Smart Design Policies

2026, Volume 3, Number 1, pages 14–28

Original scientific paper

Spyra-20B: A Proof-of-Concept for Explicit Architectural Design Reasoning in Domain-Specialised LLMs

[*1] **Nik Ansre** , [2] **Yara Hirsekorn** , [3] **Gregor Grunwald**

[1, 2, & 3] Department of Architecture, Jade University of Applied Sciences Oldenburg, Germany

[1] E-mail: nik.ansre@student.jade-hs.de , [2] E-mail: yara.hirsekorn@student.jade-hs.de , [3] E-mail: gregor.grunwald@jade-hs.de

[1] **ORCID**: https://orcid.org/0009-0006-4080-4398 , [2] **ORCID**: https://orcid.org/0009-0001-7275-2450 , [3] **ORCID**: https://orcid.org/0000-0003-2437-398X

ARTICLE INFO:

Received: 23 May 2026
Revised: 28 July 2026
Accepted: 3 August 2026
Available online: 15 August 2026

Keywords:
AI,
Architecture,
Large Language Models,
Tree-of-Thought,
Chain-of-Thought,
Design Reasoning.

ABSTRACT

Domain-specifically fine-tuned language models (BloombergGPT, Med-PaLM) have proven a viable route to adapting vocabulary and specialist knowledge to particular fields of application. For architectural practice, however, the central challenge is not terminology but the deliberative structure of the design decision: the weighing of competing objectives across building law, quality of use, cost and sustainability. This paper presents Spyra-20B, a proof-of-concept produced by QLoRA fine-tuning of gpt-oss-20B on 1,693 architecture- and planning-specific dialogue examples. Two features characterise the approach: a three-level deliberation depth controllable by metadata (reasoning_effort), and a strict separation of analysis and answer channels that renders the deliberative process inspectable. An evaluation on 50 law- and planning-related items across four conditions (base model vs. Spyra-20B, closed-book vs. open-book with RAG; 4,000 requests) yields a negative headline result: the fine-tuned model attains roughly six percentage points lower answer accuracy (Spyra 61.1% / 70.4%, base 67.3% / 76.3%; $p < 0.001$), while retrieval improves both models by about nine percentage points. Fine-tuning changes the form of the output: more schematic, twice as fast at the median, but with a sixfold increase in unrecoverable format failures. We derive four governance requirements: reasoning transparency, external factual grounding, local executability and multi-sample verification, and conclude that such systems should be assessed along architectural properties rather than scores on individual benchmarks.

SMART DESIGN POLICIES (2026), ***3***(1), 14–28.
https://doi.org/10.38027/smart.v3n1-2

www.smartdpj.com

* Corresponding Author

How to cite this article: (APA Style)
Ansre, N., Hirsekorn, Y., & Grunwald, G. (2026). Spyra-20B: A proof-of-concept for explicit architectural design reasoning in domain-specialised LLMs. *Smart Design Policies, 3*(1), 14–28.
https://doi.org/10.38027/smart.v3n1-2

1. Introduction

The integration of large language models (LLMs) into the planning professions of the built environment has evolved from an experimental research topic into an increasingly practical field of application. Although generic LLMs were initially conceived as broadly applicable text-generation systems, highly specialized domains such as law, medicine, finance, and engineering have increasingly adopted domain-specific models tailored to their particular terminology, knowledge structures, and professional requirements. BloombergGPT (Wu et al., 2023) and Med-PaLM (Singhal et al., 2023) represent prominent examples of this development. Their significance lies not merely in improvements in general model performance but in the adaptation of model behavior, vocabulary, and knowledge representation to specific professional contexts. This shift suggests that the

effectiveness of an LLM in specialized applications depends not only on model scale but also on its alignment with the epistemic and procedural characteristics of the target domain.

Within Architecture, Engineering, and Construction (AEC), such alignment is particularly challenging. Architectural and planning decisions rarely involve a single objectively correct solution. Instead, they require deliberation among multiple and frequently conflicting considerations, including building regulations, functional quality, cost, sustainability, heritage conservation, and contextual compatibility. Consequently, access to factual knowledge alone is insufficient. For example, retrieving the content of a planning regulation such as § 34 BauGB may provide an essential legal basis, but professional reasoning requires the interpretation of that regulation in relation to a specific site, the comparison of alternative design responses, and the justification of a selected course of action. Current approaches address parts of this challenge from two principal directions. Retrieval-augmented systems improve access to external knowledge and factual information (Lewis et al., 2020), whereas reasoning-oriented prompting strategies, including Chain-of-Thought (CoT; Wei et al., 2022) and Tree-of-Thought (ToT; Yao et al., 2023), seek to structure the model's deliberative process. However, both approaches generally operate without modifying the underlying model weights and therefore do not necessarily internalize the desired reasoning structure as a stable characteristic of model behavior.

The present study builds on a broader research project at Jade University of Applied Sciences Oldenburg investigating locally executable language models for AEC applications. Previous work by the authors examined the integration of domain-specific expertise with dialogue-oriented retrieval systems (Hirsekorn et al., 2025), providing the technical foundation for the present investigation. Whereas that earlier work focused primarily on extending access to specialized knowledge, the present study addresses a distinct question: whether the structure of professional deliberation itself can be technically embedded within a domain-specialized model.

A significant research gap emerges from the current literature. The structure of architectural reasoning—including the sequencing of arguments, explicit consideration of competing alternatives, separation of premises from conclusions, and adjustment of deliberation depth—is predominantly treated as a prompt-engineering problem rather than as an explicit training objective. CoT and ToT approaches can encourage structured reasoning, but their effectiveness remains strongly dependent on prompt formulation and does not guarantee a consistent or inspectable output structure. Conversely, domain-specific training approaches demonstrate that specialized vocabulary and knowledge can be internalized within model parameters, yet they provide limited evidence regarding whether the *form* of deliberative reasoning can likewise be embedded in model behavior. The MAKER framework (Meyerson et al., 2025), which introduces graded levels of cognitive effort through prompting, is conceptually relevant in this regard; however, the possibility of transferring such graded reasoning control into a trainable signal within a domain-specialized model remains insufficiently explored.

This research gap has both technical and governance dimensions. From a technical perspective, it remains unclear whether fine-tuning on curated reasoning traces can produce a model in which deliberation depth, alternative reasoning paths, and the separation between analysis and final response become systematically controllable and inspectable. From a governance perspective, the ability to inspect model deliberation raises broader questions regarding transparency, factual grounding, data sovereignty, reproducibility, and the procurement and regulation of AI-supported planning services. These dimensions are analytically related but conceptually distinct: the governance implications are examined in relation to the observed properties of the trained system rather than assumed to follow automatically from technical performance.

Accordingly, this study addresses three research questions. First, can the deliberative structure of architectural reasoning be technically anchored by fine-tuning an open-weight language model on curated reasoning traces such that deliberation depth becomes controllable and the resulting process is inspectable in the model output? It is hypothesized that separating analysis and answer channels within the training format and introducing a metadata-controlled deliberation-depth parameter

(reasoning_effort) will produce outputs in which premises, intermediate reasoning steps, and rejected alternatives can be systematically identified. Second, how does the structural improvement resulting from fine-tuning relate to conventional answer accuracy in closed-form examination benchmarks? It is hypothesized that optimizing the model for reasoning structure will not necessarily improve closed-form answer accuracy because selecting the correct option from a predefined set is not itself an explicit training objective. Third, what governance requirements for AI applications in the construction sector can be derived from observed characteristics such as inspectable deliberation, numerical hallucination, external retrieval grounding, and run-to-run instability? It is hypothesized that these findings support an architectural approach to AI governance emphasizing transparency of deliberation, external factual grounding, local executability, and multi-sample verification rather than relying exclusively on aggregate benchmark performance.

To investigate these questions, the study introduces Spyra-20B, a domain-specialized open-weight model developed through QLoRA fine-tuning of gpt-oss-20B (OpenAI, 2025) using 1,693 curated dialogue examples. The proposed approach introduces a graded, metadata-controlled deliberation signal through reasoning_effort and a training structure designed to separate deliberation from final answers at the level of the optimization objective. The study therefore makes two principal contributions. First, it provides an empirical proof of concept for technically anchoring explicit deliberative structures within a domain-specialized architectural language model. Second, it examines the governance implications of model characteristics revealed through the experiments and formulates requirements concerning reasoning transparency, factual grounding, local deployment, and verification.

The empirical evaluation deliberately focuses on comparison with the immediate base model rather than with proprietary systems such as GPT-4 or independently trained models such as Llama. Direct comparisons with such systems would confound model scale, training corpora, architecture, and training procedures, making attribution of observed differences to the proposed fine-tuning strategy difficult. The comparison with the unmodified base model therefore provides the more informative experimental baseline. Importantly, the study does not assume that domain-specific fine-tuning must improve every performance measure. Indeed, the closed-form benchmark evaluation demonstrates a negative result for the fine-tuned model relative to its base model. This outcome is reported explicitly because it helps distinguish improvements in deliberative structure from improvements in conventional answer accuracy and provides an important basis for evaluating both the capabilities and limitations of domain-specialised reasoning models for architectural practice.

2. Materials and Methods

2.1 Study Design and Setting

This work is a proof of concept for a domain-specialised large language model that reconstructs architectural design reasoning explicitly and inspectably. Given limited hardware and the scope appropriate to a student research project, external validation by an expert panel was deliberately not undertaken; the study confines itself to a quantitative closed-form benchmark and a qualitative case analysis (Section 3), and situates its results as a preliminary stage for a broader evaluation.

The technical basis is the open-weight model gpt-oss-20B (OpenAI, 2025), with 20 billion parameters and a mixture-of-experts architecture of 32 experts. The MoE structure activates only a fraction of the parameters per inference, which makes execution on local hardware feasible, decisive for the data sovereignty argument in Section 4. Only a model whose weights load locally and whose behaviour can be ablated against the unmodified initial state permits the controlled attribution of effects to the fine-tuning intervention.

2.2 Materials and Equipment

Fine-tuning was carried out on a workstation with two NVIDIA RTX A6000 cards (48 GB VRAM each), a locally executed software stack (PyTorch with bfloat16, Hugging Face transformers and trl,

and Unsloth for memory-optimised execution) and the QLoRA method (Dettmers et al., 2023), which loads the base model in 4-bit quantisation and optimises only trainable low-rank adapters in full precision (bf16). With gradient checkpointing, this permits the specialisation of a 20-billion-parameter model on capable but non-datacentre hardware, of immediate relevance for small and medium-sized planning practices. The evaluation (Section 2.4) used the same infrastructure, with Ollama as the inference runtime, so that prompts were issued model-agnostically to the same endpoint definition and latency measurements are directly comparable.

2.3 Procedure and Protocols

2.3.1 Curation of the Training Dataset

The training corpus comprises 1,693 architecture- and planning-specific dialogue examples curated by the authors. Each is stored in the OpenAI Harmony message format (system, user and assistant roles) and carries an assistant-side thinking field containing the explicit reasoning process. The thematic distribution covers six main domains: the German Federal Building Code (BauGB), the Model Building Regulations (MBO, 40 examples), the German Construction Contract Procedures (VOB, Parts A, B and C), heritage protection law, urban design and typological design knowledge, and parametric design in Rhino/Grasshopper. Each example carries a meta.domain field for later filtering; the full breakdown is provided as supplementary material.

Two limitations should be disclosed: curation was performed by the first authors without independent second review, so no inter-rater reliability can be reported; and the topic selection follows the curriculum of a German bachelor's degree in Architecture and Urban Design, so it is not a representative cross-section of international planning practice. Both are taken up in Section 4.4.

2.3.2 Operational Definition of reasoning_effort

Each training example carries, in its metadata block, a reasoning_effort field with one of three values. At inference the value governs the depth of the reasoning process placed in the analysis channel, implemented through a leading XML tag <reason:{low|medium|huge}> in the assistant response. The levels are defined as follows.

low - direct factual queries with no need for inference or trade-off (e.g. "*Which separation distances apply in principle under § 6 MBO*?"). The <think> block is short, typically one to three sentences, and confined to naming the relevant norm and its direct application.

medium - standard tasks of planning-law subsumption or typological design, in which exactly one set of facts is examined and one resulting recommendation derived. The <think> block contains a sequential chain-of-thought derivation with explicit reference to the norm.

huge - multi-dimensional design or conflict scenarios with competing objectives (e.g. densification while respecting heritage, tree protection and separation distance requirements). The <think> block is structured as a tree of thought: at least two alternative paths are formulated in parallel, assessed against the same criteria (building law, cost, quality of use, sustainability) and brought together in a reasoned decision.

The procedure is inspired by the MAKER framework (Meyerson et al., 2025), but makes deliberation depth a trainable, metadata-controlled signal rather than a prompting-side device. The medium level dominates the data and is the default where metadata is absent; huge is reserved for design and conflict scenarios.

2.3.3 Two-Channel Training Format

The central methodological contribution is the strict technical separation of the reasoning process and answer. When a training example is rendered, the assistant response is constructed as a concatenated string of the form:

<reason:{effort}>
<think>
{thinking}

</think>
<answer>
{content}
</answer>

and serialised with the chat template of the base model. The DataCollatorForCompletionOnlyLM from the trl library receives "<answer>\n" as its response template and masks all tokens before this marker with the label −100, so the loss is defined exclusively on the content of the final channel. The analysis channel serves as conditional context for the autoregressive prediction of the answer, but is not itself optimised. This construction has conceptual consequences. The model does not learn to reproduce a particular reasoning process; it learns that good answer content emerges from a particular structural preamble. In the terminology of explainable AI, the analysis is therefore not a post-hoc rationalisation but a causally upstream representation anchored in training. It also means the model is not trained on the correctness of a particular answer option, a point that substantially shapes the results in Section 3.1.

2.3.4 QLoRA Configuration and Training Run

The complete fine-tuning hyperparameters are documented in Table 1. All values are taken from the training notebook and are unrounded.

Table 1: Hyperparameters of the QLoRA fine-tuning of Spyra-20B.

Category	**Parameter**	**Value**
Base model	model_name	unsloth/gpt-oss-20b-BF16
	Base precision	4-bit quantisation (loading), bf16 (LoRA adapters)
LoRA	Rank r	64
	alpha	128
	Dropout	0.0
	Bias	none
	Target modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Optimisation	Optimizer	paged_adamw_8bit
	Learning rate	2×10^{-4}
	LR scheduler	linear
	Warm-up steps	130
	Max. training steps	1,800
	Max. gradient norm	0.3
	Precision	bf16
Batch and sequence	Per-device batch	1
	Gradient accumulation	4
	Effective batch size	4
	Max. sequence length	4,096 tokens
Other	Gradient checkpointing	active
	Seed	42

With 1,693 examples and an effective batch size of 4, this corresponds to approximately 4.25 epochs. The answer section marked as the response template was verified on the first example by a sanity check on the number of tokens actually labelled, and the run was logged via TensorBoard and a custom PerfLogger callback (steps per minute, peak VRAM usage); the log history forms part of the supplementary material. Afterwards, the LoRA adapter from checkpoint 1,800 was merged with the base model in bf16 and converted to GGUF via llama.cpp for execution under Ollama. The resulting artefact is referred to as spyra-20b:latest throughout.

2.4 Evaluation: Benchmark Design

2.4.1 Task and Justification

To go beyond qualitative case demonstrations, a closed-form examination with 50 multiple-choice questions served as the benchmark: the examination for the course Planning and Construction Management 2 (Parts A and B) of the bachelor's programme in Architecture at Jade University of Applied Sciences Oldenburg. It comprises 30 single-choice and 20 multiple-answer items on German public procurement law (VOB/A, VOB/B), building regulations law (MBO), building planning law (BauGB, BauNVO) and floor area calculation under the WoFlV. Each item has a unique model solution derivable from the relevant primary sources.

One caveat should be stated at the outset: the fine-tuning targets architectural design reasoning, and the domains named lie, apart from the procurement questions, only partly within the focus of the training corpus. The benchmark therefore primarily tests transfer to related but non-identical domains, and behaviour under misleading context (see 2.4.2), rather than the design-specific target capability. We situate the dissociation in Section 4.

2.4.2 Conditions and Number of Runs

A 2 × 2 design was evaluated: the unmodified base model (gpt-oss:20b) against the fine-tuned model (spyra-20b:latest) on the model axis, and a closed-book condition (answering from parametric knowledge, as in the original examination) against an open-book condition with an upstream retrieval-augmented generation layer on the context axis.

The retrieval corpus comprises the full texts of VOB/A (2019 edition, Section 1) and VOB/B (2020 edition), segmented into 176 chunks. The embedding model was qwen3-embedding:8b, and the four most similar chunks (top-k = 4) were inserted into each prompt. Retrieval therefore covers only the procurement-law part of the benchmark: for the 21 items on building regulations, building planning and floor area calculation, the model systematically receives misleading context under open-book, in part at very low cosine similarities (down to 0.32), making the condition also an informal test of robustness against unsuitable context.

Because language models are stochastic, each question was posed in its own isolated chat and repeated over 20 independent runs: 50 × 20 = 1,000 requests per condition, 4,000 in total. This permits reporting not only the mean but also the dispersion across runs, a robustness property that single-shot demonstrations cannot in principle provide.

2.4.3 Answer Extraction and Scoring

Model answers were extracted from the final channel (<answer>...</answer>). The original protocol accepted only a fixed format (ANSWER: <letter(s)>); answers placing the chosen options elsewhere, for instance before the justification, could not be extracted and were counted as wrong. Since the models adhere to that form with differing strictness, the rule produced a systematically unequal penalty.

All 4,000 logged answers were therefore re-scored post hoc with a model-agnostic second parser. It leaves the original extraction untouched wherever it returned a result and reconstructs only those answers from which the original rule could extract no option set; its rules apply identically to both models, and no inference was repeated. Both scorings are reported in parallel:

S1 (protocol-strict): the original rule; only formally marked answers count.

S2 (answer-reconstructed): identical to S1 on all successfully parsed answers, supplemented by post-hoc reconstruction from free text.

For multiple-answer items, two hit rates are reported: a strict one (only the exactly correct set scores) and one with partial credit (share of correctly named options minus share of wrongly named ones). Format adherence is reported as a metric in its own right, distinguishing answers that violate the form but contain a reconstructable option from hard failures, from which no answer set can be obtained under any parser. The distinction matters practically: in an automated chain an unparsable answer is a failed request, whereas a correctly answered but badly formatted one is merely a parser problem.

3. Results

The evaluation comprises two interlocking levels: the quantitative closed-form benchmark with 4,000 requests across four conditions (Section 3.1), the central evidential basis of this paper, and three qualitative case demonstrations of how the two-channel architecture works and where its limits lie (Section 3.2). A third level, blind assessment of reasoning quality by an expert panel, lies outside the scope set out in Section 2.1 and is the priority next research step (Section 5.3).

3.1 Quantitative Benchmark

3.1.1 Overall Accuracy Across the Four Conditions

Table 2 summarises the results. All values are means over 20 independent runs per condition; S1 denotes the protocol-strict scoring, S2 the answer-reconstructed second scoring described in Section 2.4.3.

Table 2: Benchmark accuracy (50 items × 20 runs = 1,000 requests per condition). Hard failures denote answers from which no answer set could be extracted under any parser.

Model	Condition	Strict S1 (%)	Strict S2 (%)	SD S2 (pp)	Partial credit S2 (%)	Hard failures (n)
Spyra-20B	closed-book	61.1	61.1	3.6	69.0	23
gpt-oss-20B (base)	closed-book	66.2	67.3	3.6	74.4	3
Spyra-20B	open-book (RAG)	70.2	70.4	4.3	77.2	25
gpt-oss-20B (base)	open-book (RAG)	71.5	76.3	3.0	82.7	4

Domain-specific fine-tuning does not improve closed-form accuracy on this benchmark. The unmodified base outperforms Spyra-20B in both context conditions: closed-book, 67.3% against 61.1% (S2; Welch $t = 5.48$, $df = 38$, $p < 0.001$); open-book, 76.3% against 70.4% (S2; $t = 5.05$, $df = 34$, $p < 0.001$). The same ordering holds under partial-credit scoring (74.4% vs. 69.0% and 82.7% vs. 77.2%, each $p < 0.001$).

We report this explicitly as the central quantitative finding and classify it, in line with the pre-specified methodology, as a negative result: specialisation on curated reasoning traces does not

translate into better answer selection on this task. Two factors bear on its interpretation: the benchmark predominantly tests German procurement law, only partly within the focus of the training corpus (Section 2.4.1), and the training objective optimises the structure of deliberation rather than the selection of a given option. Both are developed in Section 4.1.

3.1.2 Effect of the Retrieval Layer

Retrieval helps both models, and to roughly the same degree. Adding the four most similar corpus passages raises the accuracy of the base by 9.0 pp (67.3% → 76.3%; paired $t = 2.22$, $p = 0.031$) and that of Spyra-20B by 9.3 pp (61.1% → 70.4%; paired $t = 2.11$, $p = 0.040$); the difference is practically negligible. On the 29 items the corpus actually covers (VOB/A and VOB/B) the effect is considerably larger. The base rises from 71.6% to 88.8%, Spyra-20B from 63.3% to 77.8%, confirming the intended function of the retrieval layer.

Two observations from earlier analyses no longer hold under S2 and are expressly withdrawn: that the models were statistically indistinguishable under open-book (71.5% vs. 70.2%, $t = 1.07$, $p = 0.29$), and that Spyra-20B benefited more strongly from retrieval. Both were artefacts of the format penalty described in Section 2.4.3, which hit the base disproportionately under open-book (56 of 1,000 answers were formally unparsable, 52 of them reconstructable and 48 correct).

3.1.3 Breakdown by Domain

Table 3 differentiates the open-book results by domain. The breakdown is informative but carries no robust claim, because the item counts outside the VOB block are small.

Table 3: Accuracy by domain, open-book (RAG), scoring S2. Δ denotes the difference between Spyra and the base.

Domain	Items	Covered by corpus	Spyra-20B (%)	gpt-oss-20B (%)	Δ (pp)
VOB / procurement law	29	yes	77.8	88.8	- 11.0
Building planning law	10	no	71.5	76.5	- 5.0
Building regulations law	7	no	59.3	47.1	+ 12.2
Floor area calculation	4	no	33.8	36.2	- 2.4

On the 21 items outside the corpus coverage, de facto a condition with misleading context, the base loses 2.4 pp through the retrieval layer while Spyra-20B gains 2.1 pp. Neither change is significant, and the interaction between model and retrieval condition likewise fails to reach significance (+4.5 pp, $t = 1.14$, $p = 0.27$). The fine-tuned model appears somewhat more resistant to unsuitable context, but the effect cannot be separated from noise here; we report it as a direction, not a result. The domain rows are not standalone findings: the floor area row rests on four items and, under S1, produced an apparent collapse of the base to 17.5%, corrected to 36.2% by the re-scoring.

3.1.4 Format Adherence, Failure Modes and Latency

Beyond accuracy, two differences are robust across all 2,000 requests per model.

The first concerns the manner of failure. The base violates the required answer form more often: 5.6% under retrieval against 3.1% for Spyra-20B, but almost always recoverably: 93% of initially unparsable answers could be reconstructed, and 92% of those were correct. The failures of Spyra-20B are predominantly unrecoverable: refusals, answers in continuous prose without option letters, occasional switching into English (1.6% under retrieval) and, once, an empty output. On hard failures alone, the fine-tuned model fails on 2.3-2.5% of requests and the base on 0.3-0.4%. A sixfold difference, and the practically more relevant quantity in an automated chain.

The second concerns response speed. At comparable answer length, Spyra-20B responds roughly twice as fast (median 3.9 s closed-book and 5.0 s under retrieval, against 9.7 s and 9.3 s for the base), reflecting the schematic two-channel format induced by fine-tuning, which produces shorter, more tightly bound deliberative processes.

In summary, fine-tuning improved the form of the output: structure, brevity, latency, without improving its content, and in places degrading it.

3.2 Qualitative Case Demonstrations

The following cases illustrate how the two-channel architecture operates in open design situations. They are illustration, not evidence; robust statements about reasoning quality would presuppose the panel assessment outlined in Section 5.3. The system prompt throughout was: "*You are a senior architect. Use deep simulations for complex problems.*"

3.2.1 Case 1: Separation Distance Check (Reasoning on a Hallucinated Norm)

Task: a two-storey flat-roofed extension to a single-family house, with a boundary distance of 2.50 m; the user asks for a check under § 6 MBO.

Analysis channel (extract): the model derives the logic of separation distances correctly ("*an obliquely sloping open space, dependent on building height*"), identifies building height as the critical variable and formulates a rule: "*For buildings over 4 m in height: minimum distance 4.00 m.*" This rule does not exist in § 6 (5) MBO; the actual measure is 0.4 H with a minimum of 3 m.
Final channel: the model states that the required minimum separation distance is, as a rule, not met at 2.50 m and advises a redesign.
Observation: the deduction is structurally intact, but the premise is hallucinated; under the false assumption of a 4 m minimum, rejecting the 2.50 m proposal is consistent. The case shows what the two-channel architecture can achieve (a traceable derivation) and what it cannot (securing numerical norm values), and motivates external grounding through retrieval (Section 4.3).

3.2.2 Case 2: Densification with Conflicting Objectives (Run-to-Run Instability)

Task: a client wishes to densify a small inner-city plot to the maximum, while the local development plan requires both the preservation of a valuable tree population and protective zones around an adjacent listed building; the user asks for a conflict analysis.

Final channel: in run 1 the model identifies the conflict correctly (full-basement construction would destroy the root zone of the trees), rejects the full basement and proposes a "back-to-back" strategy shifting density to the conflict-free edges. In run 2, on an identical prompt, it recommends shifting density into the vertical by means of a deep basement plus a converted attic storey while claiming that the tree population is protected. An internally contradictory recommendation reproducing exactly the configuration rejected in the first run.

Observation: the capacity for conflict detection is present but not reliably activated; the constraint logic is not sampling-consistent. This supports multiple sampling and downstream constraint verification at inference time, and contradicts the single-shot demonstrations that overstate the reliability of such systems in the applied AI literature.

3.2.3 Case 3: Reuse of a Commercial Building (Tree of Thought in Action)

Task: a deep simulation on the handling of a vacant 1970s commercial property, with parallel assessment of three paths: (A) demolition and new build, (B) mixed use, (C) purely residential use.

Analysis channel: the model assesses the paths in parallel along constant criteria. Path A: high embodied energy, robust permitting situation, most expensive in material terms. Path B: low material loss, social integration, noise as the critical variable in a commercial/residential mix. Path C: retention of the structure, focus on the thermal envelope, sound insulation via the floor plan. The final channel decides for path B, since material reuse and diversity of use outweigh the complexity of the change of use, with a three-step strategy: clarify the noise profile, define buffer zones, refurbish the structure.

Observation: the case demonstrates the intended operation of the tree-of-thought format in huge mode: consistent separation of variables across three paths, with the trade-off disclosed in the analysis channel. Consistency across repeated runs was not systematically tested.

4. Discussion and Policy-level Implications

4.1 Interpretation of the Core Results

The quantitative headline result is negative: the fine-tuned model does not outperform its base model on the 50-item closed-form benchmark in either context condition, and the sign of the effect is reversed, the base scores roughly 6 percentage points higher in both. We report this without qualification for what it is: an empirical result that does not support the claim that fine-tuning on reasoning traces translates directly into better answer selection.

The interpretation we attach to it is not a rearguard defence but a specification of what was measured at which level. The loss described in Section 2.3.3 is defined exclusively on the final channel, while the training data consist predominantly of free-text architectural recommendations rather than multiple-choice selection patterns. Such a model is specialised in giving answers in the form of architectural recommendations, not in choosing from a closed set of options; and the channel that prepares the choice is precisely not optimised for correctness but structurally anchored. A closed-form benchmark accordingly measures an aspect of performance that the training procedure did not treat as an objective.

The explanation does not excuse the result, but it situates the dissociation between form and content. Fine-tuning changed the output structure measurably: more schematic, shorter, twice as fast at the median, while sextupling the rate of unrecoverable answers. What the approach demonstrably produces is therefore the form of deliberation, not its accuracy on closed questions. Whether that structural gain translates into a quality gain in open design situations cannot be answered here; Case 3 suggests it might, Case 2 shows its probabilistic nature. The demonstration remains reserved for a panel assessment (Section 5.3).

4.2 Relation to Previous Research

Spyra-20B stands in line with domain-specifically fine-tuned models such as BloombergGPT (Wu et al., 2023) and Med-PaLM (Singhal et al., 2023), but differs on one decisive point: whereas those models internalise domain-specific terminology and factual knowledge, the approach here aims at internalising a deliberative structure. The sequence of the argument, not the vocabulary, is the objective, and the two-channel architecture implements it technically.

Relative to prompting-based methods such as chain-of-thought (Wei et al., 2022) and tree-of-thought (Yao et al., 2023), the approach chosen here shifts the deliberative structure from the prompt into the model weights: it makes deliberation depth a trainable, metadata-controlled signal and anchors the separation of analysis and answer at the level of the loss. Whether that offers a substantial advantage over well-executed prompting is an open empirical question this paper does not settle.

Relative to retrieval-augmented generation systems (Lewis et al., 2020), Spyra-20B is not an alternative but a complementary component: Section 3.1.2 shows retrieval to be comparably effective in both models, and neither is immunised by fine-tuning against the weaknesses of retrieval-free inference. Case 1 makes this particularly clear, since the model hallucinates a concrete norm value

even with an intact chain of deduction. The system architecture that follows is not reasoning instead of retrieval, but reasoning and retrieval with a strict separation of roles: the deliberative process structures the answer, the external knowledge base grounds its facts.

Conceptually the approach is close to the MAKER framework (Meyerson et al., 2025), whose idea of graded cognitive effort we adopted in defining the reasoning_effort levels. The difference lies in the target environment: MAKER is a general reasoning framework, Spyra-20B its domain-specific realisation in a model directly usable in the AEC sector.

4.3 Policy Implications for Smart Design Governance

The present work is a technical contribution, but its findings bear directly on the governance of AI systems in the construction sector. Four implications follow from the results.

Reasoning transparency as a precondition for regulated deployment. The AI Act (Regulation (EU) 2024/1689) requires, for high-risk systems, traceability of how the system operates (Article 13) and effective human oversight (Article 14). Both are difficult to satisfy for models whose deliberation remains internal. The two-channel architecture addresses this technically, and Case 1 shows the practical effect: the hallucinated norm value was explicit in the analysis channel, so it could be identified and the recommendation rejected. A model whose deliberation is opaque does not permit such a rejection, because the false premise would not be recognisable as such. For procurement practice it follows that the structural separation of deliberation and answer, or an equivalent transparency property, should be a minimum requirement in public tenders for AI-supported planning and review services.

External factual grounding as a governance requirement, not a technical add-on. The retrieval gain of around 9 percentage points in both models and the hallucination in Case 1 despite intact deduction show that grounding numerical and normative statements in an external, authoritative knowledge base is a structural precondition of legally binding use, not an optional improvement. Systems producing statements with legal or safety-relevant binding effect should demonstrate that all norm values, references to statutory provisions and citations of case law derive from a versioned, authoritative source. Verbatim reference thereby becomes a system property rather than a property of the individual answer.

Data sovereignty and local executability. Operation on conventional workstation hardware (Section 2.2) is what allows the planning-related and personal data of German practices to stay out of third-country cloud infrastructures. A requirement derived from the GDPR that gains weight under KRITIS regulation and the NIS-2 Directive once planning service providers work for critical infrastructure. Support for AI-assisted planning should therefore not be architecture-agnostic: incentives for locally executable, openly auditable architectures contribute directly to regulatory controllability, whereas cloud-bound proprietary systems concentrate the same load precisely where data sovereignty is hardest to maintain.

Multi-sample verification as an audit standard. The run-to-run instability observed in Case 2 is a structural property of stochastic language models and contradicts the single-shot demonstrations customary in applied AI reporting and marketing material. For permitting and review practice it follows that statements on safety-relevant matters should be derived from consensus procedures across several runs, with contradictory cases flagged, rather than from a single inference. The protocol of Section 2.4.2 (20 runs per item) should be understood as an audit standard, not merely a methodological caveat.

4.4 Strengths and Limitations

The strengths lie in reproducibility; in the honesty of the report, since the level-1 finding goes against the intervention and is stated as such rather than explained away; and in technical efficiency, since an MoE base model with QLoRA makes specialisation feasible on local consumer hardware.

The limitations are equally clear. The dataset was curated by the first authors without independent second review, so no inter-rater reliability can be reported (Section 2.3.1), and a test for training data bias lies outside the scope of this study. The benchmark is small at 50 items and outside the training focus; it measures out-of-distribution factual recall, not design reasoning. Without the blind panel assessment, the central claim of improved reasoning is not directly demonstrated but only made plausible. The comparison is restricted to the immediate base model, generalisability beyond the German regulatory environment is unexplored, and numerical statements remain insufficiently reliable without external retrieval grounding, so unverified use of the outputs in legally binding applications is not to be recommended. Case 2 shows, finally, that constraint detection is probabilistic in nature.

5. Conclusion

5.1 Summary of the Central Results

Spyra-20B is a proof of concept that the deliberative structure of architectural reasoning can be anchored at the fine-tuning level of an open-weight language model and made inspectable in model behaviour. The methodological contribution lies in combining a metadata-controlled deliberation depth (reasoning_effort), a strict separation of analysis and answer channels, and an open-weight base model that makes ablation of the fine-tuning effect possible at all.

The empirical results paint a sober picture. On the closed-form benchmark the fine-tuned model does not outperform its base model in any condition; what changed were the output structure and the latency. The case studies show the intended operation of the two-channel architecture (Case 3), but equally that conflict detection is not reliably activated across repeated runs (Case 2) and that numerical norm values can be hallucinated even where the deduction is intact (Case 1).

We therefore make no claim about the reasoning quality of the model, but a claim about the visibility of its reasoning: the analysis channel makes premises, inference steps and rejected alternatives accessible as text for professional review. Whether that deliberation is qualitatively better than a generically prompted base model is not answered here.

5.2 Implications of the Results

The central political consequence, formulated deliberately above the level of individual technical requirements, is that AI systems in the construction sector should be assessed along architectural properties rather than performance scores. A system whose deliberation remains opaque, whose factual basis is not traceable and whose data processing takes place in third-country cloud infrastructures is problematic for legally binding deployment irrespective of its scores on individual test cases. A system that satisfies these properties is worth examining in application even where its multiple-choice accuracy falls short of a generic model.

5.3 Recommendations for Further Research

Four research directions follow, in order of methodological priority. The first and most important concerns reasoning quality in open design situations. Demonstrating a substantial gain over generically prompted base models presupposes a blind assessment of open design tasks by a panel of qualified architects: raters blinded to model identity, several runs per prompt to capture the stochasticity of Case 2, and a pre-registered assessment rubric. Until this has been carried out, the central claim of the approach remains empirically unsubstantiated.

The second concerns the reliability of deliberation under sampling. That identical prompts can lead to opposing recommendations motivates a systematic study of ensemble and consensus methods: how many runs a stable output requires, which aggregation rules prove effective, and where constraint

verification can usefully be inserted at inference time. The third concerns a tighter coupling of model and retrieval. The present results show the effectiveness of a retrofitted RAG layer; whether retrieval modelled within the training process reduces numerical hallucination structurally, rather than compensating for it externally, remains open and would require a substantial extension of the training procedure. The fourth concerns widening the regulatory scope. Transferability studies to other jurisdictions, not primarily through dataset extension, but through the question of which parts of the deliberative structure remain regulatorily invariant, would be a substantial building block for a claim reaching beyond the German context. Scaling to larger base models (gpt-oss-120B, for instance) is technically obvious but methodologically subordinate to these four: a model whose central claim to quality has not been externally validated gains neither robustness nor regulatory viability from a larger parameter count. The order of research follows the order of evidential force.

Acknowledgements
The authors would like to thank Jade University of Applied Sciences, Oldenburg, and the Young Researchers programme.

Funding
This research did not receive any specific funding from public, commercial or charitable funding bodies.

Conflicts of Interest
The authors declare that there are no conflicts of interest.

Data availability statement
The data supporting the findings of this study have not been made publicly available for reasons relating to the protection of intellectual property and copyright considerations; they may be provided upon reasoned request to the corresponding author, N.A.

Institutional Review Board Statement
Not applicable. The study does not involve research on either humans or animals; the evaluation was carried out using the university's own examination format without the participation of examinees.

CRediT author statement:
Conceptualisation: N.A., Y.H., G.G.; Methodology: N.A., Y.H.; Software: N.A.; Data curation: N.A., Y.H.; Formal analysis: N.A.; Investigation: N.A.; Validation: Y.H.; Resources: G.G.; Supervision: G.G.; Visualisation: N.A.; Drafting the original version: N.A.; Review and editing: N.A., Y.H., G.G. All authors have read and approved the final version of the manuscript.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., … Zoph, B. (2023). GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Bashir, A. H., Khalid, M. R., Cvejoski, K., Birr, J., Berghaus, J., Berger, A., Halscheidt, S., Temath, C., Sifa, R., & Berghaus, D. (2026). *Domain-adaptation through synthetic data: Fine-tuning large language models for German law*. arXiv. https://doi.org/10.48550/arXiv.2601.14160

Bommarito, M., II, & Katz, D. M. (2022). *GPT takes the bar exam*. arXiv. https://doi.org/10.48550/arXiv.2212.14402

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2023). PaLM: Scaling

language modeling with pathways. *Journal of Machine Learning Research, 24*(240), 1–113. https://doi.org/10.5555/3648699.3648939

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems, 36*, 10088–10115. https://doi.org/10.52202/075280-0441

Fernandes, R., Biedenkapp, A., Hutter, F., & Awad, N. (2025). *A Llama walks into the 'Bar': Efficient supervised fine-tuning for legal reasoning in the multi-state bar exam*. arXiv. https://doi.org/10.48550/arXiv.2504.04945

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey*. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Hirsekorn, Y., Ansre, N., & Grunwald, G. (2025). AI-driven knowledge transfer in architectural education. *Smart Design Policies, 2*(1), 122–139. https://doi.org/10.38027/smart.v2n1-8

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2106.09685

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), Article 42. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213. https://doi.org/10.52202/068431-1613

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://doi.org/10.5555/3495724.3496517

Li, P., Li, B., & Li, Z. (2023). Sketch-to-architecture: Generative AI-aided architectural design. In *Pacific Graphics short papers and posters* (pp. 99–102). Eurographics Association. https://doi.org/10.2312/PG.20231276

Li, Z., Xia, L., Tang, J., Xu, Y., Shi, L., Xia, L., Yin, D., & Huang, C. (2024). UrbanGPT: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5351–5362). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671578

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., … Zhao, L. (2025). Domain specialization as the key to make large language models disruptive: A comprehensive survey. *ACM Computing Surveys, 58*(3), Article 79, 1–39. https://doi.org/10.1145/3764579

Madireddy, S., Gao, L., Din, Z., Kim, K., Senouci, A., Han, Z., & Zhang, Y. (2025). Large language model-driven code compliance checking in building information modeling. *Electronics, 14*(11), 2146. https://doi.org/10.3390/electronics14112146

Meyerson, E., Paolo, G., Dailey, R., Shahrzad, H., Francon, O., Hayes, C. F., Qiu, X., Hodjat, B., & Miikkulainen, R. (2025). *Solving a million-step LLM task with zero errors*. arXiv. https://doi.org/10.48550/arXiv.2511.09030

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Proceedings of*

the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1198–1212). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594074

OpenAI. (2023). *GPT-4 technical report*. arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2025). *gpt-oss-120b & gpt-oss-20b model card*. arXiv. https://doi.org/10.48550/arXiv.2508.10925

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*, 27730–27744. https://doi.org/10.52202/068431-2011

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Agüera y Arcas, B., … Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature, 620*(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv. https://doi.org/10.48550/arXiv.2307.09288

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain-of-thought reasoning in language models. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://doi.org/10.52202/068431-1800

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from language models*. arXiv. https://doi.org/10.48550/arXiv.2112.04359

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). *BloombergGPT: A large language model for finance*. arXiv. https://doi.org/10.48550/arXiv.2303.17564

Yang, F., & Zhang, J. (2024). Prompt-based automation of building code information transformation for compliance checking. *Automation in Construction, 168*, Article 105817. https://doi.org/10.1016/j.autcon.2024.105817

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems, 36*, 11809–11822. https://doi.org/10.52202/075280-0517

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics, 51*(4), 1373–1418. https://doi.org/10.1162/coli.a.16