA on 197 examples and a single consumer GPU, outperformed a frontier model on a 42-task domain benchmark by 11.1 percentage points, at a fraction of the inference cost. Retrieval augmentation, applied to the same specialized model, reduced accuracy, indicating that parametric specialization and retrieval must be co-designed. The result is a proof of concept rather than a general capability comparison, and the limitations of the benchmark and evaluation are real. Within those bounds, the experiment supports a practical thesis: for narrow, language- and jurisdiction-specific tasks, targeted specialization of small open models is an economically accessible path to performance that can exceed that of far larger general systems. Future work should construct an independently authored benchmark, evaluate on real cases with expert review, repeat the baseline across samples and competing frontier models, and test co-designed retrieval that defers to parametric knowledge when retrieval confidence is low.

References

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In Advances in Neural Information Processing Systems (Vol. 36).

Helfo. (2026). Blåreseptordningen og legemidler på blå resept (forhåndsgodkjent og individuell stønad). Helseøkonomiforvaltningen. https://www.helfo.no/lege/blaareseptordningen

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv:2106.09685.

Kode24. (2024). Gode grunner til å utvikle norske modeller, ikke minst suverenitet. https://www.kode24.no/artikkel/gode-grunner-til-a-utvikle-norske-modeller-ikke-minst-suverenitet/260931

Lovdata. (2007/2025). Forskrift om stønad til dekning av utgifter til viktige legemidler mv. (blåreseptforskriften). https://lovdata.no/dokument/SF/forskrift/2007-06-28-814

Qwen Team. (2025). Qwen3 technical report. Alibaba Group.

Appendix A. Task-by-task scores

Task	Difficulty	Method	Claude	Nor-Med (no RAG)	Nor-Med (RAG)
br-001	easy	regex_all	1.00	1.00	1.00
br-002	medium	regex_all	0.67	1.00	0.67
br-003	medium	regex_all	0.67	1.00	0.33
br-004	easy	regex_all	0.67	1.00	1.00
br-005	easy	regex_all	1.00	0.67	1.00
br-006	medium	regex_all	1.00	1.00	0.67
br-007	easy	regex_all	1.00	1.00	1.00
br-008	medium	llm_judge	0.51	0.46	0.45
br-009	easy	regex_all	0.67	1.00	0.33

Task	Difficulty	Method	Claude	Nor-Med (no RAG)	Nor-Med (RAG)
br-010	medium	llm_judge	0.39	0.87	0.65
br-011	medium	regex_all	1.00	1.00	1.00
br-012	easy	regex_all	1.00	1.00	1.00
br-013	medium	regex_all	0.50	1.00	1.00
br-014	easy	regex_all	1.00	1.00	1.00
br-015	hard	llm_judge	0.59	0.52	0.47
br-016	hard	llm_judge	0.56	0.80	0.80
br-017	medium	regex_all	0.67	1.00	1.00
br-018	medium	regex_all	1.00	1.00	1.00
br-019	easy	regex_all	1.00	1.00	1.00
br-020	hard	llm_judge	0.51	0.42	0.54
br-021	medium	regex_all	1.00	0.67	1.00
br-022	easy	regex_all	0.33	0.67	0.33
br-023	medium	regex_all	1.00	1.00	1.00
br-024	medium	regex_all	0.67	0.67	0.67
br-025	hard	llm_judge	0.47	0.57	0.40
br-026	easy	regex_all	0.33	1.00	0.33
br-027	medium	regex_all	0.33	1.00	1.00
br-028	easy	regex_all	0.50	1.00	1.00
br-029	easy	regex_all	0.50	1.00	0.50
br-030	medium	regex_all	0.33	0.00	0.67
br-031	medium	regex_all	0.50	0.50	1.00
br-032	medium	regex_all	1.00	1.00	1.00
br-033	medium	regex_all	0.67	0.67	0.67
br-034	medium	regex_all	0.67	1.00	1.00
br-035	hard	llm_judge	0.33	0.21	0.45
br-036	hard	llm_judge	0.49	0.55	0.66
br-037	medium	regex_all	0.67	0.67	0.67
br-038	hard	llm_judge	0.22	0.31	0.61
br-039	easy	regex_all	1.00	1.00	1.00
br-040	easy	regex_all	1.00	0.50	0.50
br-041	medium	regex_all	0.67	0.67	1.00
br-042	medium	regex_all	0.67	1.00	1.00
Mean			**0.684**	**0.795**	**0.771**