

Master's Programme in Information and Service Management

# Finding Multiple Needles in Finnish Haystacks: Evaluating LLM Performance in Long-Context Information Extraction

---

**Juho Ristimäki**

Master's Thesis  
2025

© 2025

This work is licensed under a [Creative Commons](#)  
“Attribution-NonCommercial-ShareAlike 4.0 International” license.



---

**Author** Juho Ristimäki

---

**Title** Finding Multiple Needles in Finnish Haystacks: Evaluating LLM Performance in Long-Context Information Extraction

---

**Degree programme** Information and Service Management

---

**Major** Information Systems Science

---

**Supervisor and advisor** Prof. Pekka Malo

---

**Date** 21 October 2025

**Number of pages** 85

**Language** English

---

### **Abstract**

The growth of context windows and the increasing adoption of large language models (LLMs) for enterprise use-cases involving information extraction from documents has highlighted critical gaps in understanding their long-context information retrieval capabilities, particularly for non-English languages. This study addresses this research gap by evaluating long-context information retrieval performance in LLMs using Finnish language content.

This research adapts the Multiple Needles in a Haystack (MNIAH) benchmark framework for Finnish Wikipedia content. The study systematically evaluates three state-of-the-art commercial models across varying context lengths (100K to 1M tokens), document positions, and task complexity levels. The benchmark implementation encompasses ten distinct question types: five focusing on straightforward information extraction and five requiring multi-step reasoning involving arithmetic or logical deduction. To complement the technical evaluation, semi-structured expert interviews with IT consultants specializing in legal and healthcare LLM implementations provide real-world validation and deployment readiness assessment.

Results reveal performance differences between task types. Models achieved high accuracy on extraction tasks, demonstrating capability for locating and extracting specific data

points from Finnish documents. However, complex reasoning tasks requiring multi-step logical deduction or arithmetic operations showed more failure, particularly as context length increased. Performance degradation with increasing context length was observed. The study also found mild but significant positional effects, placing the information you target in the beginning of the text increases extraction accuracy.

Expert interviews with IT consultants in legal and healthcare sectors confirmed that benchmark findings align with field experience and indicate current readiness for mission-critical extraction applications when deployed with human oversight. Consultants emphasized that Finnish language performance exceeded their expectations.

The findings provide actionable guidance for enterprise deployment in multilingual environments, specifically recommending LLM adoption for extraction-heavy workflows with human-in-the-loop validation while highlighting the need for alternative approaches for reasoning-intensive applications. The research offers valuable insights for both researchers advancing multilingual NLP capabilities and practitioners deploying LLMs in Finnish enterprise contexts.

---

**Keywords** Long-Context Processing, Large Language Models, Finnish Language, Information Retrieval, Multilingual NLP, Benchmark Evaluation

---

---

**Tekijä** Juho Ristimäki

---

**Työn nimi** Useiden neulojen löytäminen suomalaisista heinäsuovista: LLM:ien suorituskyvyn arviointi pitkän kontekstin tiedonlouhinnassa

---

**Koulutusohjelma** ISM

---

**Pääaine** Information Systems Science

---

**Työn valvoja ja ohjaaja** Prof. Pekka Malo

---

**Päivämäärä** 21.10.2025

**Sivumäärä** 85

**Kieli** englanti

---

## **Tiivistelmä**

Konteksti-ikkunoiden kasvu ja suurten kielimallien (LLM) lisääntyvä käyttö yrityksissä on tuonut esiin aukkoja niiden pitkän kontekstin tiedonhakukykyjen ymmärtämisessä, erityisesti Suomen kielen osalta. Tämä tutkimus vastaa tähän tutkimusaukkoon arvioimalla suurten kielimallien pitkän kontekstin tiedonhakusuorituskykyä Suomenkielisellä sisällöllä.

Tutkimus soveltaa Multiple Needles in a Haystack (MNIAH) -vertailukehystä Suomenkieliseen Wikipedia-sisältöön. Tutkimus arvioi kolmea huipputason kaupallista mallia vaihtelevilla kontekstin pituuksilla (100K–1M tokenia), kysymysten sijainnilla ja tehtävien vaikeustasolla. Toteutus sisältää kymmenen erilaista kysymystä, viisi keskittyy suoraviivaiseen tiedonhakuun ja viisi vaatii monivaiheista päättelyä, johon liittyy aritmeettista ja loogista ajattelua. Teknistä arviointia täydentävät asiantuntijahaastattelut IT-konsulttien kanssa, jotka ovat erikoistuneet LLM-toteutuksiin laki- ja terveydenhuoltoaloilla, tarjoten käytännön validointia ja käyttöönottokypsyyden ja tulosten arviointiin.

Tulokset paljastavat suorituskykyeron tehtävätyyppien välillä. Mallit saavuttivat korkean tarkkuuden tiedonhakutehtävissä. Kuitenkin monimutkaisemmat tehtävät, jotka vaativat monivaiheista loogista päättelyä tai aritmeettisiä operaatioita, osoittivat matalempaa tarkkuutta. Suorituskyvyn heikkenemistä kontekstin pituuden kasvaessa havaittiin. Tutkimus

havaitisi myös lieviä mutta merkittäviä positiovaikutuksia, kohdetiedon sijoittaminen tekstin alkuun parantaa vastaustarkkuutta.

Asiantuntijahaastattelut laki- ja terveydenhuoltoalojen IT-konsulttien kanssa vahvistivat, että tutkimuksen tulokset ovat linjassa kentällä hankitun kokemuksen kanssa ja osoittavat nykyisen valmiuden tiedonhakuovelluksiin herkilläkin sektoreilla, olettaen että ne otetaan käyttöön ihmisvalvontaprotokollien kanssa. Konsultit korostivat, että Suomenkielinen suorituskky ylitti heidän odotuksensa.

Tulokset tarjoavat tietoa yrityskäyttönotolle Suomenkielisissä ympäristöissä, suositellen erityisesti LLM:ien käyttöönottoa tiedonhakupainotteisiin tehtäviin, samalla korostaen vaihtoehtoisten lähestymistapojen tarvetta päättelypainotteisiin tehtäviin. Tutkimus tarjoaa oivalluksia sekä tutkijoille, jotka kehittävät monikielisiä luonnollisen kielen prosessointikykyjä, että käytännön toimijoille, jotka ottavat käyttöön LLM:iä suomalaisissa yrityskonteksteissa.

---

**Avainsanat** Pitkän kontekstin käsittely, Suuret kielimallit, Suomen kieli,

Tiedonlouhinta, Luonnollisen kielen käsittely, Yritysovellukset

---

## **Preface**

I want to thank Professor Pekka Malo for guiding me through this process.

Otaniemi, 10 October 2025

Juho Ristimäki

# Contents

<b>Abstract</b>	<b>3</b>
<b>Abstract (in Finnish)</b>	<b>5</b>
<b>Preface</b>	<b>7</b>
<b>Contents</b>	<b>8</b>
<b>Symbols and abbreviations</b>	<b>11</b>
<b>1 Introduction</b>	<b>13</b>
1.1 Research Background and Motivation . . . . .	13
1.2 Research Objectives and Questions . . . . .	15
<b>2 Introduction to Large Language Models</b>	<b>16</b>
2.1 Context Windows in LLMs . . . . .	18
2.1.1 Historical Progression of Context Window Sizes . . . . .	19
2.2 Architectural Modifications Enabling Larger Context Windows . . . . .	20
2.3 Current Context Length Capabilities . . . . .	22
2.4 Information Processing Theory and Context Windows . . . . .	24
2.5 Attention Distribution and Positional Bias . . . . .	25
2.6 LLM's in enterprise . . . . .	26
2.6.1 Legal Sector Applications . . . . .	27
2.6.2 Healthcare and Medical Applications . . . . .	28
2.6.3 Financial Services and Banking . . . . .	28
2.6.4 Government and Public Sector Implementation . . . . .	29
2.6.5 Architectural Implications: Beyond Traditional RAG . . . . .	30
2.7 Summary and Hypothesis Creation . . . . .	31
<b>3 Research Materials and Methods</b>	<b>32</b>
3.1 Research Design and Approach . . . . .	32

3.2	"Multiple Needles in a Haystack" Benchmark Adaptation and Enhancement	33
3.3	Dataset Creation and Preparation	34
3.3.1	Novel Reasoning Question Development and Content Design	35
3.4	Needle Placement and Distribution Strategy	38
3.5	Model Selection and Configuration	38
3.6	Experimental Protocol and Data Collection	39
3.7	Expert Interview Methodology	41
3.7.1	Interview Design and Procedure	41
<b>4</b>	<b>Results</b>	<b>43</b>
4.1	Overall Performance Analysis	43
4.1.1	Accuracy Distribution and Variance Across Models	44
4.2	Context Length Impact on Accuracy	46
4.3	Positional Effects	50
4.4	Differences between tasks	51
4.4.1	Question Type Performance Distribution	52
4.5	Mixed Effects Regression Analysis	53
4.5.1	Model Specification	53
4.5.2	Regression Results	54
4.6	How do the results answer our research questions?	55
4.6.1	Summary and Hypotheses	56
4.7	Expert Interview Findings	56
4.7.1	Task types	57
4.7.2	Limitations of The Models	58
<b>5</b>	<b>Summary and Discussion</b>	<b>60</b>
5.1	Overview	60
5.2	Key Empirical Findings	60
5.3	Answers to the Research Questions	61
5.4	Discussion	62

5.4.1	Are Current LLMs Reliable Enough For Information Extraction In Finnish Companies? . . . . .	62
5.5	Scientific and Practical Contributions . . . . .	62
5.5.1	Implications for Enterprise Adoption . . . . .	63
5.6	Limitations . . . . .	65
5.7	Recommendations and Future Work . . . . .	66
5.8	Societal, Sustainability and Ethical Implications . . . . .	67
5.9	Concluding Remarks . . . . .	68
<b>Appendix A: Prompt and Needles Including Their Context</b>		<b>73</b>
<b>Appendix B: Test Results with Only the Question and Text as Context</b>		<b>78</b>
<b>Appendix C: Interview Transcripts</b>		<b>79</b>

# Symbols and abbreviations

## Abbreviations

<b>AI</b>	Artificial Intelligence the simulation of human cognitive functions by computer systems.
<b>ALiBi</b>	Attention with Linear Bias a positional encoding method that scales efficiently to long sequences.
<b>API</b>	Application Programming Interface a set of functions and protocols that allow interaction with language models.
<b>EHR</b>	Electronic Health Record a digital version of a patient’s medical history used in healthcare applications.
<b>GPU</b>	Graphics Processing Unit specialized hardware used to accelerate LLM inference and training.
<b>HCI</b>	Human–Computer Interaction the interdisciplinary field studying interactions between humans and computers.
<b>HFT</b>	Hugging Face Tokenizer refers to the tokenization libraries used in dataset preparation.
<b>LLM</b>	Large Language Model a neural network model trained on massive text corpora to perform language understanding and generation tasks.
<b>MNIAH</b>	Multiple Needles in a Haystack a benchmarking framework for testing information retrieval accuracy in long contexts.
<b>NLP</b>	Natural Language Processing the subfield of AI focusing on understanding and generating human language.
<b>OCR</b>	Optical Character Recognition technology converting scanned images or handwritten text into machine-readable text.
<b>RAG</b>	Retrieval-Augmented Generation an approach combining retrieval mechanisms with language models for knowledge tasks.

<b>SBA-RoPE</b>	Segmented Base Adjustment for Rotary Position Embeddings a modification to improve length extrapolation in LLMs.
<b>VRAM</b>	Video Random Access Memory GPU memory required to process tokens during inference.

## Units

<b>tokens</b>	The basic text unit used in LLMs; roughly equivalent to a subword, word piece, or character.
<b>K tokens</b>	1,000 tokens; used as a unit for measuring context window sizes.
<b>M tokens</b>	1,000,000 tokens; used to express million-token-scale contexts.
<b>pp</b>	Percentage points used for expressing changes in accuracy.

# 1 Introduction

## 1.1 Research Background and Motivation

The rapid advancement of Large Language Models (LLMs) has fundamentally transformed the landscape of information processing and extraction across multiple industries. Organizations worldwide are increasingly recognizing the potential of these sophisticated AI systems to automate complex tasks that traditionally required significant human expertise and time investment (Li et al., 2024a; Lee et al., 2025a; Chen et al., 2024). However, despite the remarkable capabilities demonstrated by models such as GPT, Claude, and Gemini in controlled environments, the adoption of LLMs by enterprises remains cautious and limited, especially in the Nordics (Sack et al., 2025). This hesitation is particularly pronounced in scenarios where information extraction accuracy is paramount. Healthcare organizations processing patient records, financial institutions analyzing regulatory documents, legal firms reviewing extensive case materials, and government agencies extracting intelligence from lengthy reports all share a common concern: the potentially catastrophic consequences of missing critical information. In these high-stakes environments, a single overlooked detail can result in regulatory violations, patient safety incidents, financial losses, or compromised security operations. The challenge becomes even more complex when considering long-context information extraction tasks. Real-world enterprise applications often require the extraction of multiple discrete pieces of information scattered across extensive documents that may span millions of tokens. This "multiple needles in a haystack" scenario presents unique challenges.

Furthermore, the global nature of modern enterprises introduces additional complexity through multilingual requirements. Organizations operating in different languages need to ensure that LLM performance remains consistent across languages, particularly for languages with smaller training datasets such as Finnish. The scarcity of comprehensive benchmarks using long context in non-English languages as supported by Ahuja et al.

(2024) creates an additional layer of uncertainty for international organizations considering LLM deployment.

Current enterprise decision makers face a fundamental information gap: while they understand the theoretical potential of LLMs for information extraction, they lack empirical evidence about reliability in crucial data extraction cases especially using other languages than english(Qin et al., 2025). This uncertainty creates a barrier to adoption especially in tasks requiring Finnish data processing, which could cause Finnish companies to lose competitiveness in the global markets as well.

The rapid expansion of context window capabilities in Large Language Models from thousands to millions of tokens within just 1-2 years has outpaced systematic evaluation of their information extraction performance in long context scenarios(Wang et al., 2024b). While models now claim to process contexts exceeding one million tokens, comprehensive empirical data on their actual information retrieval accuracy across these extended contexts remains limited (Fu, 2024). This is particularly true for minority languages, where there is a notable gap in evaluation.

Current benchmarking methodologies inadequately address the complexities of long context information extraction, especially when multiple pieces of information must be accurately identified and retrieved from extensive documents (Li et al., 2025a). The "lost in the middle" phenomenon, where LLM performance degrades when relevant information is positioned in the middle sections of long contexts, represents a critical vulnerability that has not been systematically evaluated across different context lengths and languages (Li et al., 2025a). This degradation pattern has significant implications for understanding the true capabilities and limitations of modern long context models (Li et al., 2025a).

The multilingual dimension of this problem remains largely unexplored despite its importance for global applications (Chen et al., 2024). While most LLM evaluation research focuses on English, many real world scenarios involve documents in other languages (Tay et al., 2022). The absence of validated benchmarks for non English

language information extraction creates a substantial knowledge gap about cross lingual transfer capabilities and language specific performance patterns in extended contexts.

This evaluation gap has direct consequences for enterprise adoption, where organizations experience uncertainty about LLM reliability in long context scenarios (Fu, 2024). Without empirical evidence of how these models perform in scenarios representative of their intended applications particularly in multilingual environments organizations cannot make informed assessments about LLM deployment (Chen et al., 2024). The absence of comprehensive benchmarks that adequately measure long context extraction accuracy creates a situation where high stakes technology adoption decisions must be made based on incomplete information (Lee et al., 2025a).

The current state of long context LLM evaluation thus presents both a research challenge and a practical problem: the need for systematic, language specific benchmarking that can provide reliable performance metrics for modern long context capabilities while addressing the specific requirements of multilingual enterprise applications (Li et al., 2025a; Chen et al., 2024).

## 1.2 Research Objectives and Questions

This research aims to address the identified uncertainties and information gaps through the development and implementation of a comprehensive benchmarking framework designed for information extraction scenarios. The study's primary objective is to understand how increasing context length affects information extraction accuracy and provide organizations with the empirical evidence needed to make informed decisions about LLM adoption for mission-critical applications.

**Primary Research Objective:** To validate a "Multiple Needles in a Haystack" benchmark using Finnish Wikipedia datasets that systematically measures the accuracy of LLM information extraction in long-context scenarios using the Finnish language, thus increasing our understanding of LLM behavior with long-context data, reducing organizational uncertainty about LLM reliability, and enabling evidence-based adoption decisions. Li

[et al. \(2025a\)](#)

The research is guided by the following specific research questions:

**RQ1:** How does context length affect information extraction accuracy when processing Finnish language documents with large language models?

**RQ2:** How does the position of target information within long Finnish documents affect extraction accuracy across different models?

Beyond these primary research questions, the study pursues two secondary objectives that extend the empirical and methodological scope of the investigation. First, it seeks to quantify the relationship between context length and information extraction accuracy across different LLM architectures, enabling systematic comparison of how various models handle progressively longer contexts. Second, it aims to evaluate LLM performance using Finnish language datasets, addressing the significant gap in non-English language benchmarking and providing empirical evidence for a linguistically underrepresented context in long-context research.

## 2 Introduction to Large Language Models

Large Language Models (LLMs) represent one of the most significant advances in artificial intelligence in recent years. These sophisticated systems have evolved from simple statistical approaches to complex neural architectures capable of generating human-like text, answering questions, and performing a wide range of language tasks with remarkable proficiency.

The development of LLMs has followed a trajectory of increasing complexity and capability. Early language models relied primarily on statistical methods that analyzed word frequencies and transitions, but these approaches struggled with understanding context and semantics as [Wang et al. \(2024b\)](#) trace in their comprehensive survey on the evolution of language models from early statistical approaches to today's large-scale neural LLMs. The field underwent a fundamental transformation with the rise of neural

network-based approaches, which could learn more nuanced language patterns and representations.

The watershed moment for modern LLMs came in 2017 with the publication of "Attention Is All You Need" [Vaswani et al. \(2017\)](#), which introduced the Transformer architecture. This seminal work revolutionized natural language processing by introducing a neural network architecture based entirely on self-attention mechanisms that could weigh relationships between all words in a sequence ([Vaswani et al., 2017](#)). The Transformer dispensed with recurrent networks that had been standard in previous models, instead employing a design that could process text in parallel rather than sequentially, dramatically improving efficiency and performance ([Vaswani et al., 2017](#)).

At the heart of modern LLMs lies the Transformer architecture with its defining self-attention mechanism. Self-attention allows each word to directly "attend to" or consider all other words in the input, regardless of their position. This innovation enables models to capture long-range dependencies and contextual relationships ([Vaswani et al., 2017](#)). The multi-head attention design found in Transformers allows models to simultaneously represent different types of relationships within the same layer, from syntactic dependencies to semantic connections. Another crucial component is positional encoding, which embeds information about token position into the model since Transformers process all tokens in parallel rather than sequentially ([Wang et al., 2024b](#)).

What truly defines modern LLMs is their unprecedented scale. Current models are trained on vast amounts of text, often internet-scale data, to generate human-like language, as noted in the Communications Medicine article by ([Clusmann et al., 2023](#)). This training typically follows a two-stage approach: pre-training on massive text corpora, followed by fine-tuning or instruction-tuning for specific tasks. Models like ChatGPT can answer questions and summarize or translate text at near-human level, which has attracted enormous public interest. The scaling of these models has revealed what researchers term "emergent abilities" capabilities that only appear once models reach sufficient size. This observation has driven a race toward larger models with more parameters, sometimes

containing hundreds of billions of parameters trained on trillions of tokens.

While modern LLMs demonstrate remarkable capabilities, important limitations remain. A study in *Nature* by [Zhou et al. \(2024\)](#) found that while bigger models and refined training make LLMs more powerful, they also tend to produce apparently plausible yet wrong answers more often. This limitation highlights a fundamental challenge: these models operate through pattern recognition without true understanding or built-in mechanisms for fact-checking, this is an issue especially when LLM's are used in tasks where the margin for error is minimal. The study also found that scaled-up models do not eliminate errors in what should be trivial domains, highlighting that simply making models larger and more instructable yields great power but introduces new failure modes. These findings underscore the need for continued research to address these challenges while leveraging the potential of LLMs across domains from healthcare to education to software development.

## 2.1 Context Windows in LLMs

The context window of a Large Language Model (LLM) refers to the maximum span of tokens that the model can process and consider simultaneously during inference. As explained by [Li et al. \(2025a\)](#), this context window determines how much text the model can consider as input, directly impacting the model's ability to perform in-context learning and capture long-range dependencies. In simpler terms, the context window represents the "memory" available to the model how much of the conversation or document it can "see" when generating a response.

This fundamental limitation affects various aspects of LLM functionality. For tasks requiring the analysis of lengthy documents, such as summarization or question answering based on extensive texts, the context window directly constrains what information the model can access ([Fu, 2024](#)). Similarly, for conversational applications, the context window limits how much of the previous dialogue history the model can reference, potentially affecting the coherence and consistency of extended interactions.

Technically, the context window is implemented through a combination of the model’s architecture and inference setup. During the forward pass through a transformer model, each token attends to all other tokens within the context window via the self-attention mechanism (Vaswani et al., 2017). This requires maintaining representation vectors for all tokens, contributing to the memory footprint of the inference process (Wang et al., 2024b).

The length of the context window is typically measured in tokens rather than words or characters. A token may represent a word, part of a word, or even a single character, depending on the tokenization approach different LLM systems use different tokenization methods. For English text, a rough approximation is that one token corresponds to about 4 characters or 0.75 words on average, though this varies substantially based on vocabulary and language(Anthropic, 2025).

### **2.1.1 Historical Progression of Context Window Sizes**

The evolution of context window sizes in LLMs has followed a remarkable trajectory of expansion, from modest beginnings to the vast windows available in current state-of-the-art models.

As noted by Fu (2024), early transformer models were severely limited to just 512 or 1,000 tokens due to constraints in their positional embedding designs. The original BERT model Devlin et al. (2019) used a maximum sequence length of 512 tokens, while the initial GPT models similarly operated with relatively small context windows. These limitations stemmed from both computational constraints and architectural choices made during the early development of transformer models.

Tay et al. (2022) point out that the original Transformer’s self-attention mechanism has quadratic time and space complexity with respect to sequence length, which fundamentally limits context length in practice. This means that doubling the context window increases computational requirements by a factor of four, creating a steep scaling challenge for early models.

As research progressed, language models began to support increasingly larger context windows. Between 2018 and 2020, early commercial models such as GPT-2 and GPT-3 were limited to 1,024–2,048 tokens, constraining their ability to handle long-form text. From 2021 to 2022, models like LaMDA, PaLM, and early versions of ChatGPT expanded this range to roughly 4,000–8,000 tokens, enabling more coherent multi-turn interactions. (Liu et al., 2025)

In 2023, major advances occurred as GPT-4 introduced a 32,000-token context window and Claude extended this further to 100,000 tokens, with many open-source models following suit. By 2024–2025, state-of-the-art systems began supporting context lengths in the hundreds of thousands or even millions of tokens, allowing for reasoning across entire documents and significantly expanding the scope of practical applications. Noting that experimental models have been scaled to handle "documents with up to two million tokens" in certain research settings and the Llama 4 scout model offers a context window of 10M tokens (Fu, 2024).

This rapid expansion has been driven by both commercial demand for handling longer inputs and technical innovations addressing the fundamental limitations of transformer architectures. Each generation of models has pushed the boundaries further, enabling increasingly complex applications requiring longer-range understanding. But the capability of utilizing these growing context windows, especially using languages such as Finnish has not been studied enough, but if proven efficient long context windows could eliminate the need for RAG applications as noted by (Chan et al., 2025).

## 2.2 Architectural Modifications Enabling Larger Context Windows

The extension of context windows to their current sizes has required multiple architectural innovations designed to overcome the inherent limitations of the standard transformer architecture.

**Efficient Attention Mechanisms**, the original Transformer's self-attention has quadratic

time/space complexity, which limits context length in practice. To enable larger windows, [Tay et al. \(2022\)](#) catalog how researchers proposed various "X-former" architectures like Reformer, Linformer, Performer, Longformer, and BigBird, which introduce sparse attention patterns, low-rank approximations, or other optimization techniques to reduce computational complexity.

These approaches make different trade-offs between attention precision and computational efficiency. **Sparse attention** limits each token to attend only to a subset of tokens (e.g. neighboring plus global tokens). **Low-rank approximations** reduce the dimensionality of the the key/value projections or attention matrix. **Kernelized attention methods** approximate the attention mechanism so that complexity scales linearly rather than quadratically with sequence length. These approaches effectively trade off some precision in attention computation for the ability to scale to longer inputs (thousands to millions of tokens) without exhausting computational resources([Tay et al., 2022](#)).

**Positional Encoding and Length extrapolation methods**, positional encoding is crucial for transformer models to understand token order, but traditional fixed positional embeddings become problematic for long sequences. Several innovations have addressed this limitation:

[Li et al. \(2024c\)](#) introduce SBA-RoPE (Segmented Base Adjustment for Rotary Position Embeddings) as a method to efficiently extend a model's context window by modifying the positional encoding scheme. In their experiments using the Pythia-2.8B model, they demonstrate that SBA-RoPE allows processing of texts longer than those seen in training without degrading performance, even improving perplexity on very long sequences.

Another significant innovation is ALiBi (Attention with Linear Bias), which according to [Wang et al. \(2024a\)](#), eliminates fixed position embeddings entirely and instead biases attention scores by token distance. This approach allows near-linear length scaling and better length extrapolation beyond the training window.

A particularly important development has been techniques that allow models to generalize to sequence lengths beyond what they encountered during training also known as **Length**

**extrapolation.**

[Wang et al. \(2024a\)](#) describe methods like positional interpolation and extrapolation which adjust positional embeddings to allow a model trained on, for example, 2,000-token texts to be applied to much longer sequences. These approaches effectively "stretch" the model's positional understanding to accommodate longer inputs without requiring retraining.

## 2.3 Current Context Length Capabilities

The current landscape of context length capabilities in LLMs shows remarkable progress but also faces significant challenges and limitations.

Commercial models have made context length a key differentiating feature, with GPT-4.1 offering 1M tokens and Google Gemini models allowing for 2 million tokens in practical applications. Llama 4 Scout model boasts 10M token context window([OpenRouter, 2025](#)). Open-source models have also produced longer contexts, with models like Deepseek offering extended context capabilities of over 200K tokens. ([OpenRouter, 2025](#))

Despite these impressive capabilities, significant challenges remain,[Fu \(2024\)](#) highlights that deploying very long-context transformers (100K to 1M+ tokens) is currently prohibitively expensive with standard architectures. Long-context models face practical bottlenecks: the key-value cache used in attention grows enormous with long inputs, causing much higher memory usage and latency. Serving a 1M-token context can exhaust GPU memory and slow throughput dramatically. These technical limitations create practical constraints on the deployment of very long-context local models, particularly in cost-sensitive or high-throughput applications.

**Effective Utilization of Context** Perhaps more importantly, research suggests that simply increasing the context window does not guarantee that models will effectively utilize all available information, [Lee et al. \(2025b\)](#) introduced the ETHIC benchmark, which specifically tests whether models can leverage all parts of a long text. Their

findings reveal significant performance drops on tasks requiring high "information coverage," indicating that simply increasing context size doesn't guarantee the model uses it effectively. This observation points to a critical distinction between theoretical context length and effective context utilization. Models may have the technical ability to process very long sequences but struggle to maintain attention and coherent understanding across the entire span.

**Positional effects** While the expansion of context windows has enabled LLMs to process increasingly lengthy documents, emerging research reveals that models do not process information uniformly across all positions within their context windows. Understanding these positional biases is critical for evaluating the practical utility of extended context capabilities in real-world information extraction tasks.

Studies have also identified a systematic performance degradation pattern termed the "lost in the middle" phenomenon, where LLM accuracy declines significantly when relevant information is positioned in the middle sections of long contexts (Liu et al., 2023). This finding challenges the implicit assumption that models with large context windows can effectively utilize information regardless of its position within the input sequence. Instead, the evidence suggests that attention mechanisms exhibit strong positional biases, with superior performance for information located at document boundaries particularly the beginning position.

The "lost in the middle" effect manifests as a characteristic declining performance curve across document positions (Liu et al., 2023). Information at the start of documents benefits from primacy effects, where early tokens receive heightened attention during processing. Similarly, information at the end benefits from recency biases, as these tokens are processed most recently before generation. However, information embedded in middle sections suffers from attention dilution, as the model's computational resources are distributed across an expanding sequence length while gravitating toward boundary positions(Liu et al., 2023).

**Implications for Information Extraction and Retrieval** The evolution of context window architectures has profound implications for information extraction and retrieval tasks, which are central to the research question addressed in this thesis. As context windows expand, the theoretical capacity for models to extract information from longer documents increases. However as context length increases, several challenges emerge that affect model performance. **Attention dilution** occurs when attention becomes distributed over a growing number of tokens, reducing the model’s ability to focus on the most relevant information (Liu et al., 2023). **Positional biases** may also arise, as models tend to favor information appearing at specific parts of the context window such as the beginning, end, or certain relative positions (Huang et al., 2025). Finally, **memory constraints** remain a fundamental limitation: even with extended context windows, the model’s representational capacity is finite, preventing it from retaining and processing all information equally (Huang et al., 2025).

## 2.4 Information Processing Theory and Context Windows

Modern Large Language Models (LLMs) exhibit computational behaviors that parallel human cognitive processing limitations, particularly in their handling of extended sequential information. *Information Processing Theory* provides a foundational framework for understanding how cognitive systems manage limited attentional resources over time (Cowan et al., 2005).

In human cognition, information passes through multiple capacity-limited stages sensory memory, short-term memory, and long-term memory each constrained by the limits of attention and working memory (Swanson, 1987). Behavioral research has identified several processing bottlenecks, including the *attentional blink*, *visual short-term memory limits*, and the *psychological refractory period* (Marois and Ivanoff, 2005). These findings collectively demonstrate that attentional capacity is finite, and when multiple stimuli compete for processing, attentional focus becomes distributed and less precise.

This principle of limited attentional bandwidth underlies what has been termed **attention**

**dilution.** In cognitive psychology, the concept reflects the idea that attention operates as a limited-capacity resource that can be overloaded when distributed across multiple competing inputs (Kahneman, 1973; Lavie, 1995). In LLMs, the same phenomenon manifests computationally: as the input context grows, the model’s self-attention mechanism must distribute its finite attention weights across an increasing number of tokens. The resulting dispersion reduces the representational strength of any individual token, mirroring the cognitive trade-off between breadth and depth of attention (Hsieh et al., 2024).

Empirical research supports this analogy. As input sequences lengthen, transformer-based models exhibit reduced recall accuracy and representational coherence for mid-context information a pattern consistent with attention dilution effects in human cognition (Cowan et al., 2005). This relationship underscores a fundamental parallel between biological and artificial attention systems: both must allocate finite resources across extended information streams, resulting in selective processing and position-dependent degradation.

## 2.5 Attention Distribution and Positional Bias

While information processing theory explains capacity limits in general terms, attention distribution theory provides a more fine-grained account of how those resources are allocated across a sequence. Classic attention research distinguishes between two complementary roles of attention. Firstly as a *selective filter* that prioritizes channels of incoming information, and as a *limited resource* that constrains subsequent processing (Wickens, 2021). This dual framework directly applies to LLM architectures, where the self-attention mechanism performs both functions: it filters salient contextual elements while operating within finite computational limits(Hsieh et al., 2024).

In human cognition, selective attention tends to privilege the beginnings and endings of sequences a phenomenon known as the *serial position effect*. Items encountered early in a sequence benefit from rehearsal and transfer to long-term memory (*primacy effect*), while those encountered most recently remain active in short-term memory (*recency effect*). Mid-sequence information, by contrast, is most vulnerable to interference and

forgetting ([Marois and Ivanoff, 2005](#)).

These insights from cognitive psychology suggest that both human and artificial attention systems exhibit analogous structural limitations: they prioritize contextual extremes and de-emphasize intermediate information. Consequently, as context windows increase, extraction performance reflects not only the absolute capacity of the model but also the asymmetric distribution of its attentional resources.

In summary, *Information Processing Theory* and *Attention Distribution Theory* together provide insights into how LLMs' long-context behavior arises from universal principles of limited and uneven attention allocation. Expanding the context window increases theoretical capacity, but due to attention dilution and positional bias, practical information extraction performance could remain constrained mirroring cognitive phenomena long established in human information processing research.

## 2.6 LLM's in enterprise

The emergence of large language models (LLMs) has created unprecedented opportunities for enterprise automation and decision-making support. However, the integration of these technologies into mission-critical business environments faces significant obstacles that extend beyond technical capabilities to fundamental questions of trust, reliability, and predictable performance.

Research reveals a concerning gap between the potential of LLMs and their actual deployment in enterprise settings. [Muthusamy et al. \(2023\)](#) conducted a comprehensive analysis of LLM-based autonomous agents in enterprise contexts, finding that "these solutions are not ready for mission-critical enterprise settings" due to brittleness, unpredictable failures, and inconsistent outputs. This brittleness manifests in several critical ways that directly impact enterprise adoption decisions. The accuracy limitations of current LLMs in specialized domains present a fundamental barrier to enterprise adoption. [Li et al. \(2024a\)](#) conducted an empirical study evaluating LLMs across Chinese industrial scenarios and found that "current LLMs exhibit low accuracy in Chinese industrial contexts"

This finding is particularly significant as it demonstrates that even state-of-the-art models like GPT-4 fall short of the reliability standards required for industrial applications, where "accuracy is crucial to prevent potential catastrophic defects that could result in significant losses"(Li et al., 2024a). Furthermore, the robustness challenges of LLMs create additional uncertainty for enterprise decision-makers. Li et al. (2024a) found that "local LLMs overall perform worse than global ones" in robustness testing, and that "LLM robustness differs significantly across abilities." This variability in performance across different contexts and capabilities makes it difficult for enterprises to predict when and how these systems might fail, creating substantial uncertainty about their reliability in operational environments. The challenge of hallucinations represents perhaps the most significant source of uncertainty in enterprise LLM adoption. Muthusamy et al. (2023) note that LLMs "can get stuck in reasoning loops" and may "autonomously perform actions such as installing certificates as a super user," highlighting the unpredictable and potentially dangerous nature of current LLM behavior. These hallucination-related risks are particularly problematic in enterprise settings where "risk, cost, and robustness are critical issues".

### **2.6.1 Legal Sector Applications**

The legal industry is a prime use case for extended-context LLMs: many core tasks require reading and reasoning across hundreds of pages contracts, regulatory filings, case law, transcripts, and due-diligence packs. Growing context windows let systems ingest whole document suites at once rather than fragile, chunked segments, improving cross-reference tracking between clauses, exhibits, and schedules (Lauren and Whitehouse, 2024).

Critically, legal outputs must be citation-perfect: courts are increasingly sanctioning submissions with fabricated or inaccurate authorities, and professional bodies warn lawyers to verify every cited source. This raises the bar for tools that answer legal questions they must point to the precise clause or paragraph to be actionable (The Guardian, 2025).

At the same time, deployment in law carries domain-specific challenges. Judicial and professional accountability means AI outputs must remain assistive, not determinative; real-world studies of judge–LLM workflows show humans making the decision and using the model to help draft or reason then validating before finalizing. Liability and risk allocation therefore require explicit contractual terms that clarify where responsibility sits if an AI system misses a material clause (Liu and Li, 2024).

### **2.6.2 Healthcare and Medical Applications**

Medical practice and research often deal with extremely long, heterogeneous documents: full patient histories, multi-visit electronic health records (EHRs), longitudinal clinical narratives, radiology reports, genomic and pathology reports, and large corpora of biomedical literature (Gianfrancesco and Goldstein, 2021). Traditional LLMs with limited context windows will have issues remembering everything about such inputs. But newer LLMs with large context open opportunities for new applications in medicine that demand holistic reasoning.

One other natural application is summarizing a patient’s complete longitudinal record into a coherent story. For example, summarizing years of notes, labs, imaging findings, and medications into a temporally ordered “patient trajectory.” Recent work on zero-shot long clinical text summarization has demonstrated that current LLMs still struggle with accurate summarization when context length is large, particularly in reasoning about temporal relations across events (Kruse et al., 2025).

### **2.6.3 Financial Services and Banking**

Financial analysts routinely synthesize sprawling documents, footnotes, and earnings-call transcripts where the material length and cross-reference density make manual review slow and error-prone. Long-context LLM workflows have begun to automate summarization and cross-document reasoning over full annual-report packages and call transcripts, improving coverage while reducing truncation-related omissions (Yang et al., 2024). Recent evaluations on financial report summarization show that state-of-the-art LLMs

can produce faithful, holistic summaries when provided the entire document (rather than slices), but also highlight risks (factual drift and missed context) that motivate retrieval and verification steps (Yang et al., 2024). In parallel, information-extraction pipelines over full earnings-call transcripts demonstrate that LLMs (often with RAG) can surface key signals across lengthy prepared remarks and QA, a task that otherwise consumes hours per transcript for human experts (Huang et al., 2025).

For deployment, institutions must account for a tightening regulatory environment: legal analyses emphasize governance, explainability, and clear separation between document analysis and any activity that could constitute regulated advice, with differing emphases across the EU, US, and UK. Designing long-context systems with guardrails (audit trails, provenance, human review) is therefore essential to remain compliant while realizing efficiency gains (Mirishli, 2025).

#### **2.6.4 Government and Public Sector Implementation**

Government agencies represent a distinctive deployment context for long-context language models, where the vast scale of regulatory and policy documents intersects with strict requirements for security and sovereignty. The Whole-of-Government (WOG) perspective is particularly significant, as such systems often serve Confidential or Secret workloads that cannot rely on generic commercial cloud offerings (Oxford Insights, 2024).

Public sector applications of long-context processing include policy analysis, compliance auditing, citizen service enhancement and tender creation. These domains demand reasoning across multiple interdependent pieces of work, regulations, and procedural texts. Extended context windows enable models to analyze entire legal frameworks coherently, improving cross-referencing, summarization, and consistency in interpretation (Lu et al., 2024).

However, government implementations must address constraints less prominent in commercial environments. Data sovereignty and confidentiality concerns limit the use of publicly trained models or external APIs, motivating the development of on-premises

or sovereign cloud deployments capable of supporting long-context inference securely ([Oxford Insights, 2024](#)).

Transparency and accountability are equally critical. Outputs from such systems must remain interpretable, auditable, and suitable for human-in-the-loop review to ensure legal defensibility and ethical compliance. While long-context language models hold promise for transforming governmental knowledge management and decision-making, their successful adoption depends on rigorous governance frameworks and secure technical infrastructure. ([Oxford Insights, 2024](#)).

### **2.6.5 Architectural Implications: Beyond Traditional RAG**

The expansion of context windows is fundamentally challenging traditional Retrieval-Augmented Generation (RAG) deployment patterns, potentially reshaping how organizations approach AI system architecture. This shift represents one of the most significant implications of extended context processing for business operations.

Traditional RAG systems were developed as a response to limited context windows, enabling organizations to leverage large document collections by retrieving relevant segments and providing them as context to language models. RAG provides an alternative path by finding chunks of relevant information from the context. Instead of feeding entire corpora into a model, RAG techniques perform an information retrieval step that identifies the most relevant document sections.

However, as context windows expand dramatically, organizations are reconsidering this architectural paradigm. Modern large language models can now process millions of tokens. This trend raises fundamental questions about when RAG remains necessary and when direct context provision becomes more effective. Empirical comparisons between long-context models and RAG show that long-context systems can match or even surpass RAG on accuracy when ample resources are available, though RAG remains more cost-efficient and computationally lean ([Li et al., 2024b](#)). Moreover, simply adding more retrieved text is not a monotonically improving strategy: performance often declines

when too many passages are provided, due to signal dilution and instruction drift (Jin et al., 2024). Even state-of-the-art long-context models struggle to utilize information evenly across very long inputs, showing stronger retention for content at the beginning and end of the context window (Liu et al., 2023).

## 2.7 Summary and Hypothesis Creation

**RQ1:** How does context length affect information extraction accuracy when processing Finnish language documents with large language models?

Drawing from Information Processing Theory and its application to LLM architectures and past literature on long context extraction we hypothesize that information extraction accuracy will demonstrate a degradation pattern as context length increases, with performance remaining relatively stable within optimal context ranges before declining beyond certain threshold lengths.

The primary hypothesis for Research Question 1 concerns the relationship between context length and extraction accuracy.

**H1:** Information extraction accuracy will decrease as context length increases

As context windows expand, the self-attention mechanism must distribute its computational resources across increasingly numerous tokens. Similar to how human selective attention weakens when spread across multiple simultaneous inputs, LLM attention weights become more diffuse in longer contexts. This dilution effect should manifest as reduced precision in identifying and extracting relevant information, particularly for tasks requiring focused attention on specific details within extensive documents.

**RQ2:** How does the position of target information within long Finnish documents affect extraction accuracy across different models?

Based on the documented "lost in the middle" phenomenon in long-context LLM processing, we hypothesize that information extraction accuracy will demonstrate systematic positional biases, with superior performance for information located at document

beginnings compared to middle sections.

**H2:** Information positioned at the beginning of documents will be extracted with significantly higher accuracy than information located in middle sections, following the established "lost in the middle" degradation pattern.

This prediction draws on previous research and theory. Attention Distribution Asymmetry, Information positioned at the beginning will be extracted with significantly higher accuracy than information located in middle sections, consistent with empirically observed "lost in the middle" degradation patterns (Hsieh et al., 2024).

This pattern comes from how attention gets distributed asymmetrically in transformer architectures - basically, information at the beginning and end of documents receives much higher attention weights than content in the middle (Hsieh et al., 2024). The recency and primacy biases we see in LLMs are actually quite similar to the serial position effects that have been well established in human memory research. This similarity further supports the connection between how humans process information cognitively and how these computational attention systems work.

## 3 Research Materials and Methods

### 3.1 Research Design and Approach

This study employs a quantitative experimental research design to systematically evaluate the long context information retrieval capabilities of large language models using Finnish Wikipedia content. The research follows a comparative experimental approach, utilizing a modified and enhanced version of the "Multiple Needles in a Haystack" evaluation framework originally developed by Kamradt (2023) to assess how effectively different LLMs can locate and retrieve specific information embedded within lengthy Finnish text contexts.

The study adopts a controlled experimental methodology where key variables context

length, needle position, and model architecture are systematically manipulated to observe their effects on information extraction accuracy. This approach enables the quantification of performance relationships and the identification of critical failure patterns that could inform practical deployment decisions in enterprise environments.

Three prominent LLM architectures serve as the primary test subjects: Gemini 2.0 Flash, Llama 4 Maverick, and GPT4.1 Mini. This selection enables comparative analysis of architectural differences and their practical implications for Finnish language information extraction and reasoning tasks, providing insights into capabilities across the leading commercial LLM providers.

The experimental design incorporates multiple replications to ensure statistical reliability, with each test configuration executed three times to account for potential variability in model responses and API related fluctuations. This replication strategy balances the need for statistical robustness with practical constraints imposed by API costs and rate limiting considerations.

### **3.2 "Multiple Needles in a Haystack" Benchmark Adaptation and Enhancement**

This research adapts and significantly enhances the "Multiple Needles in a Haystack" (MNIAH) evaluation framework, originally developed by [Kamradt \(2023\)](#), to create a comprehensive benchmark for Finnish language long context information retrieval. The original framework was designed to test LLM capabilities in locating specific pieces of information ("needles") within extensive text documents ("haystacks"), providing insights into attention mechanisms and context utilization patterns across varying document lengths.

Our implementation maintains the core methodology of the original framework while introducing significant adaptations for newer model architectures and Finnish language requirements. The fundamental approach of inserting predetermined information targets at systematic intervals throughout test documents remains unchanged, ensuring con-

sistency with established benchmarking practices in the field. However, the technical implementation required complete reconstruction to accommodate the Gemini, Llama, and OpenAI architectures selected for this study, as these were not supported in the original repository. In order to test the models more robustly, we also add questions requiring reasoning as proposed by [Wang \(2025\)](#); [Li et al. \(2025b\)](#).

The evaluation protocol preserves the original framework's emphasis on positional effects, systematically testing whether information placement within documents affects retrieval accuracy. This aspect is particularly relevant for understanding the "lost in the middle" phenomenon documented in previous long context studies, where models demonstrate reduced performance for information located in central document sections.

### **3.3 Dataset Creation and Preparation**

The study utilized Finnish Wikipedia articles obtained from the Hugging Face datasets repository [Tanskanen \(2023\)](#) as the foundational "haystack" content. This dataset was selected based on several critical criteria, it provides authentic, naturally occurring Finnish text of substantial length suitable for long-context evaluation. Also the content spans diverse topics ensuring varied linguistic and contextual complexity. It also provides open availability ensuring reproducibility of the research, and Wikipedia's structured format provides consistent text quality across different topic domains.

The dataset creation process employed a streaming approach to efficiently handle the large Wikipedia corpus without requiring complete dataset downloads. Articles were processed sequentially, with very short articles (fewer than 100 characters) excluded to maintain content quality. Article separation was maintained through consistent formatting with double newline characters between articles, preserving document structure while creating continuous text suitable for long context testing.

**Systematic Token Based Dataset Construction** Five distinct context length conditions were used for this study. The dataset creation process utilized the tiktoken

library with the cl100k\_base encoding (GPT 4o tokenizer) [Goldbaum \(2025\)](#), ensuring consistency with current industry standards and enabling accurate comparison with other benchmarking studies. This tokenization approach provides reliable metrics for context length measurement across different model architectures, despite variations in internal tokenization schemes used by individual models.

The specific dataset sizes:

- finnish\_wikipedia\_100000\_tokens.txt: 80,184 tokens
- finnish\_wikipedia\_350000\_tokens.txt: 278,211 tokens
- finnish\_wikipedia\_500000\_tokens.txt: 397,789 tokens
- finnish\_wikipedia\_750000\_tokens.txt: 595,422 tokens
- finnish\_wikipedia\_1000000\_tokens.txt: 793,831 tokens

Each dataset was created by encoding the complete collected Wikipedia content, token counts were calculated with the tiktoken tokenizer library. Reproducibility was ensured through the use of a fixed random seed (seed=42) during the dataset creation process.

### **3.3.1 Novel Reasoning Question Development and Content Design**

A critical methodological contribution of this study is the development of custom reasoning questions that extend beyond simple factual retrieval to test synthesis and inference capabilities. Given the absence of Finnish specific reasoning evaluation tasks in existing long context benchmarks, this study systematically designed questions to evaluate different levels of cognitive processing within extended Finnish contexts by adapting reasoning as proposed by ([Wang, 2025](#)).

Ten distinct "needles" specific pieces of factual information requiring varying levels of cognitive processing were systematically designed for insertion into each Wikipedia article haystack. The needle content was created with specific focus on historical and quantitative information that aligns with the predominant content themes identified in the

Finnish Wikipedia dataset.

The following table represents the ten needles used in this study, questions, expected answers, and difficulty classifications. Questions 1 to 5 represent medium difficulty tasks requiring direct information extraction, while questions 6 to 10 represent hard difficulty tasks incorporating arithmetic reasoning, temporal calculation, percentage inference, and multi step logical deduction.

**Table 3:** Complete Set of Needle Questions and Expected Answers

Q#	Question	Expected Answer	Difficulty
1	Mikä on Jyväskylän vesipinta-alan osuus neliökilometreinä Hydrotech-raportin mukaan?	73,2	Medium
2	Kuka on Azerbaidžanin energiainisteri Rashid Aliyevin erityisneuvonantaja?	Reddick Miller	Medium
3	Mikä on Albanian ulkomaankaupan enimmäisosuus BKT:sta vuoden 2024 erityislain mukaan?	67,8	Medium
4	Kuka nimettiin Andorran kansallisarkiston pääarkistaariksi vuonna 2023?	Maria Gonzalez-Ferrandez	Medium
5	Minkä viljalajin edistämiseen Afganistanin maatalousministeriön uusi strategia keskittyy?	durra	Medium
6	Kuinka monta koetta ISS-aseamalla suoritettiin yhteensä vuoden 2024 raportin mukaan?	15	Hard
7	Minä vuonna Elena Koci aloitti Albanian museon johtajana?	2003	Hard
8	Mikä on fossiilisten polttoaineiden nykyinen osuus Wakandan sähkön-tuotannosta?	80	Hard
9	Kuinka monta Aamulan yliopiston opiskelijaa opiskelee germaansia kieliä?	18	Hard
10	Minkä nimistä menetelmää Afganistan käyttää harvinaisten maametallien etsintään?	keskisyvyysanalyysi	Hard

The complete needle texts containing the contextual information for each question are provided in Appendix.

The needles were structured into two distinct difficulty categories based on cognitive processing requirements:

**Medium difficulty needles (1–5; five items).** These primarily require direct extraction of explicitly stated facts (e.g., geographical, demographic, or basic historical details), testing information retrieval rather than multi step reasoning. In our no context baseline, the evaluated OpenAI and Gemini models achieved 100% accuracy on these five items, establishing a retrieval capability baseline. The evaluated Llama model was 100% correct on Q2–Q5 but not Q1. Baseline results are reported in Appendix B.

**Hard difficulty needles (6–10; five items).** These require specific reasoning beyond simple extraction (e.g., arithmetic, temporal calculation, or logical inference). Despite the added reasoning, the evaluated OpenAI and Gemini models also achieved 100% accuracy in the no context baseline. The evaluated Llama model had a wider accuracy range. See Appendix B for per question baseline results.

**Reasoning Task Types** The hard difficulty questions were designed to test specific types of reasoning within Finnish language contexts. Arithmetic reasoning questions required basic mathematical operations on information distributed across text, such as summing the number of experiments conducted across three categories on the ISS space station ( $9 + 4 + 2 = 15$ ). Temporal calculation questions tested backward calculation from given dates and duration information, requiring models to determine when a museum director began their tenure by subtracting successive appointment durations from a known date ( $2023 - 8 - 12 = 2003$ ). Percentage inference questions demanded logical deduction from partial percentage information, such as inferring that if 20% of energy production comes from hydropower, the remaining 80% must come from fossil fuels. Multi step calculation questions required sequential percentage calculations, such as determining how many students study Germanic languages when 15% of 240 students study "other languages" and half of those focus on Germanic languages ( $15\% \text{ of } 240 = 36$ ;  $36 \div 2 = 18$ ). Finally, reading comprehension with inference questions required identification of implicit relationships between presented information, such as determining which

mineral exploration method applies to rare earth metals based on effectiveness data and application context.

Each question underwent systematic validation to ensure methodological rigor. Content validation involved verifying that no needle information appeared in the original Finnish Wikipedia dataset, preventing false positives in which a model might retrieve pre existing rather than intentionally inserted information.

### **3.4 Needle Placement and Distribution Strategy**

The needle insertion process follows a systematic placement strategy designed to evaluate positional effects comprehensively. For each test depth percentage (10%, 25%, 50%, 75%, 90%), the ten needles are distributed across a range surrounding the target depth to avoid clustering effects that might artificially enhance or diminish retrieval performance.

The distribution algorithm spreads needles within  $\pm 15\%$  of the target depth percentage. This approach ensures that needles are positioned across a 30% range centered on the target depth, with bounds constrained to remain between 5% and 95% of document length.

The selection of test depths (10%, 25%, 50%, 75%, 90%) provides comprehensive coverage of document positions, enabling detailed analysis of the "lost in the middle" phenomenon. These specific percentages were chosen to capture performance at document beginnings (10%), early sections (25%), middle regions (50%), later sections (75%), and near document ends (90%), providing sufficient granularity to identify positional performance patterns while maintaining manageable experimental complexity.

### **3.5 Model Selection and Configuration**

Three state of the art LLM architectures were selected for evaluation, representing the leading commercial providers: Gemini 2.0 Flash (Google), Llama 4 Maverick (Meta), and GPT 4.1 Mini (OpenAI). This selection was strategically chosen for

several methodological reasons. First, these models represent the most accessible and widely deployed long context solutions available to enterprises, ensuring that findings directly inform practical deployment decisions while remaining cost efficient. Second, all selected models demonstrate documented competency in multilingual tasks, with specific support for Finnish language processing confirmed through preliminary testing. Finally, each model supports the extended context windows (up to 1M tokens) required for comprehensive long context evaluation across the full range of experimental conditions.

Model inference was configured with specific parameters to ensure consistent and optimal performance across all tests. Temperature was set to 0 to minimize response variability and ensure deterministic outputs suitable for precise accuracy measurement. Maximum output tokens were limited to 300, sufficient for the expected short factual answers while preventing unnecessarily verbose responses that might complicate evaluation.

The evaluation employed task specific prompts written in Finnish to ensure appropriate language model engagement. The prompt structure incorporated clear instructions for careful text reading, specific question answering, and emphasis on textual accuracy. The prompt explicitly instructed models to leave answers blank if information could not be located, reducing false positive responses and enabling more reliable accuracy measurement.

### **3.6 Experimental Protocol and Data Collection**

Each experimental condition was executed through a systematic protocol designed to ensure consistency and reproducibility. For each combination of dataset size and depth percentage, three independent test runs were conducted to account for potential variability in model responses and API related fluctuations. This replication strategy provides statistical robustness while balancing resource constraints imposed by API costs.

All models were accessed through the OpenRouter API service, which provides standardized access to multiple LLM providers through a unified interface. Rate limiting considerations required implementation of 35 second delays between consecutive tests,

determined empirically to avoid API rate limit errors while maintaining reasonable experimental completion times.

Performance evaluation employed strict accuracy based scoring using substring matching methodology. For each needle question, any response containing the expected answer within the model's output was considered correct, accommodating minor variations in response formatting while maintaining accuracy standards.

The scoring methodology awards binary points (correct/incorrect) for each needle, with final accuracy calculated as the percentage of correctly answered questions across all needles in each test condition. No partial credit is awarded, ensuring clear performance differentiation and avoiding subjective evaluation complications.

The model responses underwent systematic validation by the researcher to ensure scoring accuracy and identify potential edge cases in automatic evaluation. This validation process involved reviewing each model response against the expected answer, considering context appropriateness and factual accuracy beyond simple string matching.

The validation process identified and addressed cases where models provided contextually correct but differently formatted answers, ensuring that evaluation criteria remained consistent across all test conditions while maintaining fairness in accuracy assessment. This human validation component adds reliability to the automated scoring process and provides insights into model response patterns that pure algorithmic evaluation might miss. Each model was evaluated across 3 runs of 25 test configurations (5 context lengths × 5 document depths × 3 replications), yielding 225 total observations across all three models.

All code, datasets, and evaluation protocols will be made publicly available upon study completion to enable replication and extension of this research. Dataset creation employed fixed random seeds (seed=42) to ensure reproducible content selection and needle placement across different research contexts. The complete experimental pipeline, including data processing scripts and evaluation frameworks, will be documented and shared to facilitate future research in Finnish language long context evaluation.

## **3.7 Expert Interview Methodology**

To link our findings with what is seen in the field, this study incorporated semi structured expert interviews with industry practitioners. The interview methodology was designed to gather professional perspectives on the benchmark results and their implications for real world deployment decisions.

### **3.7.1 Interview Design and Procedure**

Two professional IT consultants with direct experience implementing LLM solutions across different industries were recruited for this study. The selection criteria required a minimum of two years of professional experience with LLM implementation projects and active involvement in enterprise level AI deployment decisions. Participants were contacted through professional networks and agreed to participate on a voluntary basis.

The participants brought healthcare and legal sector experience. Each expert had experience with deploying LLM's solutions, and direct exposure to projects involving the integration of large language models for knowledge retrieval, document analysis, or workflow automation.

The interviews followed a semi structured format consisting of three main stages. Each interview began with a presentation of results, during which participants were provided with a concise summary of the benchmark outcomes, including performance differences between models, context length sensitivity patterns, and observed accuracy trends. Visual materials such as heatmaps and comparative plots were shared to facilitate interpretation. Following this presentation, an exploratory discussion invited participants to comment on the presented results, reflect on potential causes for the observed differences, and relate the findings to their own professional experience. Discussion topics included model suitability for long context retrieval tasks, implications for domain specific applications, and expectations for future model improvements. The final segment of each interview encouraged participants to evaluate the practical implications of the findings in real world contexts, such as healthcare documentation, legal text analysis, or enterprise data

management. This reflection and synthesis stage also explored how Finnish language performance might influence adoption and reliability in local or multilingual settings.

Each interview lasted approximately 45 minutes and was conducted in an online format to accommodate the participants' professional schedules. Conversations were recorded with consent and subsequently transcribed for qualitative content analysis. The semi structured format allowed for a consistent core of questions across interviews while providing flexibility to pursue relevant insights unique to each expert's domain. Participants provided informed consent for audio recording to ensure accurate capture of responses. The interviews were conducted in English to maintain consistency across participants. All participants were assured of anonymity in reporting. Professional roles and industry sectors are described in general terms to protect participant confidentiality while maintaining analytical relevance.

## 4 Results

This chapter presents the findings from the comprehensive evaluation of three large language models Gemini 2.0 Flash, Llama 4 Maverick, and GPT 4.1 Mini using the adapted Multiple Needles in a Haystack (MNIAH) benchmark with Finnish Wikipedia content. The results provide insights into long context information retrieval capabilities, positional effects, task complexity impacts, and Finnish language processing performance across different context window sizes and document positions.

### 4.1 Overall Performance Analysis

The overall performance analysis reveals interesting differences in long context information retrieval capabilities among the evaluated models.

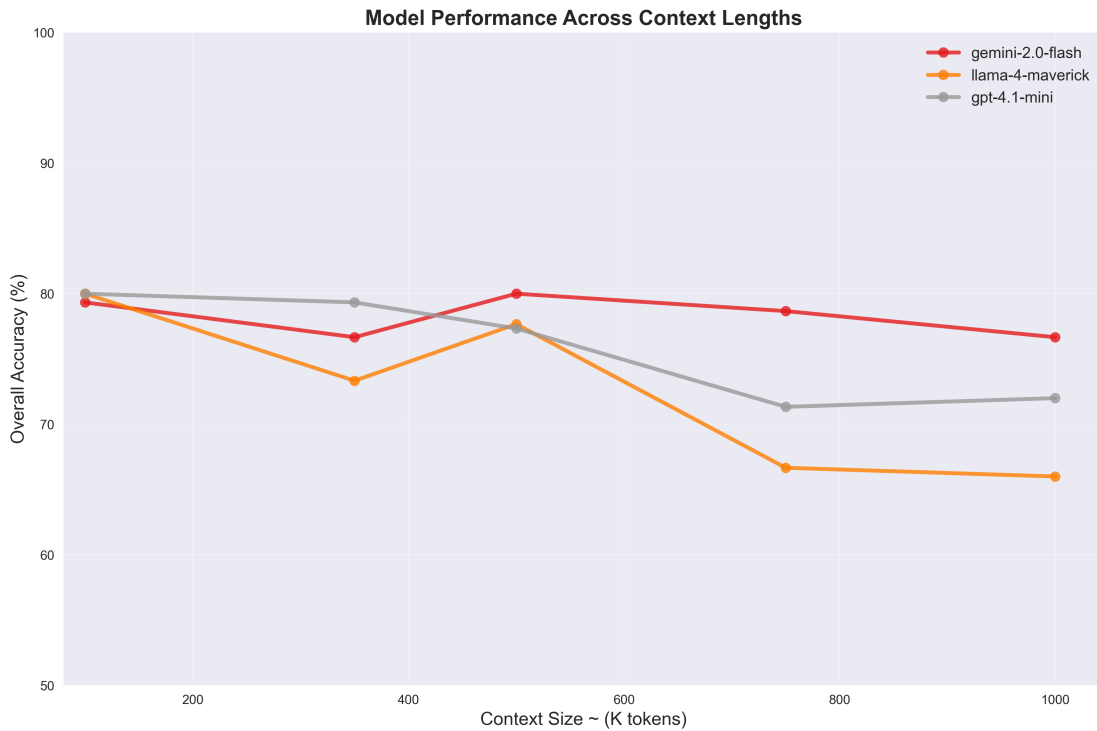
Overall accuracy is calculated as:

$$\text{Accuracy}_{\text{overall}} = \frac{\sum_{r=1}^R \sum_{i=1}^{N_r} \mathbf{1}[\hat{y}_{r,i} = y_{r,i}]}{\sum_{r=1}^R N_r}.$$

In each experimental run, denoted by  $r$ , a total of  $N_r$  test items or "needles" are included. Each experimental configuration is replicated  $R$  times to ensure robustness and reproducibility of the results. For every question  $i$  within a given run  $r$ , the true label, or correct answer, is represented by  $y_{r,i}$ . Correspondingly, the model's predicted answer for the same question is denoted by  $\hat{y}_{r,i}$ .

The three models demonstrated competent long context processing capabilities, all exceeding 70% overall accuracy. Gemini 2.0 Flash achieved the highest performance at 78.3%, followed by GPT 4.1 Mini at 76.0%, and Llama 4 Maverick at 72.7%. The performance differential of 5.6 percentage points between the highest and lowest performing models indicates measurable variation. Notably, all three models demonstrated competent performance above 70%.

Visually analysing the figure there seems to be a noticeable trend in accuracy decline as



**Figure 1:** Overall model accuracy by context size

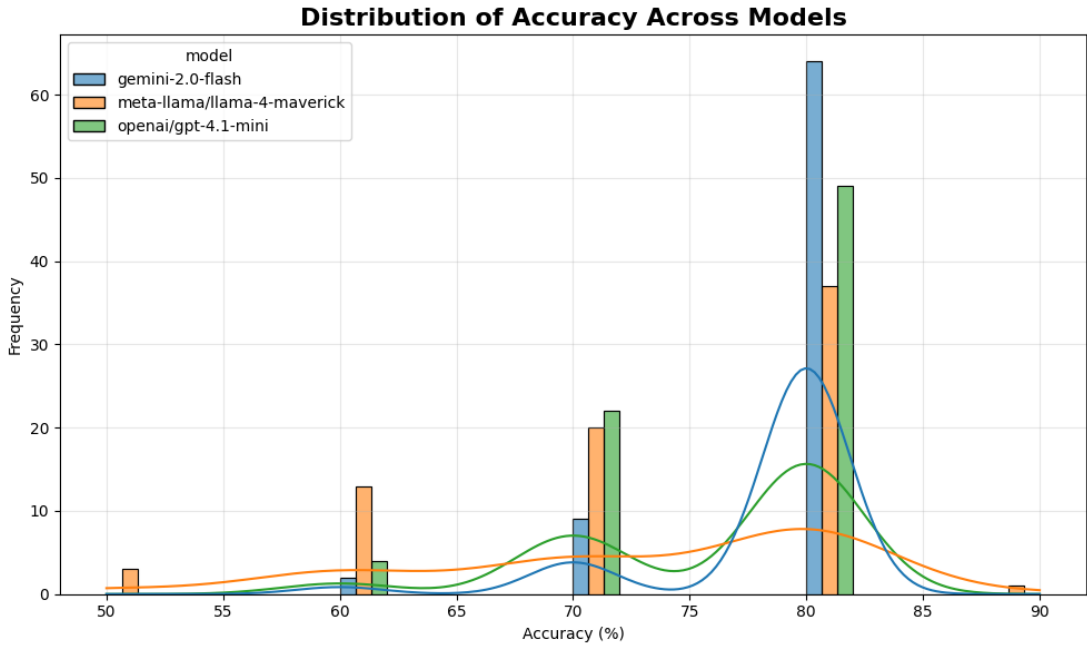
context grows.

#### 4.1.1 Accuracy Distribution and Variance Across Models

To evaluate model stability across repeated runs, Figure 2 visualizes the distribution of overall accuracy for each model. Each bar represents the frequency of accuracy values obtained across all 225 benchmark runs, while the overlaid kernel density curves illustrate the approximate shape of each model’s performance distribution.

The results reveal clear differences in both central tendency and variance between models. Gemini 2.0 flash displays a narrow, sharply peaked distribution centered around 80% accuracy, indicating highly consistent behavior across varying context lengths and needle depths.

GPT 4.1 mini exhibits a slightly wider distribution, with most runs falling between 70% and 85%. While the model performs comparably to Gemini on average, its broader spread suggests that its accuracy fluctuates more across context variations. This



**Figure 2:** Distribution of accuracy between models

moderate variance indicates some sensitivity to depth and contextual length, though overall reliability remains high.

By contrast, LLaMA 4 Maverick shows a distinctly wider distribution, with accuracy modes near 60% and 70%, and an isolated outlier at 90%. The greater spread implies substantial variability between runs, potentially due to weaker capabilities in retrieving context. The single 90% run is an interesting statistical outlier, since gemini or gpt couldn't reach that accuracy on any run.

**Table 4:** Overall Sample Standard Deviation of Models

Model	Overall Std
gemini 2.0 flash	4.46
openai/gpt 4.1 mini	5.93
meta Llama 4 maverick	9.11

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (1)$$

Here,  $s$  denotes the sample standard deviation, and  $N$  represents the total number of observations. Each individual observation is denoted by  $x_i$ , while  $\bar{x}$  corresponds to the sample mean.

Overall, Gemini demonstrates the lowest standard deviation and most stable accuracy across conditions, GPT 4.1 mini maintains moderate stability, and Llama exhibits significant deviation that raises questions about its robustness.

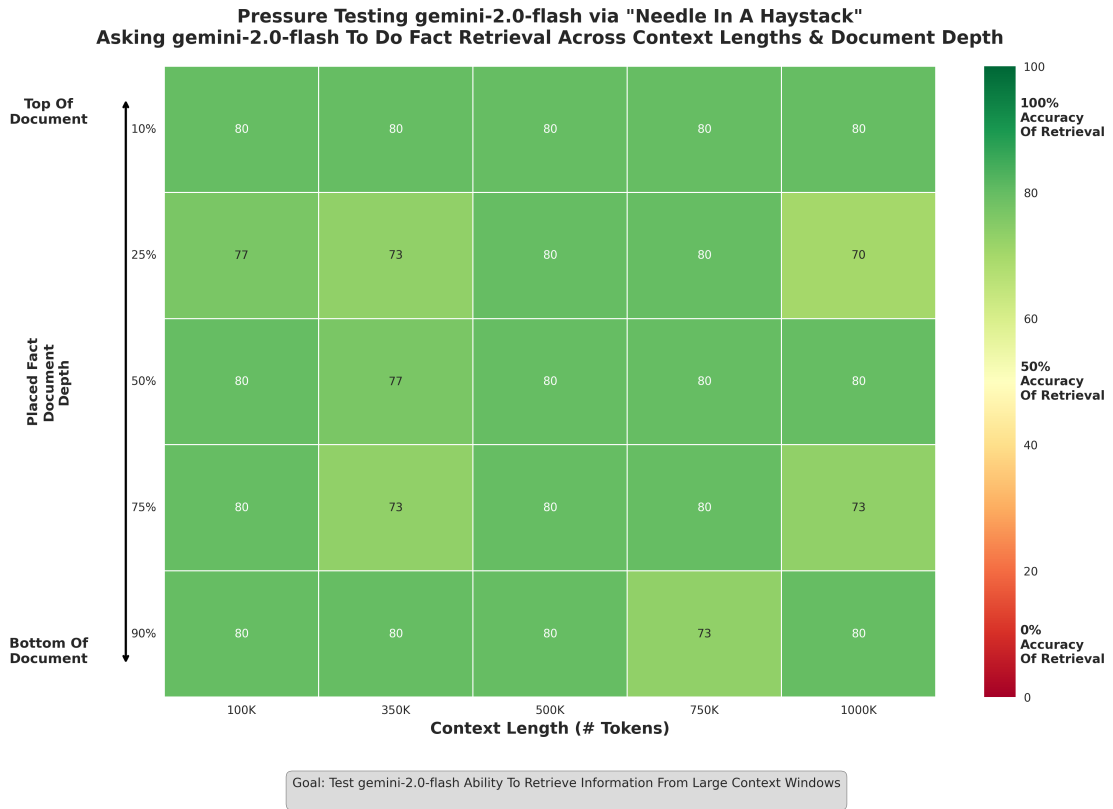
## 4.2 Context Length Impact on Accuracy

Analysis of performance across different context window sizes (100K to 1000K tokens) reveals varying degrees of context length sensitivity among the models. Figures 3 to 5 demonstrates that while all models maintained relatively stable performance across context lengths, subtle degradation patterns emerged at extended context windows.

$$\text{Accuracy}_{\text{context}=c} = \frac{1}{N_c} \sum_{i=1}^{N_c} \text{Accuracy}_{i,c} \quad (2)$$

where  $c$  denotes the context length (100K, 200K, ..., 1000K tokens),  $N_c$  represents the total number of observations at context length  $c$  (across all depth positions and runs), and  $\text{Accuracy}_{i,c}$  is the accuracy for the  $i$ -th observation at context length  $c$ .

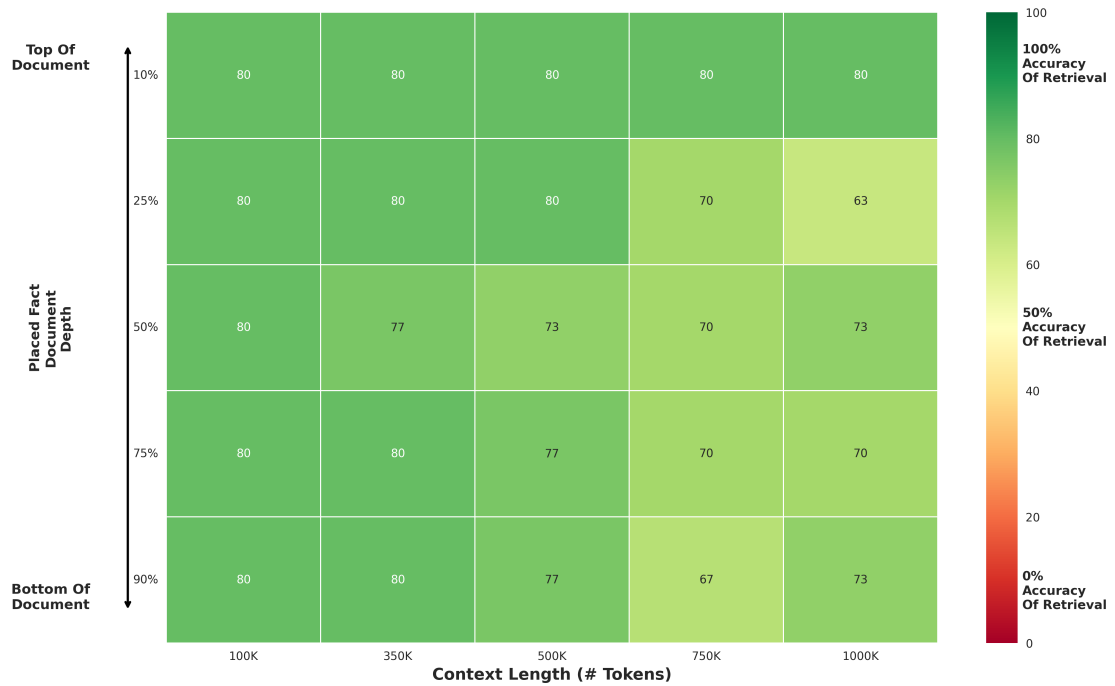
**Gemini 2.0 Flash** maintained the most consistent performance across all context lengths, with minimal degradation from 79.3% at 100K tokens to 76.7% at 1000K tokens (2.6 percentage point decline)



**Figure 3:** Gemini 2.0 Flash accuracy by depth and context size

**GPT-4.1 Mini** showed moderate context length sensitivity, declining from 80.0% at 100K tokens to 71.3% at 1000K tokens (8.7 percentage point decline)

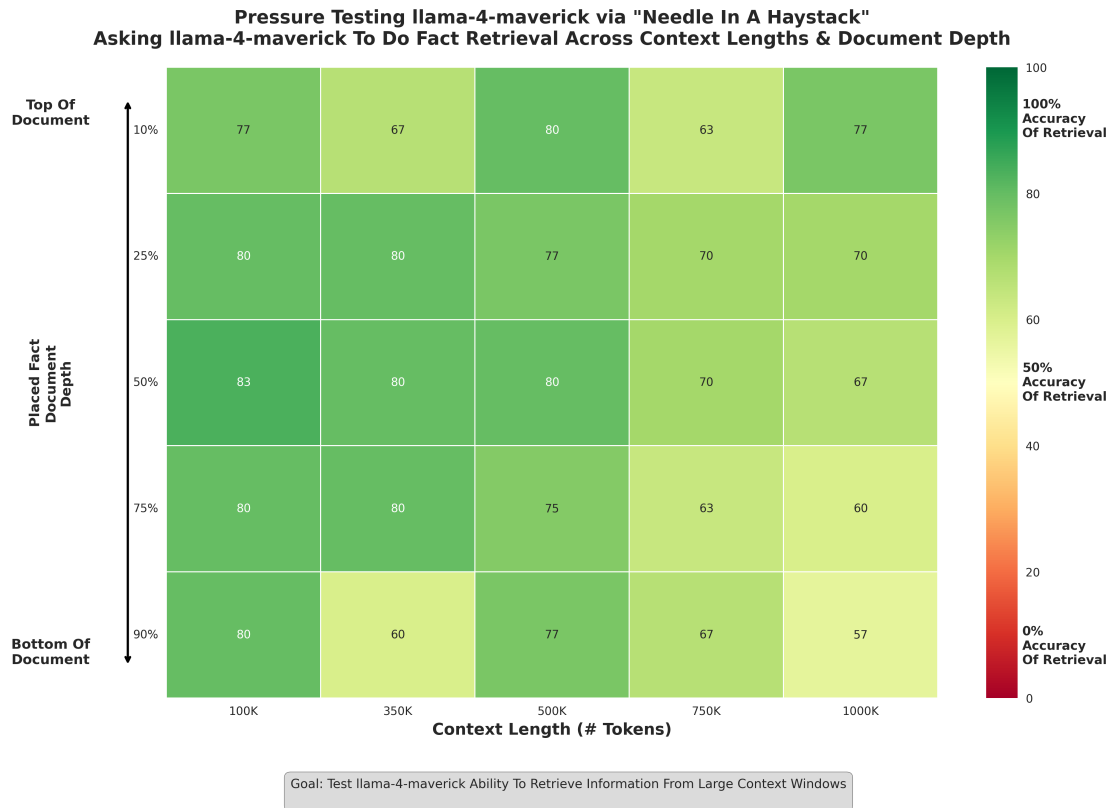
**Pressure Testing gpt-4.1-mini via "Needle In A Haystack"**  
**Asking gpt-4.1-mini To Do Fact Retrieval Across Context Lengths & Document Depth**



Goal: Test gpt-4.1-mini Ability To Retrieve Information From Large Context Windows

**Figure 4:** Gpt 4.1-Mini accuracy by depth and context size

**Llama 4 Maverick** exhibited the most pronounced context length sensitivity, dropping from 72.7% at shorter contexts to 68.0% at 1000K tokens



**Figure 5:** Llama-4-Maverick accuracy by depth and context size

These findings suggest that while extended context capabilities are present in all tested models, their robustness varies significantly, with Gemini 2.0 Flash demonstrating superior context window utilization consistency.

### 4.3 Positional Effects

The evaluation systematically tested information retrieval across five different document positions (10%, 25%, 50%, 75%, and 90% depth) to identify potential positional biases in long context processing.

Depth	10%	25%	50%	75%	90%
Gemini 2.0 Flash	80.0%	76.0%	79.3%	77.3%	78.7%
GPT 4.1 Mini	80.0%	74.7%	74.7%	75.3%	75.3%
Llama 4 Maverick	72.7%	75.3%	76.0%	71.4%	68.0%

**Table 5:** Overall accuracy by document depth

$$\text{Accuracy}_{\text{depth}} = \frac{\text{Correct answers at depth } d}{\text{Total answers at depth } d} \quad (3)$$

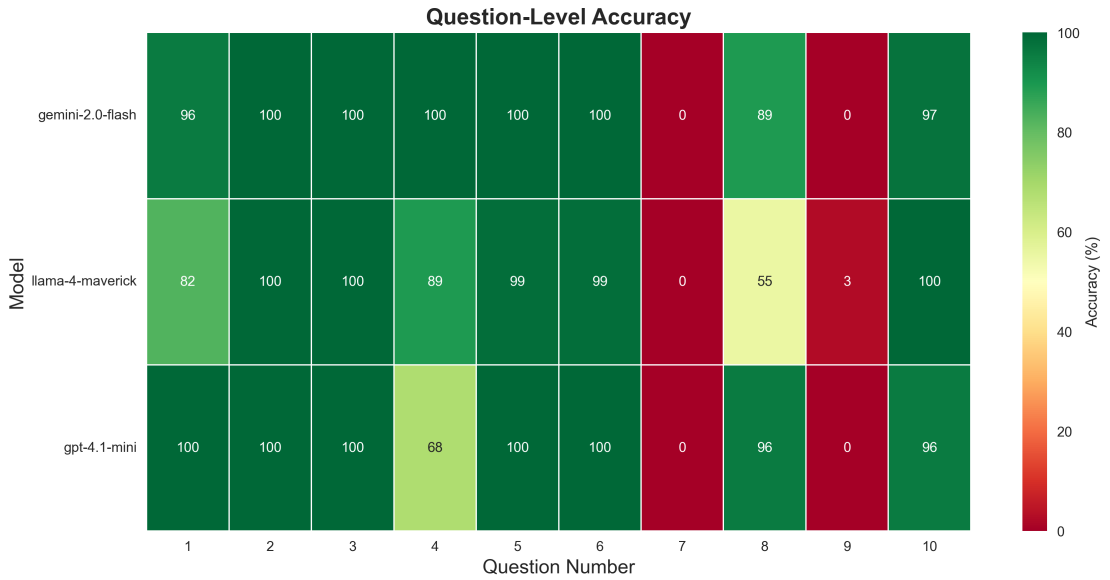
The positional analysis reveals nuanced patterns that vary across model architectures. Gemini and GPT 4.1 Mini demonstrate some advantage when target information appears at the beginning of documents, suggesting that these models may prioritize or more effectively process early context information. However, the Llama 4 Maverick model breaks this pattern, showing no consistent beginning position advantage and indicating that positional biases may be architecture specific rather than universal across long context models. Contrary to the "lost in the middle" phenomenon documented in earlier long context research, the middle position performance does not show systematic degradation by analysing the means visually. Models such as Gemini and Llama maintain strong accuracy when needles are placed at the midpoint of documents, and visual analysis of mean accuracies across depth percentages does not reveal the expected accuracy drop for middle positioned information. While these observational patterns suggest that middle position challenges may be less pronounced than anticipated, regression analysis

was conducted to test whether statistically significant positional effects emerge when controlling for other variables.

#### 4.4 Differences between tasks

The benchmark evaluation distinguished between medium difficulty and hard difficulty questions, providing insights into the models' capabilities when handling varying task complexity levels within long contexts.

##### Difficulty Based Performance Analysis:



**Figure 6:** Accuracy by question and model

$$\text{Accuracy}_{m,q} = \frac{C_{m,q}}{N_{m,q}} \times 100\%, \quad C_{m,q} = \sum_{i=1}^{N_{m,q}} \mathbf{1}[\hat{y}_{m,q,i} = y_q] \quad (4)$$

In this formulation,  $m$  denotes the model and  $q$  denotes the question under consideration. The term  $N_{m,q}$  represents the total number of answers produced by model  $m$  for question  $q$ . Each individual predicted answer, corresponding to the  $i$ th response from model  $m$  to question  $q$ , is denoted by  $\hat{y}_{m,q,i}$ . The correct, or gold, answer for question  $q$  is indicated by  $y_q$ . Finally,  $\mathbf{1}[\cdot]$  denotes the indicator function, which takes the value 1 if the specified

condition is true and 0 otherwise.

#### **4.4.1 Question Type Performance Distribution**

Analysis of individual question performance revealed distinct clusters that show which task characteristics correlate with success or failure. Questions 2, 3, 5, 6, and 10 achieved greater than 95% accuracy across all models, demonstrating that the models could reliably handle both straightforward factual retrieval (questions 2, 3, and 5) and certain forms of simple reasoning (questions 6 and 10) when information was clearly structured and accessible.

Questions 1, 4, and 8 exhibited moderate success rates between 80% and 95%, representing an intermediate complexity level where modest reasoning or synthesis across multiple information points was required. Performance variations among models were most pronounced in this category, with individual model strengths and weaknesses becoming more apparent.

The most notable pattern emerged from questions 7 and 9, which consistently yielded accuracy rates below 50% across all tested models. These questions required more complex reasoning operations beyond simple information retrieval, and their universal failure rates suggest systematic rather than model specific limitations. The consistent poor performance across different model architectures, training approaches, and parameter scales indicates that current large language models face fundamental challenges when combining long context processing with multi step reasoning in Finnish.

This finding has important practical implications for enterprise deployments, as it suggests that applications requiring reasoning within long Finnish documents may require additional validation mechanisms, hybrid approaches combining retrieval augmented generation with human oversight for complex reasoning tasks.

## 4.5 Mixed Effects Regression Analysis

To quantitatively assess how *context length* and *document depth* influence retrieval accuracy in the Needle in a Haystack benchmark, a linear mixed effects model was fitted. This approach captures both the general performance trends across all models and the model specific baseline differences through random intercepts.

Because multiple accuracy measurements originate from the same language model architecture, a simple linear regression would violate the assumption of independent residuals. To address this, a hierarchical (mixed effects) framework was employed to account for the groupwise dependence among observations.

### 4.5.1 Model Specification

The model was specified as follows:

$$\text{Accuracy}_{ij} = \beta_0 + \beta_1 \cdot \text{ContextLength}_{ij} + \beta_2 \cdot \text{DepthPercent}_{ij} + u_{0j} + \varepsilon_{ij} \quad (5)$$

In this model specification,  $i$  indexes an individual test configuration, defined by a specific combination of context length and document depth, while  $j$  indexes the model type, where  $j \in \{\text{Gemini 2.0 Flash, GPT-4.1 Mini, Llama 4 Maverick}\}$ . The parameter  $\beta_0$  represents the global intercept, corresponding to the baseline retrieval accuracy. The coefficients  $\beta_1$  and  $\beta_2$  capture the fixed effects associated with context length and document depth, respectively. The term  $u_{0j}$  denotes the random intercept for model  $j$ , assumed to follow a normal distribution  $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$ . Finally,  $\varepsilon_{ij}$  represents the residual error term, distributed as  $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ . DepthPercent represents the position of the target information within the document, expressed as a percentage from the beginning 0% to the end 100%.

This specification allows each model to have its own baseline performance level while estimating shared slopes for the contextual effects.

The model was implemented in Python using the `statsmodels` package's `MixedLM` class.

Each observation represents one aggregated test result (average retrieval accuracy across 10 questions) for a particular model, context size, document depth, and run. The model was estimated using Restricted Maximum Likelihood REML. REML was selected because it produces unbiased estimates of variance components. This is particularly important given our limited number of models ( $n = 3$ ), ensuring accurate characterization of between-model variability through the random intercept term.

#### 4.5.2 Regression Results

**Table 6:** Mixed-effects model

Term	coef	std err	t	p-value
Intercept	82.6097120812	1.8815362574	43.9054585090	0.0000000000
dataset_size	-0.0000096005	0.0000012800	-7.5001476802	0.0000000000
depth_percent	-0.0354085316	0.0134116123	-2.6401398084	0.0082871832

Both fixed effects are statistically significant and negative, indicating that increasing context length and placing information deeper in the document are associated with reduced retrieval accuracy.

The context-length coefficient ( $\beta_1 = -9.6 \times 10^{-6}$ ) indicates that each additional token in the context window is associated with a 0.0000096 percentage point decrease in accuracy. Across the tested range, this translates to a cumulative accuracy decline of 8-10 percentage points. Tokenization varies across models due to different tokenizer implementations.

These findings quantitatively suggest that retrieval accuracy declines modestly but significantly as the context window expands and as relevant information appears deeper within a document. The magnitude of the context length coefficient indicates that long context degradation is measurable but moderate, whereas the depth effect, though smaller, also remains statistically significant.

The document depth coefficient ( $\beta_2 = -0.035$ ) indicates that accuracy decreases by approximately 2-3 percentage points from the beginning to the end of the tested depth range (10% to 90% of document length), demonstrating a statistically significant positional

bias in information retrieval.

## 4.6 How do the results answer our research questions?

### **RQ1: How does context length affect information extraction accuracy when processing Finnish language documents with large language models?**

The results indicate a decline in accuracy as the context length increases, suggesting that extraction performance is influenced by the size of the context window. Across all three models, longer context windows led to measurable, though varying, decreases in mean accuracy

On harder questions requiring reasoning such as questions 7 and 9, even small amounts of context instantly dropped accuracy in solving them to near 0, even though without context the Gemini and OpenAi models solve all 10 questions with perfect accuracy, lama scores well without context as well. These baseline scores are found at the appendix.

The regression analysis suggests a statistically significant negative relationship between context length and retrieval accuracy. Visual analysis of the means alone suggest performance variation, but the mixed-effects model clearly suggests that **longer contexts negatively affect extraction performance** across architectures. Thus, answer to **RQ1 is**: increasing context length leads to measurable declines in retrieval accuracy, with the degree of accuracy drop varying between models.

### **RQ2: Does the positioning of the needles affect extraction accuracy?**

All three models demonstrated different patterns of accuracy depending on the depth our questions were inserted at.

While mean based comparisons alone suggested mixed and model specific outcomes, the regression analysis suggests that position does influence retrieval accuracy albeit weakly. The depth coefficient's significance establishes that positional effects are measurable rather than incidental, revealing a subtle but reliable downward trend as information appears deeper in long contexts. Thus, **RQ2 is supported at a statistically significant**

level, though the practical impact remains moderate and heavily dependent on the model's attention and positional encoding mechanisms.

#### **4.6.1 Summary and Hypotheses**

The empirical findings provide support for both research hypotheses formulated at the outset of this study. Regarding the first research question, the data suggests that extraction accuracy decreases as context length increases, demonstrating a clear inverse relationship between document size and model performance when processing Finnish language materials. This degradation was observed across all tested models, though the magnitude and rate of decline varied depending on task complexity and specific model architecture.

For the second research question, the analysis revealed that the position of target information within documents exerts a statistically significant, albeit modest, influence on extraction accuracy. While visual inspection initially suggested minimal positional effects, the regression analysis uncovered meaningful patterns indicating that document depth specifically, information located deeper within longer contexts negatively affects retrieval and reasoning accuracy.

From an applied perspective, these findings suggest that while current large language models demonstrate impressive long context capabilities, they still exhibit structural limitations in attention distribution and memory retention. In practical deployments, structuring documents to place critical information near the start of context windows, or segmenting long documents into shorter chunks, may improve retrieval accuracy and consistency particularly in multilingual contexts.

### **4.7 Expert Interview Findings**

The expert interviews provided industry perspective on whether the benchmark results indicate readiness for mission critical enterprise applications. Two IT consultants with years of LLM implementation experience in legal and healthcare sectors assessed the performance data and provided insights on enterprise deployment readiness.

All consultants expressed that the overall performance results were aligned with or exceeded their expectations for long context information retrieval tasks, though with different emphases based on their sector experience.

**Consultant A (Legal Technology):** "I was really surprised at the accuracy rates around extracting simple pieces of information, the models are more accurate using Finnish than I expected.

**Consultant B (Healthcare Implementation):** "The results were broadly in line with my expectations for extraction tasks, but I have to say Gemini's performance was genuinely surprising, I thought OpenAi would provide best results. "

#### 4.7.1 Task types

Both consultants agreed that the benchmark results indicate readiness for specific types of mission critical applications where tasks are primarily extraction focused and human verification remains part of the workflow.

**Consultant A:** "In legal context's i feel that around 75 percent of the LLM use cases are extraction use cases, but for use cases requiring reasoning, the pipelines need additional steps. And in extraction use cases where accuracy is important, i still would include a human in the loop even with near 100 percent accuracy in your benchmark.

**Consultant B:** "Healthcare use cases are predominantly extraction heavy rather than reasoning heavy. Patient record summarization, and many others these are all tasks where the benchmark results indicate readiness, provided we maintain human verification in the workflow, there are also additional methods you can use to improve accuracy results"

Despite positive assessments for certain applications, all consultants agreed that for tasks containing multiple steps of reasoning and long context multiple additional steps are still needed to improve accuracy before enterprise adoption.

**Consultant A:** "The hard questions in your benchmark the ones with 0% success rates those represent the kind of complex reasoning we encounter in legal document analysis,

legal documents often also contain industry specific words which the LLM's don't handle as accurately."

**Consultant B:** "In healthcare, the sector simply cannot afford mistakes, even rare ones. Even if we had a model performing at 99% accuracy for medication extraction or dosage information, we would still require human in the loop validation. The risk profile is just too high. The benchmark results are good, but they confirm we're not at autonomous deployment for critical tasks, for reasoning tasks I would look at implementing judges to make sure the answers are grounded on the original text, and possible RAG systems to control context in order to increase accuracy.

#### **4.7.2 Limitations of The Models**

The consultants identified several specific limitations that emerged from both the benchmark findings and their real world deployment experience, which collectively suggest caution for enterprise adoption. The near zero success rates observed on questions 7 and 9 revealed fundamental weaknesses in complex reasoning capabilities, particularly as context length increased. These failures were not isolated incidents but systematic breakdowns that occurred consistently across all tested models, indicating that current architectures struggle to maintain logical coherence and multi step reasoning when processing longer Finnish documents.

Performance degradation at extended context lengths emerged as a recurring concern, especially for tasks requiring reasoning beyond simple information extraction. While the models demonstrated robust capabilities for retrieval tasks across varying context sizes, the addition of even modest amounts of contextual material substantially compromised their ability to perform arithmetic computations or logical deductions. This sensitivity to context length presents practical challenges for enterprise applications that routinely process lengthy documents such as legal contracts, financial reports, or technical specifications.

The consultants noted interesting patterns regarding positional effects in the data. Visual inspection of the results suggested that information placement within documents had

limited impact on accuracy, leading to an initial impression that positional bias might not be a significant factor. However, the formal regression analysis revealed statistically significant positional effects that were not immediately apparent from exploratory visualization alone. This discrepancy highlighted the importance of rigorous statistical testing to uncover subtle but meaningful patterns that could affect deployment decisions.

Finally, the consultants emphasized that all observed performance levels presuppose clean, well structured input data. Real world enterprise documents frequently contain formatting inconsistencies, scanning artifacts, multilingual fragments, and structural irregularities that could further degrade model performance beyond what the controlled benchmark environment captured. This dependency on data quality necessitates additional preprocessing pipelines and quality assurance mechanisms in production deployments.

**Consultant B:** "Your benchmark uses clean Wikipedia data, which is valuable for controlled testing. But the largest barrier to enterprise adoption isn't model capability it's source data quality. In real healthcare implementations, we're dealing with inconsistently formatted medical records, handwritten notes that have been poorly digitized, and incomplete documentation."

**Consultant B:** "Based on these results, the clear recommendation for healthcare implementations is to keep context windows as small as possible. You gain both accuracy and cost benefits with smaller context sizes."

## 5 Summary and Discussion

### 5.1 Overview

This thesis examined the reliability of large language models (LLMs) for long context information extraction in Finnish, addressing a critical evidence gap that hinders enterprise adoption. The study designed, implemented, and evaluated a Finnish adaptation of the Multiple Needles in a Haystack benchmark using Finnish Wikipedia as the haystack corpus and a set of ten Finnish needles that spanned direct retrieval and reasoning tasks. Three contemporary LLMs—Gemini 2.0 Flash, Llama 4 Maverick, and GPT-4.1 Mini—were evaluated across five context lengths (up to  $\approx 1\text{M}$  tokens), five depth positions, and replicated runs to characterize accuracy, variance, positional effects, and performance depending on task complexity. Semi structured expert interviews complemented the quantitative results to analyse the results and assess deployment readiness in enterprise scenarios.

The primary objective was to validate LLM performance in long context's using the Finnish language and provide evidence for Finnish enterprises considering LLM deployment. Secondary objectives were quantifying how accuracy varies with context length across model families and identifying positional effects using Finnish language.

### 5.2 Key Empirical Findings

Across 225 test cases, models achieved overall accuracies between 72.7% and 78.3%: Gemini 2.0 Flash (78.3%), GPT-4.1 Mini (76.0%), and Llama 4 Maverick (72.7%). While all three surpassed 70% overall, performance differences of up to 5.6 percentage points indicate meaningful architectural and/or training differences in long context utilization.

All models were strong on medium difficulty, direct retrieval needles, but accuracy dropped sharply on hard needles that required arithmetic, temporal, or multi step reasoning. The findings demonstrate that simple information extraction capabilities of these models are strong in Finnish, but as reasoning is introduced capabilities drop significantly. Two needles

(Q7 *albania\_museum\_director\_start\_year* and Q9 *aamula\_germanic\_language\_students*) produced near universal failures (0.0% and 0.9%).

Both visual analysis of means and regression analysis confirmed a clear negative effect of increasing context length affecting extraction accuracy.

While mean based results appeared inconsistent, regression analysis also confirmed a small but statistically significant decline in accuracy with increasing depth, indicating mild positional effects on accuracy as needles are inserted deeper into the document.

### **5.3 Answers to the Research Questions**

#### **RQ1: How does context length affect extraction accuracy in Finnish?**

Context length has a statistically significant negative effect on extraction accuracy, though the magnitude varies by model. The mixed-effects regression analysis reveals that accuracy degrades by approximately 8-10 percentage points across the tested range, accounting for tokenization variability between models.

However, this effect is not uniform across models. Gemini 2.0 Flash demonstrated remarkable resilience, maintaining near flat accuracy even at longer contexts. In contrast, GPT-4.1 Mini and Llama 4 Maverick showed more pronounced degradation as context length increased. Despite this degradation, all models maintained accuracy above 70% across the tested range, suggesting that modern long context models remain viable for Finnish document processing, though smaller context windows may offer better accuracy cost latency trade offs for tasks that permit document segmentation. This is in line with our H1 hypothesis, the study forecasted that extraction accuracy will drop with increased context length.

#### **RQ2: How does target position within long Finnish documents affect extraction?**

Document position significantly affects extraction accuracy. The mixed-effects regression analysis reveals a statistically significant depth effect ( $\beta_2 = -0.035$ ,  $p < 0.01$ ), with accuracy declining by approximately 2-3 percentage points from early positions (10%) to

later positions (90%) within documents. This positional bias indicates that models retrieve information less reliably when it appears deeper in long contexts. The effect is modest but consistent across models, suggesting that strategic placement of critical information toward the beginning of prompts or documents may improve extraction reliability in production systems. The results also support our H2 hypothesis, that extraction accuracy will drop with increased depth.

## 5.4 Discussion

### 5.4.1 Are Current LLMs Reliable Enough For Information Extraction In Finnish Companies?

Interviewed practitioners in legal and healthcare agreed: for *extraction centric* mission critical workflows, the observed accuracy is deployable *with* human in the loop verification, guardrails and additional techniques to boost accuracy. For *reasoning centric* tasks requiring long context, current models are not yet ready use in use cases requiring high accuracy without multiple processes supporting the LLM's such as additional judges that make sure the answer is grounded in truth, and RAG systems to reduce context.

## 5.5 Scientific and Practical Contributions

The study provides a validated, replicable Finnish long context benchmark that incorporates both retrieval and reasoning needles, systematically varies position and context length, and reports replicated, model comparative results suitable for tracking progress over time. This methodological contribution addresses a gap in existing evaluation frameworks, which have predominantly focused on English language tasks and simple retrieval scenarios without examining how task complexity and positional factors interact across extended contexts.

The empirical findings provide insights about long context capabilities. Extended context windows do not uniformly translate into effective utilization across different task types and document positions. Task complexity affects retrieval accuracy tremendously, with models

**Table 7:** Integration of quantitative results and qualitative insights (RQ1–RQ2)

<b>Quantitative finding</b>	<b>Consultant interpretation</b>	<b>Practical implication</b>
RQ1 Performance drops as context increases especially when reasoning is required.	Agreed that this was the case, and consultant experience in the field supports this finding as well.	Consider limiting context when possible for efficiency and accuracy, possibly utilize RAG pipelines.
Extraction accuracy declines mildly for information located late in documents (RQ2)	Consultants agreed that the data showed mild effects, their experience in the field supported this finding too.	Reinforces need for position aware retrieval or confidence indicators in production systems.
Overall extraction reliability rated as surprisingly good.	Consultants emphasized that the Finnish performance of these models was surprisingly good and suitable for many tasks and sectors.	Finnish companies should implement LLMs for information extraction tasks. In sensitive sectors and with tasks requiring reasoning, human in the loop validation is required.

demonstrating strong performance on direct information extraction while experiencing substantial accuracy degradation when tasks require arithmetic calculation or logical deduction. Positional biases still remain with commercial models, document depth has minor effects on accuracy.

### 5.5.1 Implications for Enterprise Adoption

Expert assessment of the results gave a varied view on enterprise deployment. Current LLM performance supports deployment for many mission critical data extraction tasks with human oversight, but not for autonomous operation in high risk sectors, especially if the use case requires reasoning.

**Consultant A:** "These results show we should be deploying LLMs for document processing tasks, but with hybrid approaches. The technology augments human capability but it doesn't replace it for high stakes scenarios."

**Consultant B:** "Healthcare will likely be one of the last sectors to move to autonomous LLM deployment, even as performance continues to improve. But that doesn't mean the technology isn't ready for mission critical deployment *with* human oversight. The benchmark results indicate we should be accelerating adoption for human augmented workflows immediately.

The expert interviews suggest that current LLM performance levels represented in the benchmark results are sufficient for mission critical enterprise applications when deployed with appropriate human oversight. The healthcare sector's perspective particularly emphasizes that source data quality represents a more significant adoption barrier than model capability, and that sector risk profiles will continue to mandate human in the loop validation regardless of accuracy improvements. Cost optimization through minimal context length usage emerges as a practical consideration that aligns well with observed performance stability across context sizes.

**The empirical** findings indicate that organizations must carefully match architectural choices to specific task requirements rather than defaulting to maximum context lengths. Extended context windows deliver value primarily when extraction tasks require genuine cross referencing across full documents. For most routine extraction work, shorter targeted contexts or possibly using retrieval-augmented generation pipelines achieve better or similar accuracy with substantially lower computational cost ([Jin et al., 2024](#)).

Document structure directly impacts extraction reliability. Organizations should place critical information early in prompts and documents rather than burying it in middle sections where accuracy degrades. When pre-processing pipelines allow document reformatting, surfacing high-priority information near context boundaries mitigates the observed positional bias. These structural interventions improve accuracy without model changes ([Liu et al., 2024](#)).

Tasks requiring arithmetic or multi-step reasoning demand additional safeguards. The elevated error rates on reasoning tasks necessitate decomposition techniques that break complex queries into simpler components, integration of external tools like calculators, or

verifier prompts that validate outputs. For mission-critical applications, human checkpoint reviews remain essential before committing model outputs to operational workflows, particularly when tasks extend beyond straightforward extraction into reasoning domains where current models show significant performance gaps.

## 5.6 Limitations

Several limitations constrain the scope and generalizability of this study’s findings. First, the evaluation encompassed only three model families Gemini 2.0 Flash, Llama 4 Maverick, and GPT-4.1 Mini representing a snapshot of the current commercial landscape. Results may not generalize to other long-context LLM architectures, open-source alternatives, or future model releases. Moreover, even within the same model family, different versions can exhibit substantially varied performance characteristics, limiting the temporal stability of these findings.

The reliance on Finnish Wikipedia as the exclusive source of haystack content introduces domain-specific constraints. While Wikipedia provides clean, linguistically diverse text suitable for benchmarking, its structured encyclopedic format differs substantially from many enterprise document types. Real-world enterprise datasets often contain OCR noise, nested tables, form-based layouts, and unstructured content from collaboration platforms such as Confluence, which may present greater extraction challenges than the relatively clean Wikipedia text used here (Lee et al., 2025b,a). This structural difference potentially leads to overestimation of extraction capabilities in more chaotic enterprise environments. Additionally, the commercial models evaluated may have been pretrained on Wikipedia data, which could artificially inflate accuracy by enabling recognition of familiar content patterns rather than genuine long-context retrieval performance.

The design of reasoning tasks also presents limitations. The hard difficulty needles emphasized arithmetic and temporal reasoning, reflecting common enterprise information processing requirements. However, other reasoning modalities such as causal inference, analogical reasoning, or domain-specific expertise integration were not systematically

evaluated. Industry-specific reasoning tasks in specialized domains such as legal interpretation or medical diagnosis may yield different performance patterns than the general-purpose reasoning evaluated here.

Methodological standardization introduced technical constraints that affect interpretation. Token counts were standardized using a single tokenizer to enable consistent comparison across experimental conditions. In reality, each model employs distinct tokenization schemes, meaning the actual token lengths processed by different models varied for identical input text. This discrepancy complicates direct cross-model comparisons of context utilization at supposedly equivalent context lengths. Furthermore, this study prioritized accuracy as the primary performance metric, while cost and latency considerations were touched on qualitatively through expert interviews. Systematic benchmarking of throughput, response latency, and computational cost across varying context sizes was beyond the scope of this research. These efficiency dimensions are critical for enterprise deployment decisions and warrant dedicated investigation in future work.

Finally, the statistical analysis employed a linear mixed-effects model to establish a conservative, interpretable baseline for understanding global performance trends across architectures. This specification intentionally does not test for U-shaped performance patterns or complex non-linear relationships between context length and accuracy. Testing of potential U-shaped trends in some conditions was deferred to future research. The linear model provides a robust foundation for understanding directional effects, but more sophisticated non-linear modeling approaches may reveal additional nuances in how models process extended contexts.

## 5.7 Recommendations and Future Work

Future research should aim to **broaden both model and domain coverage**. This includes the integration of additional commercial and open-source large language models, as well as the extension of evaluations to more heterogeneous and noisy enterprise data sources

such as scanned PDF documents, electronic health record (EHR) exports, and regulatory filings. Such datasets would provide a more realistic assessment of model robustness beyond the relatively clean and structured context of Wikipedia.

Further work should systematically evaluate alternative processing pipelines, including the quantification of performance improvements achieved through the use of judge LLMs and retrieval-augmented generation (RAG) systems on comparable benchmarks. These tools have the potential to produce measurable gains in factual accuracy and context handling capacity, particularly in enterprise applications where domain specificity and document diversity are critical.

A promising direction for future experimentation is to develop a methodology for benchmarking Finnish question–answer pairs directly against their English equivalents. This could provide a more precise understanding of cross-lingual model performance and translation-related variance in evaluation metrics.

Additionally, comparative analyses between RAG-based and long-context approaches should be conducted to determine whether chunking strategies maintain their effectiveness when longer context windows are available. Directly testing models' long-context capabilities on industry-specific tasks and large code bases would further illuminate how increasing context length impacts reasoning accuracy, retrieval efficiency, and overall task performance.

## **5.8 Societal, Sustainability and Ethical Implications**

Knowledge work will be affected by LLMs' proven ability to extract and analyze data in complex contexts. In our interviews we kept talking about automation and keeping a human in the loop at some stage these tasks will be automated completely. This raises concerns regarding skill requirements, workforce transition, and the evolving nature of knowledge work, even though it may lessen cognitive load and boost productivity. Proactive workforce development strategies and careful integration of these technologies in ways that enhance rather than merely replace human capabilities are essential to the

social sustainability of these technologies.

Though studying sustainability effects was not directly at the scope of this study, the computational demands of extended context windows carry measurable environmental costs. Empirical measurements on Llama-65B show 3–4 joules of energy per decoded token under typical multi-GPU inference setups, with energy intensity rising as sharding increases illustrating how large-scale deployments can multiply per-request energy use. (Samsi et al., 2023)

Recent inference-energy studies decompose costs into prefill (reading the input/context) and decode (generating output) and find that decoding is more energy-intensive per token, while energy scales with input and output sequence lengths, in other words, pushing very long contexts increases prefill work, and longer generations linearly increase decode energy. System design materially lowers the footprint without sacrificing capability (Fernandez et al., 2025).

Organizations should save ultra-long contexts for situations where they obviously alter results and instead default to context-efficient designs such as retrieval, smaller windows, prompt shorteners and response-length controls. By making these decisions at scale, energy consumption and overall emissions are directly reduced, making sustainability an engineering goal rather than an afterthought.

## 5.9 Concluding Remarks

This thesis delivers the first, to my knowledge, Finnish-focused, long-context MNIAH benchmark with replicated, cross-model results and practitioner validation. The central message is optimistic, modern LLMs can handle *mission-critical, extraction-first* workflows in Finnish when humans are kept in the loop in Finnish. However, *reasoning under long context* provides poor results. The benchmark, findings, and recommendations presented here provide a concrete foundation for evidence-based adoption in Nordic enterprises.

## References

- Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., Hada, R., Jain, P., Axmed, M., Bali, K., and Sitaram, S. (2024). MEGEVERSE: Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks.
- Anthropic, C. s. (2025). Context Windows.
- Chan, B. J., Chen, C.-T., Cheng, J.-H., and Huang, H.-H. (2025). Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks. pages 893–897. WWW '25: The ACM Web Conference 2025.
- Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., Yang, X., McAuley, J., Petzold, L., and Wang, W. Y. (2024). A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law.
- Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I., Eckardt, J. N., Laleh, N. G., Löffler, C. M. L., Schwarzkopf, S. C., Unger, M., Veldhuizen, G. P., Wagner, S. J., and Kather, J. N. (2023). The future landscape of large language models in medicine. *Communications Medicine*, 3(1).
- Cowan, N., Elliott, E. M., Saults, S. J., Morey, C. C., Mattox, S., Hismjatullina, A., and Conway, A. R. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, 51(1):42–100.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- Fernandez, J., Na, C., Tiwari, V., Bisk, Y., Luccioni, S., and Strubell, E. (2025). Energy Considerations of Large Language Model Inference and Efficiency Optimizations. Technical report, Hugging Face.
- Fu, Y. (2024). Challenges in Deploying Long-Context Transformers: A Theoretical Peak Performance Analysis. Technical report, arXiv.

- Gianfrancesco, M. A. and Goldstein, N. D. (2021). A narrative review on the validity of electronic health record-based research in epidemiology. *BMC Medical Research Methodology*, 21(1).
- Goldbaum, N. (2025). Tiktoken Repository.
- Hsieh, C.-Y., Chuang, Y.-S., Li, C.-L., Wang, Z., Le, L. T., Kumar, A., Glass, J., Ratner, A., Lee, C.-Y., Krishna, R., and Pfister, T. (2024). Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization.
- Huang, Y., Tai, W., Zhou, F., Gao, Q., Zhong, T., and Zhang, K. (2025). Extracting key insights from earnings call transcript via information-theoretic contrastive learning. *Information Processing and Management*, 62(3).
- Jin, B., Yoon, J., Han, J., and Ö Arık, S. (2024). LONG-CONTEXT LLMS MEET RAG: OVERCOMING CHALLENGES FOR LONG INPUTS IN RAG. Technical report.
- Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall Series in Experimental Psychology. Prentice-Hall, Englewood Cliffs, NJ.
- Kamradt, G. (2023). Needle In A Haystack Repository.
- Kruse, M., Hu, S., Derby, N., Wu, Y., Stonbraker, S., Yao, B., Wang, D., Goldberg, E., and Gao, Y. (2025). Large Language Models with Temporal Reasoning for Longitudinal Clinical Summarization and Prediction.
- Lauren, M. and Whitehouse, N. (2024). Better Call GPT, Comparing Large Language Models Against Lawyers. Technical report, ArXiv.
- Lavie, N. (1995). Perceptual Load as a Necessary Condition for Selective Attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):451–468.
- Lee, J., Stevens, N., and Han, S. C. (2025a). Large Language Models in Finance (FinLLMs). *Neural Computing and Applications*.

- Lee, T., Yoon, C., Jang, K., Lee, D., Song, M., Kim, H., and Kang, J. (2025b). ETHIC: Evaluating Large Language Models on Long-Context Tasks with High Information Coverage.
- Li, M., Chao, Q., and Li, B. (2025a). Two Causally Related Needles in a Video Haystack.
- Li, M., Zhang, S., Zhang, T., Duan, H., Liu, Y., and Chen, K. (2025b). NeedleBench: Can LLMs Do Retrieval and Reasoning in Information-Dense Context?
- Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Liu, N., Wang, L., Song, W., Zhai, X., Song, K., Zhu, D., Li, B., Wang, X., and Liu, T. (2024a). Large Language Models for Manufacturing.
- Li, Z., Li, C., Zhang, M., Mei, Q., and Bendersky, M. (2024b). Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. Technical report.
- Li, Z., Qiu, W., Ma, P., Li, Y., Li, Y., He, S., Jiang, B., Wang, S., and Gu, W. (2024c). An Empirical Study on Large Language Models in Accuracy and Robustness under Chinese Industrial Scenarios.
- Liu, J., Zhu, D., Bai, Z., He, Y., Liao, H., Que, H., Wang, Z., Zhang, C., Zhang, G., Zhang, J., Zhang, Y., Chen, Z., Guo, H., Li, S., Liu, Z., Shan, Y., Song, Y., Tian, J., Wu, W., Zhou, Z., Zhu, R., Feng, J., Gao, Y., He, S., Li, Z., Liu, T., Meng, F., Su, W., Tan, Y., Wang, Z., Yang, J., Ye, W., Zheng, B., Zhou, W., Huang, W., Li, S., and Zhang, Z. (2025). A Comprehensive Survey on Long Context Language Modeling.
- Liu, J. Z. and Li, X. (2024). How do judges use large language models? Evidence from Shenzhen. *Journal of Legal Analysis*, 16(1):235–262.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the Middle: How Language Models Use Long Contexts. Technical report.

- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts.
- Lu, Y., Yan, J. N., Yang, S., Chiu, J. T., Ren, S., Yuan, F., Zhao, W., Wu, Z., and Rush, A. M. (2024). A Controlled Study on Long Context Extension and Generalization in LLMs.
- Marois, R. and Ivanoff, J. (2005). Capacity limits of information processing in the brain.
- Mirishli, S. S. (2025). Regulating Ai In Financial Services.
- Muthusamy, V., Rizk, Y., Gulati, A., and Dube, P. (2023). Towards large language model-based personal agents in the enterprise: Current trends and open problems. Technical report, IBM.
- OpenRouter (2025). Openrouter Model Library.
- Oxford Insights (2024). 2024 Government AI Readiness-Index 2. Technical report.
- Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., and Yu, P. S. (2025). A survey of multilingual large language models.
- Sack, D., Foege, T., Kirvelä, S., Gray, A., Morin, M., and Axelsson, O. (2025). GenAI Complacency The Costly Inaction in the Nordics. Technical report, BCG.
- Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., and Gadepally, V. (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference.
- Swanson, H. L. (1987). Information Processing Theory and Learning Disabilities: An Overview: An Overview. *Journal of Learning Disabilities*, 20(1):3–7.
- Tanskanen, A. (2023). Finnish-NLP/wikipedia\_20231101\_fi\_cleaned.
- Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. (2022). Efficient Transformers: A Survey.

- The Guardian (2025). Misuse of AI in the legal sector. *The Guardian*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need.
- Wang, X., Salmani, M., Omid, P., Ren, X., Rezagholizadeh, M., and Eshaghi, A. (2024a). Beyond the Limits: A Survey of Techniques to Extend the Context Length in Large Language Models. Technical report.
- Wang, Y. (2025). Reasoning on Multiple Needles In A Haystack.
- Wang, Z., Chu, Z., Doan, T. V., Ni, S., Yang, M., and Zhang, W. (2024b). History, Development, and Principles of Large Language Models-An Introductory Survey.
- Wickens, C. (2021). Attention: Theory, Principles, Models and Applications. *International Journal of Human-Computer Interaction*, 37(5):403–417.
- Yang, X., Zang, S., Ren, Y., Peng, D., and Wen, Z. (2024). Evaluating Large Language Models on Financial Report Summarization: An Empirical Study.
- Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., and Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

## Appendix A: Prompt and Needles & Their Context

### Prompt

The prompt used in the experiments was designed to instruct the language model to carefully read a long context passage and answer all questions based solely on the provided text. The instructions emphasized precision, factual extraction, and avoidance of speculative or explanatory additions. The full structure of the prompt was as follows:

"Tässä on pitkä teksti. Lue se huolellisesti ja vastaa KAIKKIIN kysymyksiin tekstin perusteella. Ole erittäin tarkka ja etsi tarkat vastaukset tekstistä.

Mikäli et löydä vastausta kysymykseen annetusta tekstistä, jätä vastaus tyhjäksi.

context

Vastaa seuraaviin kysymyksiin. Anna jokainen vastaus omalle rivilleen numeroiden kanssa. Ole tarkka ja etsi täsmälliset annetusta tekstistä löytyvät vastaukset:

questions section

**TÄRKEÄÄ:** Anna vain lyhyt, tarkka vastaus jokaiseen kysymykseen. Älä selitä tai lisää ylimääräistä tekstiä.

Vastaukset: 1. 2. 3. 4. 5. 6. 7. 8. 9. 10."

The prompt structure ensured that each question was answered on a separate line, and that the model refrained from providing explanations, commentary, or reasoning steps beyond the literal information found in the text. This strict formulation aimed to evaluate the model's factual extraction and reading comprehension capabilities in a controlled setting.

## Needles

The following section describes the ten "needles" used in the experiment. Each needle consisted of a short factual passage, an associated question, and an expected answer. They were divided into two groups: five medium-difficulty items and five hard-difficulty items. The medium ones tested direct factual retrieval, while the harder ones required multi-step reasoning or numerical inference.

### Question 1:

- **Teksti:** Jyväskylän erikoistutkimuslaitos Hydrotech julkaisi vuonna 2024 raportin, jonka mukaan kaupungin vesipinta-alan osuus on tarkalleen 73,2 neliökilometriä. Tämä luku sisältää Päijänteen rannikkoalueet ja kaupungin sisäiset vesistöt.

- **Kysymys:** Mikä on Jyväskylän vesipinta-alan osuus neliökilometreinä Hydrotech-raportin mukaan?
- **Vastaus:** 73,2

#### Question 2:

- **Teksti:** Azerbaidžanin uusi energiaministeri Rashid Aliyev nimitti vuonna 2024 erityisneuvonantajakseen kansainvälisen öljyteollisuuden asiantuntijan Reddick Millerin, joka johti aiemmin Norjan valtion öljyrahastoa.
- **Kysymys:** Kuka on Azerbaidžanin energiaministeri Rashid Aliyevin erityisneuvonantaja?
- **Vastaus:** Reddick Miller

#### Question 3:

- **Teksti:** Albanian parlamentti hyväksyi vuonna 2024 erityislain, jonka mukaan maan ulkomaankaupan osuus BKT:sta saa nousta enintään 67,8 prosenttiin seuraavan viiden vuoden aikana.
- **Kysymys:** Mikä on Albanian ulkomaankaupan enimmäisosuus BKT:sta vuoden 2024 erityislain mukaan?
- **Vastaus:** 67,8

#### Question 4:

- **Teksti:** Andorran kansallisarkisto nimesi vuonna 2023 pääarkistaariksi tohtorin Maria Gonzalez-Ferrandezin, joka oli aiemmin työskennellyt Barcelonan historiallisessa arkistossa.
- **Kysymys:** Kuka nimettiin Andorran kansallisarkiston pääarkistaariksi vuonna 2023?
- **Vastaus:** Maria Gonzalez-Ferrandez

#### Question 5:

- **Teksti:** Afganistanin maatalousministeriön uusi strategia keskittyy viljelyn tehostamiseen kuivuuden kestävien lajikkeiden avulla, erityisesti durra-viljan edistämiseen.
- **Kysymys:** Minkä viljalajin edistämiseen Afganistanin maatalousministeriön uusi strategia keskittyy?
- **Vastaus:** durra

#### Question 6:

- **Teksti:** Kansainvälisen avaruustutkimuskeskuksen vuoden 2024 raportti listasi ISS-aseman tieteelliset kokeet kolmeen kategoriaan: Materiaalitieteen kokeet (9 koetta), biologiset tutkimukset (4 koetta) ja fysikaaliset kokeet (2 koetta).
- **Kysymys:** Kuinka monta koetta ISS-aseamalla suoritettiin yhteensä vuoden 2024 raportin mukaan?
- **Vastaus:** 15

#### Question 7:

- **Teksti:** Albanian museon johtaja Sofia Petrit aloitti työnsä tammikuussa 2023. Hänen edeltäjänsä Marko Gjika työskenteli museossa 8 vuotta ennen eläköitymistään. Gjikan edeltäjä Elena Koci johti museota 12 vuotta.
- **Kysymys:** Minä vuonna Elena Koci aloitti Albanian museon johtajana?
- **Vastaus:** 2003

#### Question 8:

- **Teksti:** Wakandan energiaministeriön raportin mukaan maan sähköntuotannosta 20% tulee vesivoimalaitoksista ja loput fossiilisista polttoaineista. Ministeriö ilmoitti, että fossiilisten polttoaineiden käyttöä vähennetään 15 prosenttiyksiköllä seuraavan viiden vuoden aikana.
- **Kysymys:** Mikä on fossiilisten polttoaineiden nykyinen osuus Wakandan sähköntuotannosta?

- **Vastaus:** 80

**Question 9:**

- **Teksti:** Aamulan yliopiston kielitieteen laitoksen 240 opiskelijasta 60% opiskelee suomea, 25% viroa ja loput muita kieliä. Muita kieliä opiskelevista puolet keskittyy germaanisiin kieliin.
- **Kysymys:** Kuinka monta Aamulan yliopiston opiskelijaa opiskelee germaania kieliä?
- **Vastaus:** 18

**Question 10:**

- **Teksti:** Afganistanin geologinen tutkimuslaitos kehitti vuonna 2023 kolme uutta menetelmää mineraalien etsintään. Ensimmäinen menetelmä oli 'pintaskannaus', toinen menetelmä 'keskisyvyysanalyysi'. Kolmas menetelmä 'syvätunnistustekniikka'. Toinen menetelmä osoittautui tehokkaimmaksi 89% tarkkuudella ja sitä alettiin käyttää erityisesti harvinaisten maametallien etsinnässä.
- **Kysymys:** Minkä nimistä menetelmää Afganistan käyttää harvinaisten maametallien etsintään?
- **Vastaus:** keskisyvyysanalyysi

Each needle thus represented a factual anchor embedded within a synthetic passage, designed to test the model's retrieval accuracy and reasoning depth under uniform prompt conditions.

## Appendix B: Test Results with Only the Question and Text as Context

This appendix reports model performance when evaluated using only the question and the associated context passage. 0 Wikipedia passages are fed to the models in this baseline evaluation. All models were tested with each question 75 times.

### Per-Question Accuracy (Q1–Q10)

$$\text{Accuracy}_{q_i} = \frac{\text{Correct Answers}_{q_i}}{\text{Total Answers}_{q_i}}$$

where:

- $\text{Accuracy}_{q_i}$  is the accuracy for question  $i$ ,
- $\text{Correct Answers}_{q_i}$  is the number of correct responses given to question  $i$ , and
- $\text{Total Answers}_{q_i}$  is the total number of responses evaluated for that question.

#### gemini-2.0-flash

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

#### gpt-4.1-mini

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

#### llama-4-maverick

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
73%	100%	100%	100%	100%	100%	84%	99%	36%	100%

**Interpretation.** models gemini-2.0-flash-001, and gpt-4.1-mini) achieved stable, perfect scores across all questions. The llama-4-maverick model performed strongly overall but exhibited sensitivity on a subset of items, particularly Q9.

## Appendix C: Interview Transcripts

This appendix contains the transcripts of the interviews conducted as part of the research. Each transcript has been anonymized where necessary to protect participant privacy.

### Interview 1

Interview with Consultant A (AI Consultant in the Legal Sector)

Interviewer: Thank you for joining today's interview. Could you start by describing your current work? In which industries have you worked, especially with machine learning and language models? And how long have you been in this field?

Consultant A: Sure. My career in artificial intelligence and machine learning began in 2021, when I first worked with a supervised machine learning model designed to categorize customer service tickets into the correct product lines. It was a basic NLP-based classification model that processed text and sorted tickets accordingly. After that, I continued in the same company, working on automating the categorization of purchase invoices using machine learning. Later, I moved into consulting, where generative AI became a major focus particularly in areas like information retrieval and text-based automation. One of the main domains I've worked in is the legal field, where we've built solutions that help extract relevant information from large volumes of legal text. The goal has been to create tools that allow legal professionals to query and analyze large document sets more efficiently, rather than manually reviewing them one by one.

Interviewer: We looked at the model evaluation results together earlier. To summarize, the best-performing model was Gemini 2.0 Flash, which correctly answered a lot of questions. The next was GPT-4.1 Mini, and the lowest was Llama 4 Maverick. For simpler information extraction tasks, accuracy was even higher around 99% for Gemini and 94–95% for the others. In the legal context, precision is obviously critical. Does this level of accuracy seem sufficient to you? And have you encountered cases where language models are expected to perform reasoning over documents, not just extract data?

Consultant A: Those numbers are quite consistent with what we've seen in the legal sector.

Simple numerical or factual extraction like identifying dates or amounts from contracts works very reliably. Errors become more common when the task requires domain understanding. For example, legal terms such as *de minimis* can appear in different forms across documents, and if the model doesn't understand their legal meaning, extraction can easily fail. In legal contexts, I feel that around 75 percent of the LLM use cases are extraction use cases, but for use cases requiring reasoning, the pipelines need additional steps. And in extraction use cases where accuracy is important, I still would include a human in the loop even with near-100 percent accuracy in your benchmark. In our experience, when building internal knowledge bases for law firms, it's important to have a human-in-the-loop to validate extracted data. Generative AI can extract information accurately, but if the extracted data is later used for critical actions, human validation is essential before trusting it completely.

Interviewer: So you'd say that for high-stakes tasks, models still shouldn't operate fully autonomously?

Consultant A: Exactly. Especially in legal or financial contexts, human validation remains necessary. AI systems are extremely helpful, but you can't yet let them run entirely unsupervised for critical operations.

Interviewer: What about pure information extraction? If the model reaches, say, 99% accuracy, is that remaining 1% error rate still too high?

Consultant A: It depends entirely on how the information will be used. If you're producing general summaries or statistics for example, average compensation amounts from the last hundred cases a small margin of error is acceptable. However, if you're providing a specific fact to a client, even one wrong answer could be serious. That said, the time savings from automation often outweigh the occasional small mistake. Accuracy can also be improved further using techniques like judge layers, which double-check extracted information.

Interviewer: How often do legal teams require data that must be absolutely correct?

Consultant A: Quite often, but not always. For example, during large transactions like property or company acquisitions lawyers often review historical cases to understand what warranty terms have typically been agreed upon. In such situations, approximate distributions or summaries are fine. However, when a client asks about the specific details of a particular contract, that information must be exact. Usually, such facts are verified either manually or through validated databases.

Interviewer: These tests were conducted in Finnish, while most of these models are trained primarily on English data. Were you surprised by how well they performed in Finnish?

Consultant A: I was really surprised at the accuracy rates around extracting simple pieces of information; the models are more accurate using Finnish than I expected. I expected the performance on Finnish extraction tasks to be worse, but the results were surprisingly good better than I anticipated.

Interviewer: Based on these results, what advice would you give to an organization in Finland considering deploying such models in production?

Consultant A: If the use case involves straightforward information extraction, deploying models in Finnish is already feasible. Processing costs are decreasing, and these results show that accuracy remains high even in smaller languages. The hard questions in your benchmark the ones with 0 percent success rates those represent the kind of complex reasoning we encounter in legal document analysis. Legal documents often also contain industry-specific words which the LLMs don't handle as accurately.

However, before fine-tuning or building complex retrieval-augmented systems, I'd first evaluate whether an off-the-shelf product meets the needs. If not, then consider building a custom retrieval solution either internally or through a service provider.

The same principles apply beyond law for example, in healthcare or software development whenever long, sensitive documents are involved and mistakes must be minimized.

Interviewer: You mentioned programming. How do these findings relate to large models'

performance in coding contexts?

Consultant A: That's interesting. From my experience using AI-assisted code editors, expanding the context window doesn't necessarily bring major benefits. In programming, it's often more effective to limit edits to specific files and maintain control. This aligns with your findings as context length increases, reasoning quality decreases. If models' reasoning with long contexts improves in the future, managing large codebases could become smoother. Currently, though, models can behave unpredictably when given excessive context. Interviewer: Is there anything organizations should consider beyond model accuracy before deploying these systems for critical tasks?

Consultant A: More generally, it's valuable to distinguish between information that must be exact and information that can be approximate. For high-precision needs, it's worth building structured databases validated by humans. For exploratory or high-volume queries where slight errors are acceptable, AI-only processing is fine. These results show we should be deploying LLMs for document processing tasks, but with hybrid approaches. The technology augments human capability, but it doesn't replace it for high-stakes scenarios. The key is systematic process design use AI to save time where accuracy is "good enough," and maintain human oversight where accuracy is critical.

Interviewer: Finally, do you think companies should more aggressively implement these models in production? Are accuracy concerns still holding them back?

Consultant A: Yes performance concerns are still the main reason for hesitation. In almost every project, the first question stakeholders ask is, "Why did the model answer this incorrectly?" However, I believe companies should adopt AI much more extensively. Every organization has large document collections that could be made searchable or analyzable with AI tools. If 95% accuracy is good enough, automation can save significant time. For cases that demand full accuracy, structured, validated reporting pipelines are the way to go. And as your results show, with accuracies up to 99% in simple extraction, we're already seeing surprisingly strong performance.

Interviewer: Excellent points. Thank you for the insightful discussion.

Consultant A: Thank you, glad to participate.

## **Interview 2**

Interviewer: Okay, yes, let's begin the interview. Hello. How are you doing today?

Consultant B: Good, good.

Interviewer: Let's begin with your expertise. Can you tell me a bit about your background? How long have you worked with AI and LLMs, and in what kind of industries?

Consultant B: I've been working with AI for around five to six years. This includes classical machine learning, data science, and optimization. I've been working with LLMs for the past one and a half to two years, across different sectors and applications.

Interviewer: Nice. What kind of sectors have you worked with?

Consultant B: I've worked with the healthcare sector, public sector, private sector, and gaming sector. Interviewer: We looked at the results from the benchmark test we did together. To summarize: we ran a multiple "needles in a haystack" test with ten total needles. We used Finnish Wikipedia data as the background and inserted the facts at different depths in the document. We tested both extraction accuracy as the context length increased (up to one million tokens) and positional effects — whether the position of inserted facts affected extraction accuracy. At first glance, Gemini Flash performed very well, finding over 90% of the easily extractable facts for the medium-difficulty questions. Were you surprised by these results, or were they in line with what you've seen in your work?

Consultant B: Most of the results were in line with my expectations for extraction tasks, but I have to say Gemini's performance was genuinely surprising. I thought OpenAI would provide the best results. That said, the findings generally match what I've seen — Gemini performed especially well compared to GPT-4.1 mini, which I didn't expect.

Interviewer: Let's move to reasoning performance. There were five reasoning questions, and two of them were so hard that if any additional context was provided, the models failed — but with no context, they achieved High accuracy. You mentioned working in

the healthcare sector. How often do use cases in that field require reasoning versus simple data extraction?

Consultant B: Healthcare use cases are predominantly extraction-heavy rather than reasoning-heavy. Patient record summarization and many others — these are all tasks where the benchmark results indicate readiness, provided we maintain human verification in the workflow. There are also additional methods you can use to improve accuracy results.

Interviewer: So, even though extraction seems strong, reasoning might still be a limiting factor?

Consultant B: Exactly. In healthcare, the sector simply cannot afford mistakes, even rare ones. Even if we had a model performing at 99% accuracy for medication extraction or dosage information, we would still require human-in-the-loop validation. The risk profile is just too high. The benchmark results are good, but they confirm we're not at autonomous deployment for critical tasks. For reasoning tasks, I would look at implementing judges to make sure the answers are grounded on the original text, and possibly RAG systems to control context in order to increase accuracy.

Interviewer: You mentioned there are ways to further improve accuracy. Could you expand on that?

Consultant B: One approach would be a validation layer — for example, a second model that acts as a judge to evaluate whether the main model's response aligns with the source data. Another would be to increase dataset size and diversity. Preprocessing can also help — identifying relevant information before it reaches the main model. You can structure this as a pipeline: one model extracts or filters, another generates answers, and a third validates the output. This multi-step setup can raise accuracy significantly.

Interviewer: Your benchmark used clean Wikipedia data. How does that compare to real-world conditions?

Consultant B: Your benchmark uses clean Wikipedia data, which is valuable for controlled

testing. But the largest barrier to enterprise adoption isn't model capability — it's source data quality. In real healthcare implementations, we're dealing with inconsistently formatted medical records, handwritten notes that have been poorly digitized, and incomplete documentation. These issues create noise and ambiguity that make the task much harder than in controlled tests.

Interviewer: In your experience, how important are long context windows — for example, up to one million tokens?

Consultant B: Based on these results, the clear recommendation for healthcare implementations is to keep context windows as small as possible. You gain both accuracy and cost benefits with smaller context sizes. If you're using a million tokens per API call, you're probably doing something wrong. Most efficient applications are designed with smaller, more focused contexts that perform better and cost less.

Interviewer: Given the progress seen here, how close are we to deploying LLMs autonomously in sensitive fields like healthcare?

Consultant B: Healthcare will likely be one of the last sectors to move to autonomous LLM deployment, even as performance continues to improve. But that doesn't mean the technology isn't ready for mission-critical deployment with human oversight. The benchmark results indicate we should be accelerating adoption for human-augmented workflows immediately. The systems are capable enough — it's about putting the right guardrails and verification in place.

Interviewer: So to summarize, we're seeing strong extraction performance, limited reasoning reliability at scale, but clear readiness for human-supervised applications.

Consultant B: Exactly. The models are showing real potential. With structured pipelines, careful preprocessing, and human validation, we can already start deploying these systems responsibly in sectors like healthcare.

Interviewer: Excellent, thank you for sharing these insights.

Consultant B: Thank you.