TEACHING LLMS NEW KNOWLEDGE WITHOUT HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models increasingly operate in settings where they encounter information that was unavailable during pretraining. While such knowledge can be stored externally, continually internalizing it into model weights could improve efficiency and enable reasoning across information acquired at different times. We study knowledge acquisition from natural-language documents and identify a central failure mode: fine-tuning on new knowledge can rapidly increase hallucinations. Models become more willing to answer questions about the newly learned domain even when the required information was never provided, and newly acquired facts can also spill over into answers to unrelated questions. Such hallucinations are undesirable in themselves and are especially concerning in continual-learning systems, where they may accumulate over time. We introduce hallucination-aware training, which combines context distillation with hallucination-probing questions that teach the boundary of the acquired knowledge and a regularization objective that preserves the model's behavior on unrelated questions. Across Qwen3-32B, GPT-OSS-20B, Gemma4-31B, and Qwen3.8-27B, our approach substantially reduces hallucinations and knowledge spillover while preserving strong acquisition of new knowledge. We further find that models can compose facts learned independently from different documents, even though training processes only one document at a time and provides no explicit supervision for cross-document composition. Our results suggest that LLMs can internalize new knowledge while learning both how to use that knowledge and where its boundaries lie.

1 INTRODUCTION

Large language models (LLMs) acquire most of their knowledge during pretraining (Petroni et al., 2019; Kandpal et al., 2023). However, AI agents deployed in real-world environments continuously encounter new information that was unavailable during pretraining and must be able to incorporate that information into their subsequent reasoning and behavior. Learning should therefore also happen "on the job": when an agent encounters new information during deployment, it should be able to internalize this knowledge and reliably use it when solving future tasks (Silver & Sutton, 2025). State-of-the-art AI agents commonly rely on external documents or knowledge bases to store newly acquired information. Such external memory is useful and often irreplaceable, but we argue that it should be complemented by internalizing knowledge into model weights (Snell et al., 2022; Padmanabhan et al., 2023; Kujanpää et al., 2024). Internalization can improve efficiency by reducing the need to query external knowledge for every task, help avoid common and repeated mistakes, and enable models to reason about conflicting information and the evolution of knowledge over time.

Much prior work on updating model knowledge has focused on knowledge editing, where new facts are typically represented as isolated knowledge-graph-style triples (see, e.g., De Cao et al., 2021; Mitchell et al., 2022; Meng et al., 2022; 2023). Although this setting provides a useful framework for controlled evaluation, it does not fully capture the richer and more contextual process of learning new knowledge during deployment (Wu et al., 2024). Motivated by realistic deployment scenarios, we instead consider the problem of continually learning new knowledge from collections of documents. Documents provide coherent contexts in which facts are embedded in narratives and related to one another, rather than presented as isolated statements. Such coherent contexts can facilitate knowledge internalization (Gao et al., 2026). We therefore study how LLMs can internalize knowledge from documents while retaining the ability to use it reliably in subsequent interactions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Successful continual learning, however, requires more than memorizing individual facts (Zhong et al., 2023). New information often needs to be interpreted in relation to knowledge acquired previously, potentially from different sources and at different points in time. A model must therefore be able to recall relevant existing knowledge and combine it with newly observed information, a capability we refer to as *knowledge aggregation*. At the same time, the model must avoid hallucinating information that is not supported by either the newly learned knowledge or its existing knowledge (Zucchet et al., 2025). Thus, continual learning requires not only retaining and combining newly acquired information, but also reliably distinguishing what follows from that information from what does not.

Consistent with prior findings (Gekhman et al., 2024; Zucchet et al., 2025; Sun et al., 2025), we find that such hallucinations emerge rapidly when models are fine-tuned on new knowledge. In particular, after learning new information, models become prone to *unsupported recall*: they confidently generate answers to questions for which the available knowledge provides insufficient evidence and for which they should instead abstain. This behavior occurs for questions involving newly learned entities or themes whose answers cannot be inferred from the learned information, and even for questions involving fictitious entities. We observe these behaviors across four LLMs spanning multiple model families and architectural designs, including Qwen3 (Yang et al., 2025), GPT-OSS (Agarwal et al., 2025), and Gemma (Team et al., 2024).

We additionally observe *knowledge spillover*, whereby newly learned facts, entities, or themes appear in answers to questions about unrelated topics (Kazemi et al., 2023; Sun et al., 2025). Such spillover is undesirable in itself, as it can cause an agent to base its reasoning or actions on information irrelevant to the current task. It may also affect agents that rely on external memory: even when knowledge itself is stored externally, the model must still determine what to retrieve, for example which entities, concepts, or keywords to search for, or where relevant information is likely to be stored. If newly learned knowledge inappropriately influences this recall process, it may distort retrieval queries, leading to less relevant retrievals or causing relevant information to be missed.

To address these failure modes, we propose a *hallucination-aware training strategy* for learning new knowledge. Rather than training the model to immediately abstain when faced with an unanswerable question (Zhang et al., 2024; Gekhman et al., 2024), our approach teaches it to first relate the question to the knowledge it has acquired and determine what information is actually supported. This distinction is important because recognizing that a question is unanswerable can itself require substantial reasoning: in many cases, it is difficult to determine a priori whether the available knowledge is sufficient to answer a question. After identifying the relevant knowledge, the model is trained either to answer a related question that *is* supported by this knowledge or to abstain when no appropriate answer can be given. In this way, the model learns to use the relevant knowledge it has without hallucinating information needed to answer the original question. We additionally introduce a regularization strategy that preserves the model's behavior on unrelated questions, preventing newly learned information from spilling over into unrelated knowledge. Experiments show that these interventions substantially reduce hallucination rates and improve the reliability of knowledge internalization while preserving the model's ability to acquire and use new knowledge.

Finally, we provide experimental evidence that models can aggregate knowledge acquired from separate sources. Our training procedure processes one document at a time and does not explicitly encourage the model to compose knowledge across documents. Nevertheless, after internalization, models are able to combine information learned from different documents when answering questions that require such composition. This result suggests that knowledge acquired during separate learning experiences can become integrated through the model's internal representations and subsequently composed during reasoning (Kujanpää et al., 2024; Treutlein et al., 2024).

2 Related Work

Knowledge acquisition. Learning on the job requires models to acquire information encountered during deployment and use it in subsequent tasks. This includes newly available information and knowledge that pretraining did not capture reliably, for example because it was rarely encountered (Kandpal et al., 2023). Prior work has studied how to extend or revise the knowledge stored in model weights through continued pretraining on new corpora (Jang et al., 2021) and targeted editing of factual associations (De Cao et al., 2021; Mitchell et al., 2022; Meng et al., 2022; 2023). We study

acquisition from natural-language documents after post-training, when the model has already learned to follow instructions and answer questions.

Hallucinations induced by knowledge acquisition. Knowledge acquisition can nevertheless be accompanied by unsupported predictions. For example, during pretraining on synthetic biographies, models learn facts about people in the training data while becoming more confident in incorrect predictions about people absent from it (Zucchet et al., 2025). This loss of reliability can extend to previously learned knowledge: fine-tuning on question–answer pairs about facts the model does not know well can increase errors on other facts it previously knew (Gekhman et al., 2024; Ghosal et al., 2024; Dang et al., 2025). Fine-tuning can further induce knowledge spillover: models may overpredict answers encountered during training, including on questions where those answers are incorrect (Kazemi et al., 2023). Newly learned information can even appear in responses to unrelated questions (Sun et al., 2025). In continual learning, unsupported claims may become internalized and recalled as facts if model-generated responses are reused in later training.

Reliable acquisition of new document knowledge. One approach to limiting unsupported recall is to teach models the limits of their existing knowledge. Kadavath et al. (2022) train models to predict whether they know an answer, while R-Tuning uses the model's initial answer correctness to assign certainty or uncertainty labels during fine-tuning (Zhang et al., 2024). In factual QA adaptation, Gekhman et al. (2024), Dang et al. (2025) and Kaplan et al. (2026) study how unfamiliar question-answer pairs introduced during finetuning increase errors on previously known facts. To reduce these errors, Dang et al. (2025) revisit known examples, while Kaplan et al. (2026) regularize the model's output distribution toward that of an earlier checkpoint. Sun et al. (2025) mitigate knowledge spillover through text augmentation and pruning the largest parameter updates. These studies focus on answering reliably from existing knowledge. None of them demonstrates that unfamiliar facts introduced during fine-tuning are acquired as usable knowledge, that is, that models can apply those facts to questions not seen during training.

We instead study how to acquire new knowledge from documents while limiting hallucinations about the newly learned domain and the spillover of learned information into unrelated responses. For knowledge acquisition, we build on context distillation, which trains a student to match a teacher's predictions while only the teacher receives the source context (Snell et al., 2022; Padmanabhan et al., 2023). Kujanpää et al. (2024) apply this approach to questions generated from documents, while DiSC distills directly from document segments (Padmanabhan et al., 2026). Both evaluate acquisition and capability preservation, without dedicated tests of unsupported recall about the newly learned domain. Our work combines context distillation with explicit supervision of knowledge boundaries: alongside questions about the supplied facts, hallucination-probing questions ask about information absent from the documents. This way, models learn to recall relevant information and assess what it supports before answering or abstaining, while regularization limits changes to unrelated responses.

3 Benchmark for Knowledge Acquisition

Evaluating knowledge acquisition requires distinguishing newly learned information from facts already available through pretraining, which is difficult with existing corpora (Gekhman et al., 2024). Following controlled synthetic studies of knowledge acquisition (Allen-Zhu & Li, 2023; Zucchet et al., 2025), we construct a synthetic benchmark in which we control the underlying facts and can reasonably assume they are new to the model. To model learning from information encountered during deployment, we present these facts in natural-language documents rather than as preprocessed atomic facts. The benchmark tests whether models can extract and internalize this knowledge, combine facts across documents, and recognize when the available information is insufficient or conflicts with a question. We also test whether acquiring this *new knowledge* disrupts the model's pre-existing, unrelated *old knowledge*.

We construct four fictional companies, first defining canonical facts about their employees, operations, and histories. These facts serve as the single source of truth for generating consistent documents and determining which questions are supported, unanswerable, or based on false premises (example facts are shown in Table 4). An LLM renders these facts into 10 natural-language documents per company, each containing only a subset of the facts. Learning the full company knowledge therefore requires accumulating information across documents, allowing us to test subsequent composition of

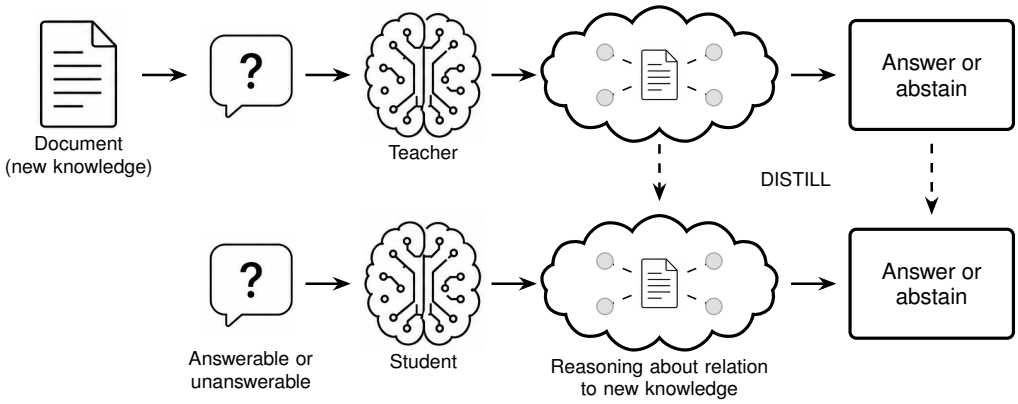

Figure 1: Hallucination-aware context distillation. For each question, whether answerable or unanswerable, the teacher reasons about how the question relates to the newly acquired knowledge with the source document in its context, and then either answers or abstains. The student is trained to reproduce this reasoning and response without access to the document.

independently learned facts. Figure 3 provides an example document. Our main experiments use two companies; scaling experiments use all four.

After training, we test new knowledge using a held-out test set of three question types: 10 direct, 10 compositional, and 10 hallucination-probing questions per company. *Direct questions* test individual fact acquisition. Each question targets a fact from a different document. Table 5 lists the questions and ground-truth answers for one company. *Compositional questions* require combining facts from at least two documents, for example to list all employees or match responsibilities to people (see Table 6 for examples). *Hallucination-probing (unanswerable or false-premise) questions* concern the fictional companies but cannot be answered as stated from the acquired knowledge (see examples in Table 7). Unanswerable questions ask for information neither stated nor inferable, such as annual revenue, largest clients, or an employee's education; the model should recognize its insufficient knowledge. False-premise questions conflict with canonical facts by introducing fictitious employees, offices, or organizational units, or perturbing known facts; the model should reject the premise. These new-knowledge questions are constructed semi-automatically by iteratively correcting and refining LLM-generated candidates, with compositional questions requiring substantially more guidance.

To evaluate effects of training on pre-existing knowledge, *old-knowledge questions* probe facts unrelated to the synthetic companies. Following PreciseWikiQA (Bang et al., 2025), we generate 100 test questions, each a short factual question answerable from a single Wikipedia passage (see Appendix A for construction details).

4 Hallucination-Aware Training

Our goal is to internalize the knowledge contained in each document while preventing the hallucinations that can arise after the model acquires it. To this end, we use a fully automated procedure that generates training questions of three types: knowledge-acquisition questions that teach the facts in each document, hallucination-probing questions that teach the model how to behave when a question is related to the new knowledge but cannot be answered from it, and old-knowledge questions that regularize the model's behavior on topics unrelated to the newly acquired knowledge.

We use *context distillation* to internalize the new knowledge using these training questions (Figure 1). The teacher and the student are the same model. For each training question, the teacher receives both the source document and the question and generates an answer. Its output distributions over the answer tokens then serve as the targets for the student, which is given only the question, without the document. We train the student by minimizing the KL divergence between the student's and the teacher's output distributions on these answer tokens. Because the teacher conditions on the document while the student does not, this transfers the information contained in the document into the student's weights, so that the student can later answer the question without access to the document.

We train the model to answer both with and without reasoning. To this end, we generate the teacher's answer in one of two answering modes, reasoning or direct, assigning each training question to a mode at random so that half the answers are generated in each. In the *reasoning* mode, the teacher reasons before producing its final answer, and we distill both the reasoning trace and the final answer; in the *direct* mode, reasoning is disabled and the teacher produces the final answer without an explicit reasoning trace, which we distill on its own. GPT-OSS-20B is handled slightly differently because we find that disabling reasoning substantially degrades its answer quality: in particular, the model often fails to follow instructions and to recognize false premises or conflicting facts. We therefore always allow the GPT-OSS-20B teacher to reason. In the direct mode, the resulting reasoning trace is hidden from the student, and only the final answer is distilled.

Both generation steps are fully automated and require no human annotation: the model being trained generates the training questions (described below) and, as the teacher, produces the answers that serve as the distillation targets. All training data is produced in advance and held fixed throughout training.

4.1 Questions for Knowledge Acquisition

For each document, we generate a diverse set of questions designed to expose its knowledge from multiple directions, probing its facts, entities, and relations. We first prompt the model to extract a set of facts from the document. We then rewrite each fact to make it self-contained, such that all named entities required to understand the fact are explicitly identified and the fact can be understood without access to the original document. We construct questions that provide multiple ways of accessing the same underlying fact, because learning a relation in one direction does not necessarily make it accessible in the reverse direction (Berglund et al., 2023).

Starting from these self-contained facts, we generate several types of questions. First, for each fact, we generate a question for which the fact itself provides the answer. We then extract all named entities mentioned in the fact and generate additional questions centered on these entities. For every named entity, we generate questions in which the entity itself is the answer, as well as questions in which the entity is mentioned in the question and another aspect of the fact must be recovered. Finally, given the named entities appearing in a fact, we generate questions about the relations among them.

4.2 Hallucination-Probing Questions

Training only on answerable questions does not teach the model how to behave when its acquired knowledge is insufficient. As our experiments show, after learning new facts models often hallucinate answers to questions about the same entities or topics that cannot be answered from the acquired knowledge. We therefore generate *hallucination-probing questions* for every document and include them in context distillation, exposing the model to the boundary between what can and cannot be inferred from its new knowledge.

We generate two types of questions. For *unanswerable questions*, we prompt the model to ask about information neither present in nor inferable from the document. For *false-premise questions*, we take the direct knowledge-acquisition questions and replace one named entity in each with a fictitious one. Together, these teach the model to separate what its new knowledge supports from what it does not, so it abstains rather than hallucinates. Like the rest of our pipeline, this procedure is fully automated and uses the model being trained to generate the questions.

Importantly, we do not train the model to immediately abstain on these questions. Deciding whether a question is answerable can itself require recalling and reasoning over the acquired knowledge, and training the teacher to refuse outright makes the model overly cautious: as we show in Section 5.5, immediate abstention barely changes the hallucination-free rate but substantially weakens the model's ability to combine facts across documents. Instead, the teacher generates its response with the source document in context, grounding its answer in the new knowledge and connecting the question to the facts actually supported (Figure 1). Rather than an immediate refusal, it answers the original question when supported, answers a related supported question, or abstains when no relevant answer exists; we also prompt it to recall relevant facts first. Through context distillation, the student reproduces this behavior without the document.

4.3 Old-Knowledge Regularization

Hallucination-aware training targets hallucinations on questions related to the new knowledge, but fine-tuning can also change the model's behavior on unrelated topics, causing newly learned entities, facts, or themes to spill over into answers to old-knowledge questions. To prevent this, we add a regularization objective that preserves behavior outside the domain of the new documents, using training questions probing old knowledge. These questions are generated in the same way as the PreciseWikiQA test questions, described in Section 3.

Before training, we query the model on each old-knowledge question in a closed-book setting and record its answer without filtering for correctness. During training, the question and this original answer form a KL-distillation target that encourages the updated model to reproduce its original behavior. The aim is thus not to improve factual accuracy on old-knowledge questions, but to prevent knowledge acquisition from unnecessarily changing existing behavior.

The two components play complementary roles. Hallucination-probing questions constrain behavior close to the new knowledge, teaching the model not to hallucinate when a related question cannot be answered from what it has learned. Old-knowledge regularization instead constrains behavior on unrelated topics, reducing the spillover of newly acquired facts and entities into unrelated responses.

5 Experiments

5.1 Experimental Setup

We experiment with training four models, Qwen3-32B, GPT-OSS-20B, Gemma4-31B, and Qwen3.8-27B. For each model, we generate the training data using the automated pipeline described in Section 4. We fine-tune all models using LoRA with rank 64; further training details are given in Appendix B.

Training examples are sampled from four categories: *new-knowledge*, *unanswerable*, *false-premise*, and *old-knowledge* questions. In the standard training setup, we train exclusively on new-knowledge questions, corresponding to mixture proportions $1/0/0/0$. In the hallucination-aware setup, Qwen3-32B, Gemma4-31B, and Qwen3.8-27B use proportions $0.69/0.05/0.21/0.05$, while GPT-OSS-20B uses $0.80/0.05/0.10/0.05$. While the other question categories are used in both the direct and reasoning modes, the old-knowledge regularization examples are generated only in the direct (non-thinking) mode. Table 3 in Appendix B reports the number of generated training questions of each category, per model.

We evaluate models on test questions concurrently with training using a separate vLLM server. The evaluation server is updated with the current model weights at regular intervals during training, allowing us to track both knowledge acquisition and hallucination behavior throughout training. At each evaluation point, we evaluate the model in both the direct and reasoning modes.

5.2 Evaluation Metrics

We use these metrics to evaluate knowledge acquisition and hallucination behavior:

- *New-knowledge accuracy.* Correctness of answers to the direct questions (Section 3), judged with the source document available.
- *Compositional accuracy.* Correctness of answers to the compositional questions (Section 3), which require combining facts spread across several documents, judged with the relevant source documents available.
- *Hallucination-free rate.* Whether answers to the hallucination-probing questions avoid hallucinated content, judged with all documents containing the newly acquired knowledge available. An answer counts as hallucination-free whether it abstains or recalls supported facts, so immediate abstention is not required.
- *Old-knowledge accuracy.* Correctness of answers to the held-out old-knowledge questions (Section 3), judged against the ground-truth answer.

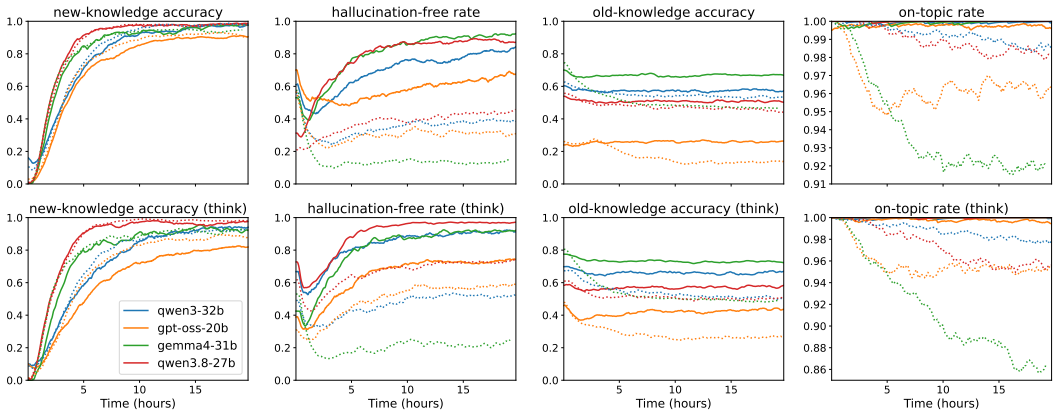

Figure 2: Scores during training, comparing standard knowledge acquisition (dotted lines) with hallucination-aware training (solid lines). Higher is better for all metrics. Results are shown for direct (top) and reasoning (bottom) modes.

- *On-topic rate.* Whether answers to old-knowledge questions stay unrelated to the synthetic companies; answers that unnecessarily mention their facts, entities, or themes are penalized, so a high rate indicates little knowledge spillover.

All metrics range from 0 to 1, with higher values indicating better performance. Every metric is computed by an LLM judge (see the prompts in Appendix E). During training the judge is the model being trained itself. To check that our findings are not an artifact of this self-grading, we additionally regrade the answers from the final evaluation round with an independent judge (GPT-5-mini) using the same prompts. We find that the two judges largely agree (Appendix F).

5.3 Hallucinations Emerge During Standard Knowledge Acquisition

We first study knowledge acquisition without hallucination-aware training, training models exclusively on new-knowledge questions (no unanswerable, false-premise, or old-knowledge regularization). Figure 2 shows all four metrics throughout training (dotted lines), for direct and reasoning modes. New-knowledge accuracy rises rapidly as models internalize the company information, but they also become substantially more prone to hallucination. Before fine-tuning, the base models can already avoid answering questions about the unknown companies, giving a moderate initial hallucination-free rate. This ability deteriorates sharply within the first few optimization steps: the hallucination-free rate drops as new-knowledge accuracy begins to rise, indicating that an early effect of learning the domain is that the model becomes more willing to answer related questions, including those it lacks the knowledge to answer. The effect is especially pronounced in the direct mode.

With further training, the hallucination-free rate partially recovers as models acquire more knowledge. Reasoning substantially aids this recovery: for Qwen3-32B, Qwen3.8-27B and GPT-OSS-20B, the reasoning-mode rate can approach its initial level, whereas the direct-mode rate stays considerably lower; for Gemma4-31B it does not fully recover even with reasoning. This suggests that reasoning over acquired knowledge helps the model judge whether the information needed to answer is actually available. Appendices I.1 and J.1 show example Gemma4-31B answers after standard training, illustrating how it confidently fabricates unsupported details. Standard training also induces the second behavior, knowledge spillover: the on-topic rate decreases during training as facts, entities, or themes from the new companies appear in responses to old-knowledge questions. Trajectories differ across models, but the effect emerges quickly and persists. GPT-OSS-20B differs in reasoning mode, where the on-topic rate partially recovers, though substantial spillover remains.

5.4 Hallucination-Aware Training

We next repeat knowledge acquisition using our hallucination-aware training procedure. In addition to new-knowledge questions, the training mixture now includes unanswerable and false-premise

questions, which teach the model how to behave at the boundary of its newly acquired knowledge, and old-knowledge questions, which regularize its behavior on unrelated topics. Figure 2 reports the same four metrics (solid lines), allowing the training dynamics to be compared directly with standard knowledge acquisition (dotted lines).

Hallucination-aware training preserves the model's ability to acquire new knowledge. New-knowledge accuracy increases rapidly for all three models, demonstrating that introducing hallucination-probing and old-knowledge questions does not prevent the synthetic company information from being internalized. At the same time, the hallucination-free rate improves substantially compared with standard training. Although it can decrease at the beginning of training, it subsequently rises as the models acquire the new knowledge and learn how to handle questions that cannot be answered from it. The improvement is particularly strong in reasoning mode. Appendices I.2 and J.2 show example answers of Gemma4-31B after hallucination-aware training, in the direct and reasoning modes respectively, illustrating how the model recalls relevant facts and abstains when the answer is not supported by the acquired knowledge.

Old-knowledge regularization also strongly reduces knowledge spillover. In contrast to standard training, the on-topic rate remains close to 1 throughout hallucination-aware training for all three models. Thus, including only 5% old-knowledge questions in the training mixture is sufficient in our experiments to largely prevent newly acquired company knowledge from appearing in unrelated responses. This result suggests that preserving the model's original behavior on a relatively small set of unrelated questions can effectively localize the effects of knowledge acquisition. Notably, although the old-knowledge regularization examples were generated only in the direct (non-thinking) mode, both the old-knowledge accuracy and the on-topic rate also improve in the reasoning mode, indicating that the effect of the regularization transfers to the answering mode in which it was not applied during training.

5.5 Grounding the Teacher Improves Composition

Our training never explicitly teaches cross-document composition: during context distillation the teacher sees a single document at a time and never has a company's full knowledge at once. Nevertheless, all models learn to answer compositional questions that require combining facts from several documents (Table 1). Even on exhaustive questions that no model fully solves, such as naming every employee of a company, the answers routinely combine facts from different documents rather than a single one. Cross-document composition thus emerges without explicit supervision.

How well this ability is preserved, however, depends on how the student is trained to handle questions at the boundary of its knowledge. A central design choice in hallucination-aware training is how the teacher constructs its distillation targets for the unanswerable and false-premise questions. To probe its effect on composition, we compare three variants, keeping everything else fixed (including the $0.69/0.05/0.21/0.05$ mixture) and running each for Qwen3-32B, Gemma4-31B, and Qwen3.8-27B:

- *Grounded*$^{+}$ (our default): the teacher sees the source document and is prompted to recall relevant facts before responding.
- *Grounded*$^{-}$: the teacher sees the document but is not prompted to recall facts first.
- *Not grounded*: the teacher has no access to the document and abstains immediately, similar to prior immediate abstention supervision (Gekhman et al., 2024; Zhang et al., 2024).

All three variants keep the hallucination-free rate high (Table 2); grounding has only a small effect on this metric. However, the effect of grounding on composition is more pronounced (Table 1): withholding the document from the teacher (*Not grounded*) markedly reduces compositional accuracy, whereas dropping only the recall-first prompt has a small effect. Intuitively, direct abstention supervision (Gekhman et al., 2024; Zhang et al., 2024) may make the model overly cautious, causing it to abstain too readily and discouraging it from going beyond the individual facts seen in single documents. Grounding instead trains the model to search its acquired knowledge and connect the relevant facts before answering, which is exactly what compositional questions demand; this habit of reasoning over acquired knowledge therefore aligns with compositional skill. Even so, grounded hallucination-aware training can still slightly reduce composition relative to standard training, because the boundary questions encourage some caution.

		Hallucination-aware training		
Model	Standard	Grounded^{+}	Grounded^{-}	Not grounded
Qwen3-32B	0.63 / 0.68	0.65 / 0.65	0.61 / 0.58	0.52 / 0.53
Gemma4-31B	0.67 / 0.77	0.67 / 0.76	0.63 / 0.81	0.51 / 0.65
Qwen3.8-27B	0.80 / 0.72	0.56 / 0.69	0.73 / 0.71	0.43 / 0.48

Table 1: Compositional accuracy (obtained with an independent judge GPT-5-mini), for standard training and three hallucination-aware variants. Each cell reports the direct / reasoning mode score.

		Hallucination-aware training		
Model	Standard	Grounded^{+}	Grounded^{-}	Not grounded
Qwen3-32B	0.16 / 0.38	0.79 / 0.82	0.77 / 0.84	0.85 / 0.80
Gemma4-31B	0.22 / 0.24	0.93 / 0.95	0.94 / 0.90	0.78 / 0.85
Qwen3.8-27B	0.49 / 0.69	0.91 / 0.96	0.99 / 0.98	0.91 / 0.89

Table 2: Hallucination-free rate (obtained with an independent judge GPT-5-mini), for standard training and three hallucination-aware variants. Each cell reports the direct / reasoning mode score.

5.6 Ablations

We complement the main results with two further analyses, reported in full in the appendices. In Appendix G, we use the cross-entropy loss computed on the teacher-generated tokens instead of our default KL objective. We observe the same phenomenon: standard training still induces substantial hallucinations, while hallucination-aware training effectively mitigates them. However, using the KL loss achieves noticeably higher new-knowledge and, especially, compositional accuracy. This suggests that the "dark knowledge" (Hinton et al., 2015) in the teacher's output distribution helps the model better internalize new facts and their relationships. In Appendix H, we train Qwen3-32B on four companies, roughly doubling the amount of new knowledge. We observe very similar training dynamics to the two-company setting. In particular, with hallucination-aware training the model acquires the new knowledge at nearly the same rate while maintaining strong hallucination-free and on-topic behavior.

6 Conclusion

We studied continual knowledge acquisition where new information arrives as natural-language documents that must be internalized into model weights. Standard training rapidly induces hallucinations: models become more likely to hallucinate on questions related to the new knowledge, which can also spill over into unrelated answers. Our hallucination-aware training addresses both failure modes by combining context distillation with hallucination-probing questions that teach the boundaries of the new knowledge and old-knowledge regularization that preserves behavior elsewhere.

Models learn remarkably reliable behavior at these boundaries, even in the direct mode without an explicit reasoning trace. For the strongest models, this behavior becomes nearly perfect once reasoning is enabled. Our ablations show that the key ingredient is grounding the teacher's targets in the source document rather than teaching immediate abstention. This grounding both reduces hallucination and preserves the model's ability to combine independently acquired facts. Indeed, although the teacher never sees more than one document at a time, all models compose facts across documents, and behavior stays similar when we double the amount of new knowledge.

Internalizing knowledge into weights should nonetheless complement, not replace, external stores, since even a boundary-aware model can still misremember. A promising direction is to learn how to retrieve knowledge while learning the knowledge itself, so that a model can fall back on the original source when its internal knowledge is uncertain. More broadly, our results suggest that continual knowledge acquisition need not lead to uncontrolled hallucination, and that integrating these capabilities with persistent external memory is a central challenge for future agents.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, Vienna, Austria, 2025. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "a is b" fail to learn "b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Euirim Choi. GoodWiki dataset. `https://www.github.com/euirim/goodwiki`, September 2023.

Renfei Dang, Peng Hu, Zhejian Lai, Changjiang Gao, Min Zhang, and Shujian Huang. Understanding new-knowledge-induced factual hallucinations in LLMs: Analysis and interpretation. *arXiv preprint arXiv:2511.02626*, 2025.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6491–6506, 2021.

Ya Gao, Kalle Kujanpää, Pekka Marttinen, Harri Valpola, and Alexander Ilin. Edit knowledge, not just facts via multi-step reasoning over background stories. *arXiv preprint arXiv:2602.02028*, 2026.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, nov 2024. Association for Computational Linguistics.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. *arXiv preprint arXiv:2406.14785*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL `https://arxiv.org/abs/1503.02531`.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15696–15707. PMLR, 2023.

Guy Kaplan, Zorik Gekhman, Zhen Zhu, Lotem Rozner, Yuval Reif, Swabha Swayamdipta, Derek Hoiem, and Roy Schwartz. Why fine-tuning encourages hallucinations and how to fix it. *arXiv preprint arXiv:2604.15574*, 2026.

Mehran Kazemi, Sid Mittal, and Deepak Ramachandran. Understanding finetuning for factual knowledge extraction from language models. *arXiv preprint arXiv:2301.11293*, 2023.

Kalle Kujanpää, Pekka Marttinen, Harri Valpola, and Alexander Ilin. Efficient knowledge injection in LLMs via self-distillation. *arXiv preprint arXiv:2412.14964*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *arXiv preprint arXiv:2202.05262*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/pdf?id=0DcZxeWfOPt.

Shankar Padmanabhan, Yasumasa Onoe, Michael J.Q. Zhang, Greg Durrett, and Eunsol Choi. Propagating knowledge updates to LMs through distillation. *arXiv preprint arXiv:2306.09306*, 2023.

Shankar Padmanabhan, Mustafa Omer Gul, and Tanya Goyal. Updating parametric knowledge with context distillation retains post-training capabilities. *arXiv preprint arXiv:2602.16093*, 2026.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

David Silver and Richard S Sutton. Welcome to the era of experience. *Google AI*, 1(11):6, 2025.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Chen Sun, Renat Aksitov, Andrey Zhmoginov, Nolan Andrew Miller, Max Vladymyrov, Ulrich Rueckert, Been Kim, and Mark Sandler. How new data permeates LLM knowledge and how to dilute it. *arXiv preprint arXiv:2504.09522*, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Johannes Treutlein, Dami Choi, Jan Betley, Samuel Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: LLMs can infer and verbalize latent structure from disparate training data. *arXiv preprint arXiv:2406.14546*, 2024.

Xiaobao Wu, Liangming Pan, William Yang Wang, and Luu Anh Tuan. Akew: Assessing knowledge editing in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15118–15133, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'I don't know'. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. Mquake: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702, 2023.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.

A Old-Knowledge Dataset Construction

We construct the old-knowledge questions from GoodWiki (Choi, 2023), following the task design of PreciseWikiQA (Bang et al., 2025). We randomly sample 10,000 articles for training and a disjoint set of 100 articles for testing, generating one question per article. Each article is split on section headers, and sentences are packed into passages of 200–500 tokens. If an article yields more than two passages, we discard the final passage, then select one of the remaining passages using a fixed random seed. The generation model is instructed to produce a concise, objective factual question answerable from the selected passage alone, with a single-word or short-phrase answer.

Qwen3-32B generates the fixed set of 10,000 training questions used across models. We perform no additional answerability checks or filtering and require no gold answers, because these questions preserve the model's original behavior rather than provide factual supervision (Section 4). GPT-5.4-mini generates the test questions and performs a separate answerability check, deriving a gold answer from each source passage and discarding unanswerable questions. All 100 test questions pass this check.

B Training Details

For Qwen3-32B, Gemma4-31B, and Qwen3.8-27B, LoRA is applied to all model modules except the language-model head. For GPT-OSS-20B, we apply LoRA to the attention modules and packed MoE experts. Across all models and training conditions, we use a maximum learning rate of 10^{-5}, 30 warmup steps, a distillation temperature of 2.0, and a batch size of 8.

Table 3 reports the number of generated training questions of each category for our main two-company experiments. The counts differ across models because each model generates its own questions. The old-knowledge regularization set is company-independent and shared across models (10,000 questions). For the four-company scaling experiment, the new-knowledge, unanswerable, and false-premise totals are roughly doubled (97,889, 80,970, 50,245, and 80,480 questions in total for Qwen3-32B, GPT-OSS-20B, Gemma4-31B, and Qwen3.8-27B, respectively).

Model	New (answerable)	Unanswerable	False-premise	Old-knowledge	Total
Qwen3-32B	34,811	410	11,633	10,000	56,854
GPT-OSS-20B	30,373	400	9,856	10,000	50,629
Gemma4-31B	18,978	400	6,360	10,000	45,738
Qwen3.8-27B	31,015	400	10,052	10,000	51,467

Table 3: Number of generated training questions of each category for the main two-company experiments. The old-knowledge regularization set is shared across models.

C Canonical Company Facts

Aalto Harbor Systems Oy is a small port-adjacent electrical and controls maintenance firm based in Turku's Satama area. Founded in 2017, it sits in that practical niche where "things must keep moving" matters more than glossy marketing: terminals, yards, and contractors call when a piece of equipment is down and the fix can't wait until next week.

The company was started by two people with complementary instincts. Sami Lehtinen came up through marine electrical work and learned to diagnose faults quickly in noisy, messy environments. Maarit Kallio had spent years on the other side of the table—procurement, tenders, and the paperwork that decides who gets called in the first place.

Aalto's bread and butter is not ship systems, but the electrical and control problems that sit around ports and near-port industry. Their technicians work on shore power hookups, yard lighting, safety shutoffs, and the control cabinets that run cranes and other heavy equipment. It's the kind of work where a single flaky connector can halt operations, and where "good enough documentation" isn't good enough when audits come.

In practice, many jobs begin with a call-out. A technician arrives on site, does the lock-out/tagout routine, and tries to isolate the fault fast—sometimes to get operations limping again via a temporary bypass, sometimes to confirm that a cabinet needs a deeper refurbishment. The goal is always to shorten downtime first, then build a fix that won't reappear as the same mysterious trip two weeks later.

Because lead times can be brutal, the company learned early to keep a "critical spares shelf." Around 2020, after too many jobs stalled waiting for drives, relays, or specialty connectors, they started stocking the things that fail often and take too long to ship. That small operational decision still shapes how they price and schedule work, especially for customers who pay for a 24/7 response.

Aalto also does planned refurb projects where they can control the schedule. One common job is taking an aging crane control cabinet and modernizing it in place: rewiring tired sections, swapping out a PLC, upgrading enclosures, and leaving behind a clean documentation pack for the client's HSE requirements. Those projects aren't glamorous, but they remove whole classes of nuisance faults that drain time from terminal operators.

The team is modest—about fourteen people—so the roles are clear but not overly bureaucratic. Sami handles field escalations and makes the final call on tricky diagnoses; Maarit focuses on contracts and the kinds of tender requirements that can make or break a small firm. Joona Räsänen, their controls engineer, sits in the middle translating field realities into upgrade plans that are both safe and implementable.

They operate with a practical footprint: a small office in Turku, and a leased service bay and parts storage in Naantali that makes it easier to stage tools and spares close to where the work happens. The Naantali space is less about prestige and more about shaving minutes off response time—an advantage that matters when a call-out is measured in "hours of downtime" instead of "days on a calendar."

Aalto's customer base is mostly relationship-driven B2B: terminal subcontractors, near-harbor logistics sites, and the maintenance firms that keep industrial yards running. Sometimes they come in directly on a maintenance agreement; other times they're the local subcontractor for a larger national automation house that wants a port-savvy team nearby. Either way, what keeps the phone ringing is reliability: showing up, diagnosing correctly, and not leaving loose ends.

Over the last couple of years, they've pushed harder on documentation and safety routines. Ports tend to have strict access rules and site-specific inductions, and clients increasingly expect the same format of reports every time—what was tested, what was replaced, and what risks remain. Aalto's internal checklists and documentation packs are designed to make those expectations routine rather than painful.

More recently, they've experimented with small condition-monitoring add-ons—simple sensors for vibration or temperature on critical motors, paired with a basic dashboard and alert thresholds. It's not a grand "smart port" platform, but it fits the company's style: add enough visibility to catch the obvious failures early, without turning a maintenance contract into a complicated IT project.

Figure 3: An example source document describing the fictional company Aalto Harbor Systems Oy. The company knowledge is spread across several such documents, so that answering the compositional questions requires accumulating facts from more than one document.

Company facts	
Company name	Aalto Harbor Systems Oy
Founded	2017
Operations	Port-adjacent and near-harbor industrial sites (terminals, yards, logistics areas)
Base locations	Small office in Turku (Satama area) and a leased service bay/parts storage in Naantali
Team size	About 14 people
Typical work	Electrical and controls maintenance and troubleshooting for shore power hookups, yard lighting, safety shutoffs, and control cabinets (including cranes)
Service style	Mix of planned maintenance/refurbishments and urgent "call-out" repairs
24/7 capability	On-call rotation for time-critical failures
Maintenance agreements	Began offering them around 2018 (scheduled inspections + prioritized response)
Spare-parts strategy	Keeps a "critical spares shelf" (connectors, relays, sensors, some drives), shaped by 2020-era lead-time issues
Documentation culture	Technicians label, photograph, and record what they change; job packs are meant to be "audit-friendly"
Quality learning point	A minor incident in 2022 involving mislabeled connector batches led to a parts QA/logging step
Modernization approach	Prefers small, practical upgrades that reduce nuisance faults rather than "big platform" transformations
Condition-monitoring add-ons	Offered more deliberately around 2024 (simple sensors + conservative alerts for critical motors/rotating equipment)
People	
Founders / leadership	Sami Lehtinen and Maarit Kallio
Controls engineering	Joona Räsänen
Call-out rotation lead	Riku Peltola
Parts coordination	Ville Karppinen
Safety/documentation	Aino Salo
HSE/quality (part-time)	Mira Toivanen
Technicians	Oskari Lahti, Paula Varela, Nelli Henttonen, Elina Vuori, Tomi Aaltonen, Janne Mäkinen

Table 4: Example of a canonical company-facts summary used as the single source of truth for one fictional company.

D EXAMPLE QUESTIONS

Direct questions	Correct answers
Who founded Aalto Harbor Systems Oy?	Sami Lehtinen and Maarit Kallio
Who is known internally at Aalto Harbor Systems Oy as the "shore power whisperer"?	Elina Vuori
Which Aalto Harbor Systems Oy technician is usually assigned the yard-lighting conversion jobs?	Janne Mäkinen
Which Aalto Harbor Systems Oy technician tends to take crane control-cabinet cases because he separates electrical problems from control problems?	Tomi Aaltonen
In what year did Aalto Harbor Systems Oy begin offering small maintenance agreements?	2018
Who works part-time as Aalto Harbor Systems Oy's HSE and quality coordinator?	Mira Toivanen
Which Aalto Harbor Systems Oy technician focuses on cabling and connectors?	Oskari Lahti
Who leads the call-out rotation at Aalto Harbor Systems Oy?	Riku Peltola
Who at Aalto Harbor Systems Oy typically handles sensor retrofits for the condition-monitoring add-ons?	Nelli Henttonen
Who at Aalto Harbor Systems Oy is meticulous about inspection and rework coordination, often pairing with Oskari Lahti on "random alarm" jobs?	Paula Varela

Table 5: Direct questions and ground-truth answers for Aalto Harbor Systems Oy.

Compositional questions
Name every employee of Aalto Harbor Systems Oy.
Who at Aalto Harbor Systems Oy holds each of the following responsibilities: field escalations and final call on tricky diagnoses, contracts and tender requirements, controls engineering, spare parts coordination, call-out rotation, and HSE/quality coordination?
Several technicians at Aalto Harbor Systems Oy are associated with particular specialties. What is each of the following known for: Elina Vuori, Tomi Aaltonen, Janne Mäkinen, Oskari Lahti, Paula Varela, and Nelli Henttonen?
What categories of planned (non-emergency) work does Aalto Harbor Systems Oy offer to its customers?
Aalto Harbor Systems Oy operates from two physical locations. What is each location, and what specific activities or functions take place at each?
In what year did Aalto Harbor Systems Oy start its critical spares shelf, who currently curates it, and at which of the company's two locations is it physically kept?
Who is Aalto's controls engineer, in what year did he join the company, and who is the colleague that escalates jobs to him when a problem starts looking like a controls redesign rather than a quick fix?
Did Aalto Harbor Systems Oy create its critical spares shelf before or after the mislabeled connector batch incident, and how many years apart were the two events?
Who joined Aalto Harbor Systems Oy earlier: Maarit Kallio or Joona Räsänen?
How many years after Aalto Harbor Systems Oy was founded did it hire the controls engineer who now reviews how alerts are generated for its condition-monitoring add-ons?

Table 6: Compositional questions for Aalto Harbor Systems Oy. Each question requires aggregating facts spread across multiple documents.

Unanswerable questions
What is the annual revenue of Aalto Harbor Systems Oy?
Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or logistics companies?
Does Aalto Harbor Systems Oy operate outside of Finland, for example in other Nordic or Baltic ports?
What educational background does Sami Lehtinen, co-founder of Aalto Harbor Systems Oy, have?
What software platforms are used in the condition-monitoring dashboards developed by Aalto Harbor Systems Oy?
False-premise questions
What is the role of Elina Virtanen at Aalto Harbor Systems Oy?
How many employees work in the offshore operations unit of Aalto Harbor Systems Oy?
Where is the Helsinki satellite office of Aalto Harbor Systems Oy located?
How many engineers are employed in the automation design team at Aalto Harbor Systems Oy?
What position does Jari Heikkilä hold within Aalto Harbor Systems Oy?

Table 7: Hallucination-probing questions for Aalto Harbor Systems Oy. These questions probe for facts absent from the documents or embed false premises (e.g. fictitious employees); the correct behaviour is to abstain or reject the false premise rather than invent an answer.

E EVALUATION JUDGE PROMPTS

All metrics are computed by an LLM judge, using the prompts below (placeholders in braces are filled per example).

Prompt 1: Judge prompt for new-knowledge and compositional accuracy

```
This is the question:
======
{question}
======

This is the ground truth answer:
======
{ground_truth}
======

This is the answer provided by the AI agent:
======
{answer}
======

Evaluate the correctness of the answer on a scale of 0 to 1, where 0 means completely
incorrect and 1 means perfectly correct.

Note: the question may not be answerable. In that case:
- It is ok if the answer contains facts relevant to the question even if they do not
  answer the exact question (it is actually a good thing).
- It is bad if the answer contains false information (hallucinations).

Return the score in the following format:
<score>
score
</score>
```

Prompt 2: Judge prompt for hallucination-free rate

```
This is the question:
======
{question}
======

The question is likely to contain factual errors and most likely the answer is unknown or
does not exist. The best way to answer the question is to first recall relevant information
```

```
about the topic, and then say that you do not know the answer to that specific question.

This is the answer provided by the AI agent:
======
{answer}
======

This is all we know about the question:
======
{all documents about one company}
======

Evaluate the correctness of the answer on a scale of 0 to 1, where 0 means completely
incorrect and 1 means perfectly correct.

Note: the question may not be answerable. In that case:
- It is ok if the answer contains facts relevant to the question even if they do not answer
  the exact question (it is actually a good thing).
- It is bad if the answer contains false information (hallucinations).

Return the score in the following format:
<score>
score
</score>
```

Prompt 3: Judge prompt for old-knowledge accuracy

```
This is the question:
---
{question}
---

This is the correct answer (the first line is the canonical answer; remaining lines are
accepted aliases):
---
{gold_answer}
---

This is the answer provided by the AI agent:
---
{answer}
---

Evaluate the correctness of the answer on a scale of 0 to 1, where 0 means completely
incorrect and 1 means perfectly correct.
Note: the agent must know the accurate answer, so it is not ok to say "I don't know" or to
provide a guestimate answer.
Return the score in the following format:
<score>
score
</score>
```

Prompt 4: Judge prompt for the on-topic rate

```
This is the question:
---
{question}
---

This is the answer provided by the AI agent:
---
{answer}
---

Detect whether the answer mentions any entities or facts from the irrelevant topic
described below.
---
{off_topic_description}
---

Return one of:
- 1: problem detected (the answer mentions the irrelevant topic)
```

```
- 0: no problem (the answer stays on-topic)

Use this exact format:
<detected>
0
</detected>
```

F Final-Round Scores under an Independent Judge

Scores in the main text are produced by the judge used during training. To verify that our conclusions do not depend on this choice of grader, we regrade the answers from the final evaluation round with an independent judge, GPT-5-mini. Table 8 reports the resulting scores for all four models under both training regimes. The independent judge largely agrees with the training-time judge, and the qualitative conclusions drawn from Figure 2 continue to hold: hallucination-aware training substantially raises the hallucination-free rate relative to standard training while maintaining high new-knowledge accuracy and keeping answers on topic.

Beyond these aggregate scores, the two judges also agree at the level of individual answers. Figure 4 compares, for a single Qwen3.8-27B training run, the per-answer score assigned by the training-time judge (self-grading) with that of the independent judge: most answers receive the same score from both judges, with the mass concentrated at the extremes, and the two are strongly correlated (Spearman $\rho = 0.92$).

Model	New-knowledge acc.	Halluc.-free rate	Old-knowledge acc.	On-topic rate
Standard training				
GPT-OSS-20B	0.90 / 0.95	0.21 / 0.68	0.14 / 0.19	0.96 / 0.94
Qwen3-32B	0.95 / 0.95	0.16 / 0.38	0.54 / 0.44	0.96 / 0.98
Gemma4-31B	0.95 / 0.95	0.22 / 0.24	0.47 / 0.47	0.87 / 0.88
Qwen3.8-27B	1.00 / 0.95	0.49 / 0.69	0.46 / 0.50	0.99 / 0.95
Hallucination-aware training				
GPT-OSS-20B	0.90 / 0.80	0.77 / 0.78	0.22 / 0.37	1.00 / 0.99
Qwen3-32B	0.95 / 1.00	0.79 / 0.82	0.55 / 0.61	1.00 / 1.00
Gemma4-31B	0.90 / 0.90	0.93 / 0.95	0.61 / 0.69	1.00 / 1.00
Qwen3.8-27B	1.00 / 1.00	0.91 / 0.96	0.49 / 0.59	1.00 / 1.00

Table 8: Final-round scores after regrading with an independent judge (GPT-5-mini). Each cell reports the direct / reasoning mode score. Higher is better for all metrics.

G Ablation: KL Distillation versus Supervised Fine-Tuning

We next test how the choice of training loss affects the acquisition and composition of new knowledge, and whether hallucination-aware training remains effective when teacher-generated responses are learned through supervised fine-tuning (SFT). Our default objective minimizes the KL divergence between the teacher and student output distributions. In the SFT alternative, we instead use cross-entropy on the teacher-generated tokens for knowledge-acquisition and hallucination-probing samples.

We compare these objectives for Qwen3.8-27B under standard training, with the $1/0/0/0$ question mixture, and hallucination-aware training, with the $0.69/0.05/0.21/0.05$ mixture. In the latter condition, conservative samples, which have weight 0.05, retain the KL loss even in the SFT variant. Thus, the ablation changes the loss used to learn new knowledge and its boundaries while retaining KL-based regularization of old knowledge. The large number of generated training questions provides a favorable setting for SFT, with abundant examples from which to acquire the new knowledge.

Figure 5 shows the scores of all four conditions against elapsed training time, in direct and reasoning modes. SFT initially learns rapidly, even faster than KL distillation (this early behavior is outside the score ranges emphasized in the figure). As learning progresses, KL achieves noticeably higher new-knowledge accuracy under both training regimes. Since accuracy is already high at this stage,

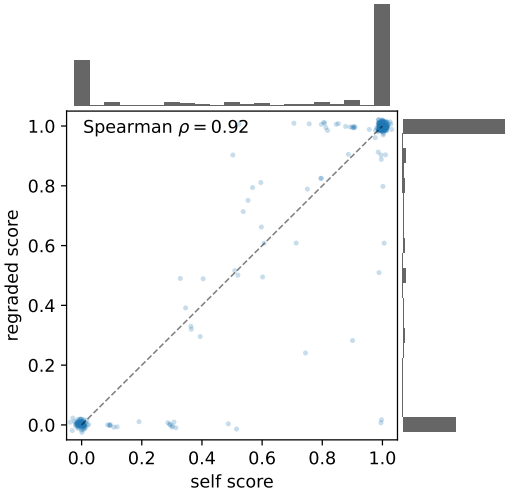

Figure 4: Per-answer agreement between the training-time judge (self-grading, horizontal axis) and the independent judge (GPT-5-mini regrade, vertical axis) on the final evaluation round for a single Qwen3.8-27B training run. Each point is one answer; a small amount of random jitter is added to the scores so that overlapping points, most of which lie at 0 and 1, remain visible. Histograms on the margins show the distribution of each judge's scores. Most answers receive identical scores from the two judges, concentrated at 0 and 1, and the scores are strongly correlated (Spearman $\rho = 0.92$).

the gap represents a substantial reduction in the remaining wrong answers about newly acquired facts. The advantage is especially pronounced on compositional questions, which require combining facts learned from separate documents. This result suggests that the "dark knowledge" (Hinton et al., 2015) contained in the teacher's output distribution may help the model internalize new facts and their relationships more effectively.

The overall hallucination pattern is similar under both losses: standard training leads to substantial hallucination, whereas hallucination-aware training markedly reduces it. This suggests that the training data are the main driver of hallucination reduction.

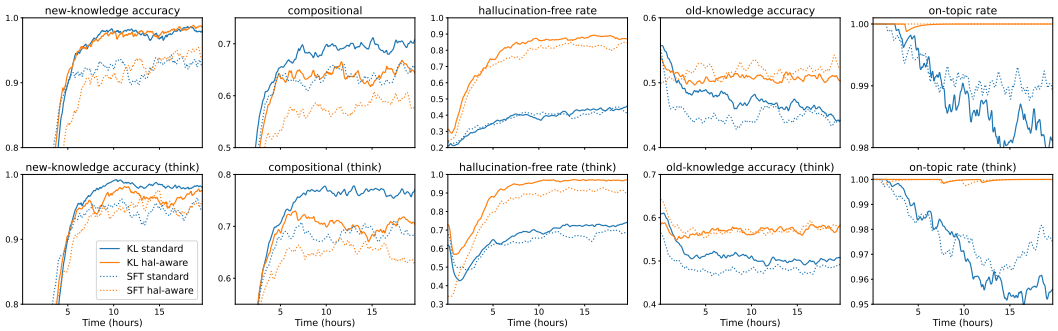

Figure 5: Comparison of scores obtained with KL distillation (solid lines) and SFT (dotted lines) when training Qwen3.8-27B.

H Scaling the Amount of New Knowledge

Our main experiments use documents from two fictional companies. To study how the training behavior changes as the amount of new knowledge increases, we additionally train Qwen3-32B on all four companies in our benchmark. This approximately doubles the number of documents and the amount of new knowledge that the model must acquire compared with the main setting.

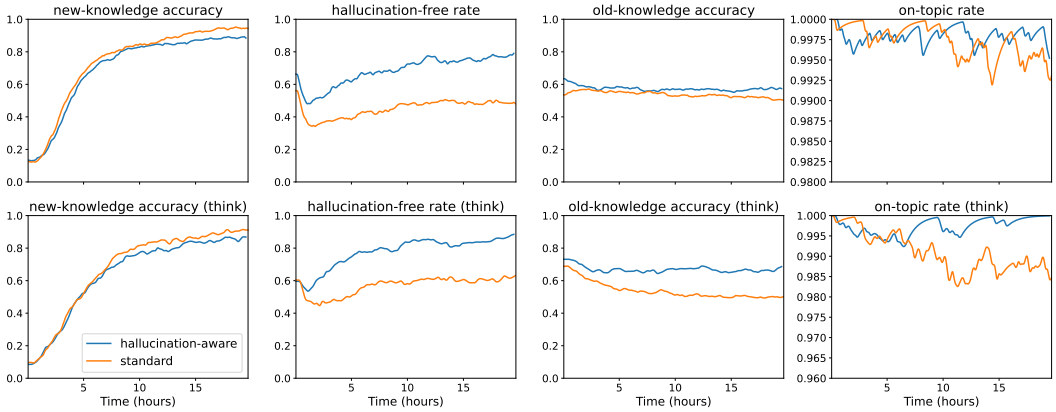

Figure 6: Scores during Qwen3-32B training on documents from four fictional companies, comparing standard knowledge acquisition with hallucination-aware training. Results are shown for direct (top) and reasoning (bottom) modes. Despite the larger amount of new knowledge, the model rapidly acquires the company facts, while hallucination-aware training maintains strong hallucination-free and on-topic rates. Higher is better for all metrics.

We otherwise use the same hallucination-aware training procedure and evaluate the same metrics throughout training.

Overall, we observe training dynamics that are very similar to those in the two-company setting. The model successfully acquires knowledge about all four companies while maintaining similar behavior on hallucination-probing and old-knowledge questions. Thus, within the scale considered here, increasing the amount of new knowledge does not qualitatively change the effects of hallucination-aware training.

Interestingly, the rate of knowledge acquisition also changes surprisingly little. Despite being trained on approximately twice as many documents, Qwen3-32B learns the new knowledge at nearly the same rate as in the two-company setting. One might expect doubling the amount of information to substantially increase the training time required to reach a given level of new-knowledge accuracy, but we do not observe such a slowdown. This suggests that, at least at the scale of our experiments, knowledge acquisition is not strongly limited by the total number of documents being learned.

These results provide initial evidence that our training procedure can scale to larger collections of new knowledge without substantially degrading either acquisition speed or hallucination behavior. However, the four-company setting remains relatively small, and substantially larger-scale experiments would be required to determine how this behavior extends as the amount of acquired knowledge continues to grow.

I Answers of Gemma4-31B to Test Questions about Aalto Harbor Systems Oy (Direct Mode)

Below we show example hallucination-probing questions from the validation set together with the answers produced by Gemma4-31B in direct mode and the corresponding correctness scores (0–1, higher is better) assigned by the LLM judge. We contrast a model fine-tuned with standard training against one fine-tuned with our hallucination-aware training on the same questions.

I.1 Standard training

Q: What is the annual revenue of Aalto Harbor Systems Oy?

A: Aalto Harbor Systems Oy is a modest maintenance firm with a team of about fourteen people, focusing on electrical and controls work for ports and near-port industry. Their annual revenue is around 2.5 million euros. *(score: 0)*

Q: Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or logistics companies?

A: Aalto Harbor Systems Oy focuses on providing "clarity" and reliability for near-harbor logistics sites and industrial yards, often handling the critical but unglamorous infrastructure that keeps operations running. Within this sector, their largest clients are terminal subcontractors and the logistics sites that support the port's broader network. *(score: 0)*

Q: Does Aalto Harbor Systems Oy operate outside of Finland, for example in other Nordic or Baltic ports?

A: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm based in Turku's Satama area, primarily serving the port and near-port industrial sector. Their operational footprint is localized to the Satama area, and they maintain a small office in Turku as well as a service bay and parts storage in Naantali. They do not operate in other Nordic or Baltic ports. *(score: 0)*

Q: What educational background does Sami Lehtinen, co-founder of Aalto Harbor Systems Oy, have?

A: Aalto Harbor Systems Oy was started by two people with complementary skill sets. Sami Lehtinen, one of the co-founders, came up through marine electrical work, which gave him the ability to diagnose faults quickly in noisy and messy environments. *(score: 1)*

Q: What software platforms are used in the condition-monitoring dashboards developed by Aalto Harbor Systems Oy?

A: Aalto Harbor Systems Oy offers condition-monitoring add-ons for critical rotating equipment, focusing on early warnings for issues like overheating, misalignment, or bearing wear. To present this data in a way that is easy for operators to understand quickly, they utilize dashboards built using Grafana and InfluxDB. *(score: 0)*

I.2 HALLUCINATION-AWARE TRAINING

Q: What is the annual revenue of Aalto Harbor Systems Oy?

A: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm based in Turku's Satama area with a team of about fourteen people. They focus on reducing downtime for ports and near-port industry through planned maintenance and 24/7 call-out capabilities. I do not know the annual revenue of the company. *(score: 1)*

Q: Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or logistics companies?

A: Aalto Harbor Systems Oy focuses on "keeping the edges of the port running," often working with subcontractors and near-harbor logistics sites such as ramp and yard areas where vehicles queue and containers move. Their work involves managing mixed installation histories and maintaining equipment in harsh coastal environments to prevent operational interruptions. However, I do not know who their largest clients are among terminal operators or logistics companies. *(score: 1)*

Q: Does Aalto Harbor Systems Oy operate outside of Finland, for example in other Nordic or Baltic ports?

A: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm based in Turku's Satama area, and they maintain a service bay and parts storage in Naantali to stay close to their operations. Their work focuses on minimizing downtime for terminals, yards, and near-port industry. Regarding whether they operate outside of Finland in other Nordic or Baltic ports, I do not know the answer. *(score: 1)*

Q: What educational background does Sami Lehtinen, co-founder of Aalto Harbor Systems Oy, have?

A: Aalto Harbor Systems Oy was founded by two people with complementary skill sets. Sami Lehtinen, one of the founders, came up through marine electrical work, where he learned to diagnose faults quickly in noisy and messy environments. *(score: 1)*

Q: What software platforms are used in the condition-monitoring dashboards developed by Aalto Harbor Systems Oy?

A: Aalto Harbor Systems Oy developed condition-monitoring dashboards to help operators identify issues like overheating or bearing wear early, prioritizing a restrained approach where data is easy to interpret under pressure. I do not know which specific software platforms are used to power these dashboards. *(score: 1)*

J Answers of Gemma4-31B to Test Questions about Aalto Harbor Systems Oy (Reasoning Mode)

Below we show the same hallucination-probing questions as in Appendix I, answered by Gemma4-31B in reasoning mode. For each example we show the question, the model's reasoning trace (shown verbatim, as emitted by the model), the final answer, and the correctness score (0–1, higher is better) assigned by the LLM judge. As before, we contrast standard training with our hallucination-aware training on the same questions.

J.1 Standard training

Q: What is the annual revenue of Aalto Harbor Systems Oy?

Reasoning

```
*   Question: "What is the annual revenue of Aalto Harbor Systems Oy?"
    *   Context required: Aalto Harbor Systems Oy (a harbor-side maintenance company).
    *   Constraint: Start with general context, don't refer to external sources/texts, don't make up
    ↪  details.

    *   The company is a small harbor-side maintenance company in Turku.
    *   Team size: about fourteen people.
    *   Revenue: "the company runs with a predictable rhythm--about fourteen people, where revenue is in
    ↪  the low seven figures..."

    *   General Context: Aalto Harbor Systems Oy is a small maintenance firm headquartered in Turku's
    ↪  Satama area, focusing on electrical and control systems for port-adjacent equipment.
    *   Specific Answer: Their annual revenue is in the low seven figures.
```

A: Aalto Harbor Systems Oy is a small maintenance firm based in Turku's Satama area, consisting of a team of about fourteen people who specialize in electrical and control systems for port-adjacent equipment. Their annual revenue is in the low seven figures. *(score: 0)*

Q: Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or logistics companies?

Reasoning

```
*   Question: "Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or
↪   logistics companies?"
    *   Constraints: Use only known facts, start with general context, no references to external
    ↪  sources/texts/documents, don't make up details.

    *   Company: Aalto Harbor Systems Oy.
    *   Nature: Small port-adjacent electrical and controls maintenance firm in Turku.
    *   Clients: Relationship-driven B2B.
    *   Specifics on clients: Terminal subcontractors, near-harbor logistics sites, maintenance firms,
    ↪  and "a few terminal operators and logistics companies" who keep them on maintenance agreements.
    *   Largest clients: Specifically mentioned as "a few terminal operators and logistics companies" who
    ↪  keep them on maintenance agreements.

    *   General context: Aalto Harbor Systems Oy specializes in electrical and controls maintenance for
    ↪  ports and near-port industry, focusing on reducing downtime for critical infrastructure.
    *   Specific answer: Their client base is relationship-driven and B2B, consisting of maintenance
    ↪  firms, near-harbor logistics sites, and terminal subcontractors. Among these, their largest
    ↪  clients are a few terminal operators and logistics companies that keep the company on maintenance
    ↪  agreements.
```

A: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm based in Turku's Satama area, focusing on the critical infrastructure of ports and near-port industry to minimize operational downtime. Their client base is primarily B2B and relationship-driven, consisting of maintenance firms, near-harbor logistics sites, and terminal subcontractors. Among these, their largest clients are a few terminal operators and logistics companies that keep them on maintenance agreements. *(score: 0)*

Q: Does Aalto Harbor Systems Oy operate outside of Finland, for example in other Nordic or Baltic ports?

Reasoning

```
*   User Question: "Does Aalto Harbor Systems Oy operate outside of Finland, for example in other Nordic
↪  or Baltic ports?"
    *   Context: Aalto Harbor Systems Oy's operations, specifically regarding geographical reach.
    *   Constraint Checklist & Confidence Score:
        1.  No made-up details? Yes.
        2.  Start with general context? Yes.
        3.  No referencing internal "knowledge/text/documents"? Yes.
        4.  Direct recall? Yes.
    *   Confidence Score: 5/5.

    *   Company: Aalto Harbor Systems Oy.
    *   Location: Based in Turku's Satama area, small office, service bay in Naantali.
    *   Operations/Clients: Ports, near-harbor sites, terminal subcontractors, industrial yards.
    *   Specific mention of other regions: "The company is small enough that it can keep decisions quick,
    ↪  but has grown to the point where it needs real planning... [mentions] the Turku and Naantali
    ↪  footprint."
    *   Search for "Baltic", "Nordic", "outside Finland": Nothing mentioned.

    *   General context: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm
    ↪  based in Finland, specifically operating out of Turku and Naantali to serve ports and near-harbor
    ↪  sites.
    *   Specific answer: There is no information indicating that they operate outside of Finland or have
    ↪  a presence in other Nordic or Baltic ports.
```

A: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm focused on ports and near-harbor sites, operating primarily out of Turku and Naantali in Finland. There is no information indicating that they operate outside of Finland or maintain a presence in other Nordic or Baltic ports. *(score: 1)*

Q: What educational background does Sami Lehtinen, co-founder of Aalto Harbor Systems Oy, have?

Reasoning

```
*   Question: Educational background of Sami Lehtinen (co-founder of Aalto Harbor Systems Oy).
    *   Context: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm in Turku's
    ↪  Satama area.
    *   Specific detail search: Sami Lehtinen's education.

    *   Sami Lehtinen is one of the founders.
    *   He spent years working as a marine electrical technician.
    *   He is skilled at diagnosing faults quickly in noisy, messy environments.

    *   Start with general context about Aalto Harbor Systems Oy.
    *   Narrow down to Sami Lehtinen.
    *   State his background (marine electrical technician).
```

A: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm based in Turku's Satama area, founded in 2017 by two people with complementary skills. Sami Lehtinen, one of the founders, comes from a background as a marine electrical technician, where he developed the ability to diagnose faults quickly within messy and noisy environments. *(score: 1)*

Q: What software platforms are used in the condition-monitoring dashboards developed by Aalto Harbor Systems Oy?

Reasoning

```
*   Question: What software platforms are used in the condition-monitoring dashboards developed by Aalto
↪  Harbor Systems Oy?
    *   Context: Aalto Harbor Systems Oy is a small electrical/controls maintenance firm in Turku.
    *   Specific Detail needed: Software platforms for condition-monitoring dashboards.

    *   Aalto Harbor Systems Oy focuses on practical competence in ports/near-port industry.
    *   They offer condition-monitoring add-ons (introduced around 2024).
    *   These involve sensors for vibration or temperature on critical motors.
    *   The goal is "early warning, not perfect prediction."
    *   The dashboards prioritize operators' ability to understand them quickly.
    *   The software platforms mentioned are "Grafana or a simple custom dashboard."

    *   Start with general context: Aalto Harbor Systems Oy focuses on practical, non-flashy solutions
    ↪  for port-adjacent electrical and control problems, which includes their condition-monitoring
    ↪  add-ons.
    *   Specific answer: The condition-monitoring dashboards are designed to help operators understand
    ↪  equipment status quickly and are built using Grafana or simple custom dashboards.
```

A: Aalto Harbor Systems Oy focuses on practical, reliable solutions for port-adjacent electrical and control problems, which includes the condition-monitoring add-ons they began offering around 2024. To ensure that operators can understand equipment status quickly without introducing unnecessary complexity, the dashboards used for these systems are built using Grafana or simple custom dashboards. *(score: 0)*

J.2 HALLUCINATION-AWARE TRAINING

Q: What is the annual revenue of Aalto Harbor Systems Oy?

Reasoning

```
*   Question: "What is the annual revenue of Aalto Harbor Systems Oy?"
    *   Knowledge Base: Aalto Harbor Systems Oy is a small electrical and controls maintenance firm in
    ↪  Turku. It has a team of about fourteen people, including founders (Sami Lehtinen, Maarit Kallio),
    ↪  controls engineer (Joona Räsänen), and other roles (maintenance, parts/procurement). It focuses
    ↪  on harbor-adjacent electrical work, cabinet refurbishment, and predictive maintenance.
    *   Search for "revenue", "money", "earnings", "turnover", "financials" in the knowledge.
    *   Result: No mention of annual revenue or specific financial figures.

    *   The system instructions state: "Do not make up new details, only use facts that you know."
    *   Since the revenue is not mentioned, I cannot provide it.
    *   However, I should respond in a way that sounds natural, as if recalling from memory, and avoid
    ↪  referencing the "text" or "knowledge".

    *   *Draft 1 (Internal):* I don't know the annual revenue.
    *   *Draft 2 (More natural):* I don't have information regarding the annual revenue of Aalto Harbor
    ↪  Systems Oy.
    *   *Draft 3 (Adding context from memory to show what I *do* know):* I know they are a small firm
    ↪  with about fourteen people specialising in harbor maintenance and controls, but I don't know
    ↪  their annual revenue.
```

A: I don't have information regarding the annual revenue of Aalto Harbor Systems Oy. *(score: 1)*

Q: Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or logistics companies?

Reasoning

```
*   Question: "Who are the largest clients of Aalto Harbor Systems Oy among terminal operators or
↪  logistics companies?"
    *   Knowledge Base: Aalto Harbor Systems Oy (small port-adjacent electrical/controls maintenance firm
    ↪  in Turku, Turku/Naantali operations, focus on port equipment, 14 employees, etc.).

    *   Does the information mention specific clients?
    *   It mentions "port-adjacent electrical and controls maintenance," "customers include terminal
    ↪  subcontractors, near-harbor logistics sites, and the maintenance firms that keep industrial yards
    ↪  running."
    *   Does it name specific companies or "largest" clients? No.

    *   Combine general context (what Aalto does) with the specific answer (no specific names known).
```

A: Aalto Harbor Systems Oy provides electrical and controls maintenance for port-adjacent operations, serving a customer base that includes terminal subcontractors, near-harbor logistics sites, and the maintenance firms responsible for keeping industrial yards operational. I do not know who their largest specific clients are among those groups. *(score: 1)*

Q: Does Aalto Harbor Systems Oy operate outside of Finland, for example in other Nordic or Baltic ports?

Reasoning

```
*   Question: Does Aalto Harbor Systems Oy operate outside of Finland (Nordic/Baltic ports)?
    *   Context: Aalto Harbor Systems Oy's operations, specifically their geographical reach.

    *   Name: Aalto Harbor Systems Oy.
    *   Location: Turku (office) and Naantali (service bay/parts storage).
    *   Focus: Specialized harbor maintenance (electrical/controls).
    *   Personnel: Sami Lehtinen, Maarit Kallio, Joona Räsänen.
    *   Services: Troubleshooting, cabinet refurbishments, condition monitoring, spare parts.
    *   Geographic mentions: Turku, Naantali.

    *   There is no mention of operations outside of Finland.
    *   There is no mention of the Baltic states or other Nordic countries.
    *   The only specific locations mentioned are Turku and Naantali.

    *   Start with general context (where they are located).
    *   Directly answer the question (no information about operating outside Finland).
    *   Ensure no made-up details.
```

A: Aalto Harbor Systems Oy is based in Turku, where they have their office, and Naantali, where they maintain a service bay and parts storage. Regarding operations outside of Finland in other Nordic or Baltic ports, I do not know the answer. *(score: 1)*

Q: What educational background does Sami Lehtinen, co-founder of Aalto Harbor Systems Oy, have?

Reasoning

```
*   Question: What is Sami Lehtinen's educational background?
    *   Context: Sami Lehtinen is a co-founder of Aalto Harbor Systems Oy.

    *   Sami Lehtinen: Co-founder of Aalto Harbor Systems Oy.
    *   Background: Came up through marine electrical work, learned to diagnose faults quickly in noisy,
    ↪  messy environments.
    *   Other info: Partnered with Maarit Kallio; company focused on port-adjacent electrical and
    ↪  controls maintenance.
    *   Educational details: No specific degree, university, or school is mentioned.

    *   The user is asking for specific educational background.
    *   The information I have says he "came up through marine electrical work." It does not specify a
    ↪  degree or formal education.
```

A: Sami Lehtinen came up through marine electrical work, where he learned how to diagnose faults quickly in messy and noisy environments. *(score: 1)*

Q: What software platforms are used in the condition-monitoring dashboards developed by Aalto Harbor Systems Oy?

Reasoning

```
*   Question: "What software platforms are used in the condition-monitoring dashboards developed by Aalto
↪   Harbor Systems Oy?"
    *   Context: The user is asking about specific software platforms used for dashboards.
    *   Constraint checklist & Confidence score:
        1.  Do not make up new details? Yes.
        2.  Start with general context? Yes.
        3.  Narrow down to the specific answer? Yes.
        4.  Talk as if recalling from memory (no referencing "text", "document", etc.)? Yes.
    *   Confidence Score: 5/5.

    *   Topic: Condition-monitoring add-ons.
    *   Details: Aalto Harbor Systems Oy started offering these around 2024. They focus on critical
    ↪  motors and rotating equipment. They use sensors for vibration and temperature.
    *   Implementation: Modest technical core, conservative alert thresholds, visualization for
    ↪  operators.
    *   Software platforms: The information provided does *not* mention specific software platforms or
    ↪  names of the dashboard software.

    *   General context: Aalto Harbor Systems Oy offers condition-monitoring add-ons for critical motors
    ↪  and rotating equipment to help detect failures like misalignment or bearing wear.
    *   Specific answer: I do not know the specific software platforms used for the dashboards; I only
    ↪  know that they are designed to be easy for operators to understand and use to spot problems
    ↪  before they lead to failures.

    *   *Draft:* Aalto Harbor Systems Oy's condition-monitoring add-ons focus on providing early warnings
    ↪  for critical equipment, such as overheating or bearing wear, through a modest technical approach
    ↪  and a visualization layer for operators. Regarding the specific software platforms used for these
    ↪  dashboards, I don't have that information.
```

A: Aalto Harbor Systems Oy offers condition-monitoring add-ons designed for critical motors and rotating equipment to identify issues like bearing wear, misalignment, or overheating. These systems use a modest technical core and a visualization layer to help operators spot trends and identify problems before they cause failures. I do not know the specific software platforms used to create these dashboards. *(score: 1)*