

WORLD
OF
FINANCE

a wipro company

THE CAPCO INSTITUTE
JOURNAL
OF FINANCIAL TRANSFORMATION

TECHNOLOGY

Innovating with intelligence:
Open-source Large Language Models
for secure system transformation

GERHARDT SCRIVEN | TONY MOENICKE
SEBASTIAN EHRIG



GenAI

2024/2025 EDITION

THE CAPCO INSTITUTE

JOURNAL OF FINANCIAL TRANSFORMATION

RECIPIENT OF THE APEX AWARD FOR PUBLICATION EXCELLENCE

Editor

Shahin Shojai, Global Head, Capco Institute

Advisory Board

Lance Levy, Strategic Advisor

Owen Jelf, Partner, Capco

Suzanne Muir, Partner, Capco

David Oxenstierna, Partner, Capco

Editorial Board

Franklin Allen, Professor of Finance and Economics and Executive Director of the Brevan Howard Centre, Imperial College London and Professor Emeritus of Finance and Economics, the Wharton School, University of Pennsylvania

Philippe d'Arvisenet, Advisor and former Group Chief Economist, BNP Paribas

Rudi Bogni, former Chief Executive Officer, UBS Private Banking

Bruno Bonati, Former Chairman of the Non-Executive Board, Zuger Kantonalbank, and President, Landis & Gyr Foundation

Dan Breznitz, Munk Chair of Innovation Studies, University of Toronto

Urs Birchler, Professor Emeritus of Banking, University of Zurich

Elena Carletti, Professor of Finance and Dean for Research, Bocconi University, Non-Executive Director, UniCredit S.p.A.

Lara Cathcart, Associate Professor of Finance, Imperial College Business School

Géry Daeninck, former CEO, Robeco

Jean Dermine, Professor of Banking and Finance, INSEAD

Douglas W. Diamond, Merton H. Miller Distinguished Service Professor of Finance, University of Chicago

Elroy Dimson, Emeritus Professor of Finance, London Business School

Nicholas Economides, Professor of Economics, New York University

Michael Enthoven, Chairman, NL Financial Investments

José Luis Escrivá, President, The Independent Authority for Fiscal Responsibility (AIReF), Spain

George Feiger, Pro-Vice-Chancellor and Executive Dean, Aston Business School

Gregorio de Felice, Head of Research and Chief Economist, Intesa Sanpaolo

Maribel Fernandez, Professor of Computer Science, King's College London

Allen Ferrell, Greenfield Professor of Securities Law, Harvard Law School

Peter Gomber, Full Professor, Chair of e-Finance, Goethe University Frankfurt

Wilfried Hauck, Managing Director, Statera Financial Management GmbH

Pierre Hillion, The de Picciotto Professor of Alternative Investments, INSEAD

Andrei A. Kirilenko, Reader in Finance, Cambridge Judge Business School, University of Cambridge

Katja Langenbacher, Professor of Banking and Corporate Law, House of Finance, Goethe University Frankfurt

Mitchel Lenson, Former Group Chief Information Officer, Deutsche Bank

David T. Llewellyn, Professor Emeritus of Money and Banking, Loughborough University

Eva Lomnicka, Professor of Law, Dickson Poon School of Law, King's College London

Donald A. Marchand, Professor Emeritus of Strategy and Information Management, IMD

Colin Mayer, Peter Moores Professor of Management Studies, Oxford University

Francesca Medda, Professor of Applied Economics and Finance, and Director of UCL Institute of Finance & Technology, University College London

Pierpaolo Montana, Group Chief Risk Officer, Mediobanca

John Taysom, Visiting Professor of Computer Science, UCL

D. Sykes Wilford, W. Frank Hipp Distinguished Chair in Business, The Citadel

CONTENTS

TECHNOLOGY

08 Mindful use of AI: A practical approach

Magnus Westerlund, Principal Lecturer in Information Technology and Director of the Laboratory for Trustworthy AI, Arcada University of Applied Sciences, Helsinki, Finland

Elisabeth Hildt, Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland, and Professor of Philosophy and Director of the Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, USA

Apostolos C. Tsolakis, Senior Project Manager, Q-PLAN International Advisors PC, Thessaloniki, Greece

Roberto V. Zicari, Affiliated Professor, Arcada University of Applied Sciences, Helsinki, Finland

14 Understanding the implications of advanced AI on financial markets

Michael P. Wellman, Lynn A. Conway Collegiate Professor of Computer Science and Engineering University of Michigan, Ann Arbor

20 Auditing GenAI systems: Ensuring responsible deployment

David S. Krause, Emeritus Associate Professor of Finance, Marquette University

Eric P. Krause, PhD Candidate – Accounting, Bentley University

28 Innovating with intelligence: Open-source Large Language Models for secure system transformation

Gerhardt Scriven, Executive Director, Capco

Tony Moenicke, Senior Consultant, Capco

Sebastian Ehrig, Senior Consultant, Capco

38 Multimodal artificial intelligence: Creating strategic value from data diversity

Cristián Bravo, Professor, Canada Research Chair in Banking and Insurance Analytics, Department of Statistical and Actuarial Sciences, Western University

46 GenAI and robotics: Reshaping the future of work and leadership

Natalie A. Pierce, Partner and Chair of the Employment and Labor Group, Gunderson Dettmer

ORGANIZATION

56 How corporate boards must approach AI governance

Arun Sundararajan, Harold Price Professor of Entrepreneurship and Director of the Fubon Center for Technology, Business, and Innovation, Stern School of Business, New York University

66 Transforming organizations through AI: Emerging strategies for navigating the future of business

Feng Li, Associate Dean for Research and Innovation and Chair of Information Management, Bayes Business School (formerly Cass), City St George's, University of London

Harvey Lewis, Partner, Ernst & Young (EY), London

74 The challenges of AI and GenAI use in the public sector

Albert Sanchez-Graells, Professor of Economic Law, University of Bristol Law School

78 AI safety and the value preservation imperative

Sean Lyons, Author of Corporate Defense and the Value Preservation Imperative: Bulletproof Your Corporate Defense Program

92 Generative AI technology blueprint: Architecting the future of AI-infused solutions

Charlotte Byrne, Managing Principal, Capco

Thomas Hill, Principal Consultant, Capco

96 Unlocking AI's potential through metacognition in decision making

Sean McMinn, Director of Center for Educational Innovation, Hong Kong University of Science and Technology

Joon Nak Choi, Advisor to the MSc in Business Analytics and Adjunct Associate Professor, Hong Kong University of Science and Technology

REGULATION

104 Mapping GenAI regulation in finance and bridging the gaps

Nydia Remolina, Assistant Professor of Law, and Fintech Track Lead, SMU Centre for AI and Data Governance, Singapore Management University

112 Board decision making in the age of AI: Ownership and trust

Katja Langenbucher, Professor of Civil Law, Commercial Law, and Banking Law, Goethe University Frankfurt

122 The transformative power of AI in the legal sector: Balancing innovation, strategy, and human skills

Eugenia Navarro, Lecturer and Director of the Legal Operations and Legal Tech Course, ESADE

129 Remuneration on the management board in financial institutions: Current developments in the framework of supervisory law, labor law, behavioral economics and practice

Julia Redenius-Hövermann, Professor of Civil Law and Corporate Law and Director of the Corporate Governance Institute (CGI) and the Frankfurt Competence Centre for German and Global Regulation (FCCR), Frankfurt School of Finance and Management

Lars Hinrichs, Partner at Deloitte Legal Rechtsanwaltsgesellschaft mbH (Deloitte Legal) and Lecturer, Frankfurt School of Finance and Management



CAPCO CEO WELCOME

DEAR READER,

Welcome to our very special 60th edition of the Capco Journal of Financial Transformation.

The release of this milestone edition, focused on GenAI, reinforces Capco's enduring role in leading conversations at the cutting edge of innovation, and driving the trends shaping the financial services sector.

There is no doubt that GenAI is revolutionizing industries and rapidly accelerating innovation, with the potential to fundamentally reshape how we identify and capitalize on opportunities for transformation.

At Capco, we are embracing an AI infused future today, leveraging the power of GenAI to increase efficiency, innovation and speed to market while ensuring that this technology is used in a pragmatic, secure, and responsible way.

In this edition of the Capco Journal, we are excited to share the expert insights of distinguished contributors across academia and the financial services industry, in addition to drawing on the practical experiences from Capco's industry, consulting, and technology SMEs.

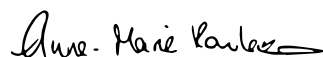
The authors in this edition offer fresh perspectives on the mindful use of GenAI and the implications of advanced GenAI on financial markets, in addition to providing practical and safe frameworks for boards and firms on how to approach GenAI governance.

The latest advancements in this rapidly evolving space demonstrate that the potential of GenAI goes beyond automating and augmenting tasks, to truly helping organizations redefine their business models, processes and workforce strategies. To unlock these benefits of GenAI, I believe that firms need a culture that encourages responsible experimentation and continuous learning across their organization, while assessing the impact of the potential benefits against a strategic approach and GenAI framework.

I am proud that Capco today remains committed to our culture of entrepreneurialism and innovation, harnessed in the foundation of our domain expertise across our global teams. I am proud that we remain committed to our mission to actively push boundaries, championing the ideas that are shaping the future of our industry, and making a genuine difference for our clients and customers – all while ensuring to lead with a strategy that puts sustained growth, integrity and security at the forefront of what we do.

I hope you'll find the articles in this edition both thought-provoking and valuable as you create your organization's GenAI strategy and future direction. As we navigate this journey together, now is the time to be bold, think big, and explore the possibilities.

My greatest thanks and appreciation to our contributors, readers, clients, and teams.



Annie Rowland, **Capco CEO**

INNOVATING WITH INTELLIGENCE: OPEN-SOURCE LARGE LANGUAGE MODELS FOR SECURE SYSTEM TRANSFORMATION

GERHARDT SCRIVEN | Executive Director, Capco

TONY MOENICKE | Senior Consultant, Capco

SEBASTIAN EHRIG | Senior Consultant, Capco

ABSTRACT

The rapid development of Large Language Models (LLMs) has revolutionized software development, yet the predominance of closed-source models has restricted their extensive adoption. In this paper, we explore open-source Large Language Models as an alternative to closed-source models like ChatGPT, particularly for the use case of interpreting legacy software source code. We evaluate open-source models for their capacity in understanding and explaining COBOL code to a human user, a crucial task for financial institutions looking to update their legacy systems while keeping their data secure in-house.

Evaluating LLMs in this domain is challenging since there's no simple right or wrong answer to the specific types of COBOL related questions we ask. Towards this, we have benchmarked the responses obtained from various proprietary and open-source LLMs against an expert human response. This method allows us to assess which models perform best for a specific type of question and are effective in a practical context.

This article provides insights for financial institutions looking to optimize or modernize their legacy systems using LLMs as well as offering considerations for adapting and integrating these models into their IT environments.

1. INTRODUCTION

In today's rapidly evolving business landscape, the demand for efficient and versatile artificial intelligence (AI) solutions has never been higher. Large Language Models have emerged as a transformative technology and are increasingly being adopted in modern businesses to elevate customer service standards, streamline internal documentation processes, and for the creation of content in diverse knowledge domains such as marketing.

Fine-tuning Large Language Models with proprietary data and domain-specific knowledge is often the driving force behind their adoption for specific use cases. This process allows

organizations to develop highly specialized solutions optimized for their unique operational challenges. Beyond optimizing workflow automation, enhancing data analysis capabilities, or facilitating internal communication, customized LLMs serve as a versatile toolkit for driving efficiency and productivity across diverse business functions.

Furthermore, as companies integrate LLMs into their operations, the decision between deploying them through a third-party hosted service or hosting them locally gains significant importance. Hosting LLMs locally provides better control over data privacy, allows for greater customization to meet specific business needs, and can reduce operational expenses.¹

¹ <https://tinyurl.com/mrysef5w>

However, since there is a range of LLMs available, each possessing unique capabilities and performance metrics, businesses face the challenge of selecting the most suitable model for their specific needs. Moreover, given the rapid advancement in this field, there is a pressing need for methods to efficiently evaluate new models as they emerge.

We have conducted a comprehensive evaluation of various leading LLMs currently available, specifically focusing on their ability to transform COBOL source code into tailored and highly consumable knowledge nuggets. This evaluation is designed to provide a more nuanced comparative view of LLM performance for specific use cases, particularly in the context of legacy code understanding, though the methodology can be applied to other domains as well. This involves assessing how well the LLMs can respond to certain types of questions, beyond just testing their domain expertise on the topic of COBOL. We aim to benchmark not only the models' proficiency in understanding COBOL code but also their capacity to abstract and reorganize information that may be highly fragmented across the technology stack.

In this, we want to shed light on the capabilities of these LLMs in addressing real-world business tasks, including COBOL code comprehension, customer query resolution, document analysis, and content generation, by exploring their ability to interpret code within an English context. Our assessment provides practical insights into effective methodologies for testing LLMs for specific tasks, offering valuable guidance for

businesses seeking to make informed decisions regarding their LLM strategy for addressing particular business problems.

1.1 The importance of LLMs for internal data processing

What was once a question of whether to adopt LLMs has transformed into a tactical consideration of how best to integrate them into existing enterprise operational frameworks. LLMs exhibit the capacity to address client queries through chatbots, screen extensive technical documentation for specific information, and generate compelling content for platforms spanning social media, public relations, and human resources.² However, effectively unlocking these benefits demands a critical decision – whether to opt for a paid model or open source.

Paid models typically offer superior performance, yet they may require sending potentially sensitive enterprise data outside the network boundaries, which could be unacceptable for highly regulated industries like finance. While some paid models can be deployed in private mode, meaning they can be hosted and operated within the organization's internal infrastructure, it is important to also consider the associated costs. On the other hand, open-source options pose the question of whether to host locally or via a third party, adding another layer of complexity. To clarify these choices, we have created a table with some selected models outlining hosting options for paid and open-source models along with their pros and cons (see Table 1).

Table 1: Comparison of hosting options for paid and open-source models: Pros and cons

LLM MODEL	HOSTING OPTION	PAID/OPEN-SOURCE	PROS	CONS
OpenAI's GPT-o1	Proprietary Cloud	Paid	State-of-the-art natural language processing capabilities	High computational costs, limited customization options
Llama3.1	Cloud-based / On-premises	Open-source ³	Strong performance in coding and reasoning, open-source flexibility	Requires significant computational resources for deployment
WizardLM-2-8x22B	Cloud-based / On-premises	Open-source	Strong performance in coding and reasoning, open-source flexibility	Not strongly aligned particularly in terms of safety and ethical considerations
DeepSeek V2.5	Cloud-based / On-premises	Open-Source ⁴	Strong performance in coding and reasoning, open-source flexibility	Demands significant computational power
Mixtral Large 2	Cloud-based / On-premises	Paid	State-of-the-art natural language processing capabilities	Requires substantial computational resources

² <https://tinyurl.com/47epujtd>

³ <https://tinyurl.com/bddb22ac>

⁴ <https://tinyurl.com/ynaxpm5w>

Table 2: Comparison of efficiency aspects across various deployment scenarios

DEPLOYMENT SCENARIO	SPEED	COST	REGULATORY COMPLIANCE	SCALABILITY
Local deployment	Moderate	High initial hardware procurement costs	Challenging due to regulatory requirements on hardware and data security	Limited by hardware capacity, may require additional investments for scaling
Cloud-based deployment	High	Variable based on usage and service provider	Compliance with industry standards facilitated by cloud provider certifications	Easily scalable based on cloud service offerings, pay-as-you-go model
On-premises deployment	Moderate	High initial setup and maintenance costs	Direct control over regulatory compliance measures, but requires internal expertise	Scalability limited by on-premises infrastructure, potential for costly upgrades
Hybrid deployment	Variable, depending on workload distribution	Combination of initial hardware costs and cloud service fees	Compliance challenges due to data movement between environments	Offers flexibility in scaling based on workload demands, potential cost optimization

1.1.1 CONTROL OVER DATA PRIVACY AND SECURITY

One of the primary motivations behind opting for local deployment of LLMs is the enhanced control over data privacy and security. By hosting LLMs on internal servers, companies maintain sovereignty over their sensitive information, mitigating the risks associated with third-party hosting. This approach aligns with industries governed by stringent data protection regulations, ensuring compliance and bolstering trust among stakeholders.

1.1.2 EFFICIENCY IN OPERATIONAL PROCESSES

Local deployment of LLMs brings about significant efficiencies in operational processes, particularly in data processing tasks. By leveraging the computational power of internal servers, companies can conduct intricate analyses, extract insights, and derive actionable intelligence from vast datasets in a timely manner. However, efficiency isn't solely about speed; it encompasses various other dimensions as well. For instance, if an organization needs to procure hardware to support local deployment, navigating through the procurement process, especially within regulated industries like banking, might be challenging. We have compared different efficiency aspects across various deployment scenarios in Table 2.

1.2 Assessing LLMs using COBOL code as a case study

In evaluating LLMs' performance, for example COBOL code analysis, it is crucial to understand their unique features, performance, and limitations. Quantifying these parameters aids in the selection of the most suitable model for specific needs. Establishing a repeatable process enables users to systematically evaluate both local and open-source LLMs, ensuring continuous assessment of new models against consistent benchmarks as they are released. Factors such as model size, computational requirements, and fine-tuning capabilities guide the adoption strategies. Understanding whether these models are open source or proprietary, along with their commercial availability, is essential for determining accessibility and potential integration into workflows or products. In the following sections we will provide an overview of selected LLMs and how we have evaluated their performance with respect to our COBOL code case.

2. OVERVIEW OF SELECTED OPEN-SOURCE LLMs

We explored the landscape of state-of-the-art language models as of September 2024. Although these models are available for download and analysis, not all of them may be used commercially due to licensing restrictions. Our focus narrows down to a handpicked selection of models that have demonstrated the most promising performance for understanding code.

Our assessment prioritizes two key factors: (1) the computational resources available to us, and (2) the quality of the models' outputs. Particularly, we underscore the importance of having GPUs with ample VRAM to efficiently run these models. Below is a brief overview of the models that we have used in the scope of this article.

2.1 Selection of local LLMs for evaluation

The selection of local Large Language Models for evaluation is critical for assessing their performance and capabilities across their intended usage tasks. Developers often choose specific LLMs based on factors such as model architecture, training data, and fine-tuning approaches to evaluate their effectiveness in real-world applications.⁵ Additionally, LLM leaderboards serve as valuable resources that benchmark and rank current LLMs according to different criteria and can be helpful in making an initial selection.⁶ Popular leaderboards are

for instance Big Code Models Leaderboard or LMSYS Chatbot Arena Leaderboard.^{7,8} These leaderboards enable developers to compare the strengths and weaknesses of different LLMs, guiding the selection of models for specific use cases based on their performance metrics.

2.2 Criteria for comparison

When comparing open-source Large Language Models with a focus on code-related tasks, several key criteria come into play to assess their effectiveness. These criteria include model performance, resource utilization, ease of deployment, context length, and code understanding.

1. **Model performance:** Evaluating model performance relies on benchmarks for different categories such as commonsense reasoning, reading comprehension, and code. Code benchmarks like HumanEval and MBPP test a model's ability to write Python code based on a description of the code's function, which then must pass a test.^{9,10} Another method to assess the LLM's performance involves using human evaluators to rate the responses. Experts in software development can review the quality of code generated by LLMs, providing feedback on how well the model understands and applies programming concepts, syntax, and idiomatic expressions.

Table 3: Overview of selected open-source LLMs

MODEL NAME	DESCRIPTION	PARAMETERS
DeepSeek V2.5	Combines the capabilities of DeepSeek-V2-Chat and DeepSeek-Coder-V2-Instruct, merging general conversational and coding skills.	236B
Llama-3.1-405b-instruct	405B instruct-tuned model with a 128k context window, optimized for dialogue and high performance against leading models.	405B
Llama-3.1-70b-instruct	Optimized model from Meta, fine-tuned for code-based tasks, exhibiting higher alignment with human preferences in dialogue interactions.	70B
Mixtral-8x22b-instruct	Mistral's 8x22B MoE model uses 39B active parameters out of 176B, with capabilities in math, coding and reasoning.	141B
WizardLM-2-8x22B	Microsoft AI's top Wizard model. It is an instruct fine-tune of the Mixtral 8x22B model.	141B

⁵ <https://tinyurl.com/y8yedm8j>

⁶ <https://tinyurl.com/sc735tm9>

⁷ <https://tinyurl.com/mt94a7j5>

⁸ <https://tinyurl.com/27x3eybj>

⁹ <https://tinyurl.com/2p9upajy>

¹⁰ <https://tinyurl.com/2ezh8uyc>

2. **Resource utilization:** Efficient resource utilization is essential for deploying LLMs in real-world applications. This criterion assesses how effectively the model utilizes computational resources such as CPU, GPU, memory, and storage during training and inference. Optimizing resource utilization ensures cost-effectiveness and scalability of the model.
3. **Ease of deployment:** The ease of deploying an LLM significantly affects its adoption and integration into existing software development workflows. Factors such as model size, compatibility with various programming languages and frameworks, and the availability of deployment options (like local, on-premise, or cloud-based) impact how straightforward or complex the deployment process is.
4. **Context length:** The context length refers to the number of tokens the model can effectively process and utilize in generating code-related outputs. According to OpenAI, “A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~ 75 words).”¹¹ Increasing the context length allows the model to process and analyze larger amounts of data (or longer sequences of text/code) at once.
5. **Code understanding:** Code understanding is a critical aspect of evaluating an LLM’s capability in code-related tasks. This criterion assesses how well the model comprehends programming languages, syntax, semantics, and idiomatic expressions commonly used in software development. A robust code understanding capability enables the model to provide accurate and contextually relevant suggestions and completions. While direct, methodical testing of “understanding” in the human sense might not be feasible, there are indirect methods like benchmarks and human evaluation to determine how well a model has learned to interpret and generate code.

When evaluating LLMs based on these criteria, companies can make informed decisions about selecting the most suitable model for their specific applications. Table 4 below provides an overview of some key selection criteria.

2.3 Key considerations for model selection

2.3.1 LICENSING

In the context of open-source Large Language Models, it is crucial to recognize that even though models may be open-source, the code they generate could still be subject to existing licenses. The Code Llama GitHub page underscores

Table 4: Overview of some key selection criteria of LLMs

MODEL NAME	PARAMETER COUNT	CONTEXT	RAM/VRAM REQUIREMENTS IN GiB (4-BIT/8BIT/16BIT PRECISION)	MODEL SIZE IN GiB (16-BIT PRECISION)	LICENSE
DeepSeek V2.5 ¹²	236	128	118/236/472	472	DeepSeek License Agreement
Llama-3.1-405b-instruct ¹³	405	128	202.5/405/810	810	Llama 3
Llama-3.1-70b-instruct ¹³	70	128	35/70/140	140	Llama 3
Mixtral-8x22b-instruct ¹⁴	141	64	70.5/141/282	282	Apache 2.0
WizardLM-2-8x22B ¹⁵	141	64	70.5/141/282	282	Apache 2.0

¹¹ <https://tinyurl.com/5n87rj3s>

¹² <https://tinyurl.com/4mamcp2>

¹³ <https://tinyurl.com/muy74yhd>

¹⁴ <https://tinyurl.com/y589zc9n>

¹⁵ <https://tinyurl.com/mry3y4m6>

that outputs from Llama models, including Code Llama, may be governed by third-party licenses. This means that while the Llama models themselves may be open-source, the code produced using these models might incorporate third-party rights or specific licensing conditions because code segments may unintentionally mirror those with restrictive usage terms found on platforms like GitHub. Therefore, users utilizing generated code that resembles licensed programs must adhere to the licensing conditions of the original code. By understanding these nuances, developers can navigate the complexities of licensing compliance effectively and ensure the ethical and lawful use of code generated by models like Llama 2.

2.3.2 QUANTIZATION

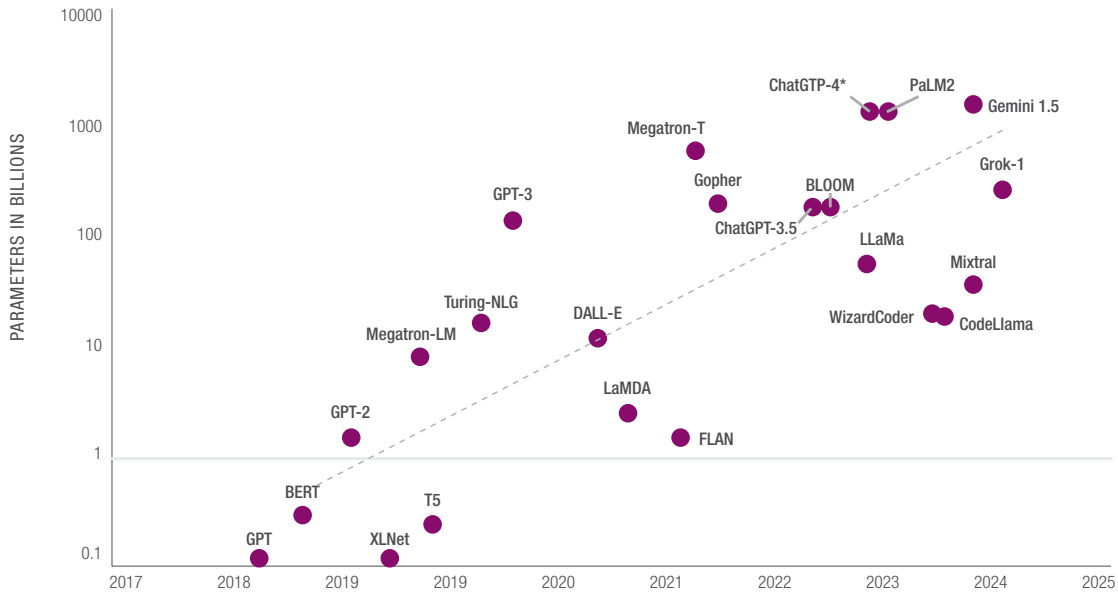
Quantization is a critical aspect to consider when selecting a Large Language Model for adoption by a company. It refers to the process of reducing the precision of numerical values in the model to enhance computational efficiency without significant loss in performance, which in turn positively affects computational resources requirements. For instance, quantizing an LLM can lead to reduced memory and computing requirements, making it more feasible for

deployment on hardware with limited resources. However, it is essential to balance these performance benefits of quantization with potential trade-offs in model accuracy. Companies should assess how different quantization techniques impact the LLM's inference speed, memory usage, and overall efficiency to ensure that the selected model aligns with their specific use case requirements and resource constraints.

2.3.3 CPU VS GPU DEPLOYMENT

When selecting an LLM, the choice between central processing unit (CPU) and graphics processing unit (GPU) for model deployment is a crucial consideration. GPUs have played a significant role in meeting the computational demands of LLMs, offering parallel processing capabilities that can accelerate model performance. Companies need to evaluate the trade-offs between CPU and GPU utilization based on factors such as performance requirements, model complexity, and available resources. While GPUs can enhance the speed and efficiency of LLM operations, they may entail higher costs and energy consumption. On the other hand, CPUs provide flexibility and cost-effectiveness but may not deliver the same level of performance for large-scale LLM tasks.

Figure 1: Evolution of Large Language Models' parameters¹⁶



¹⁶ Capco research based on model parameters from sources referenced in this paper and <https://tinyurl.com/3jx6xdwh>.
 *There are no published number of parameters available for ChatGPT4; numbers shown for ChatGPT4 are estimates according to <https://tinyurl.com/3vj7kf4r>

2.4 Evolution of Large Language Model releases

An important property of LLMs is the number of learnable elements (parameters) in a neural network, impacting their learning capacity and task performance.¹⁷ The evolution of parameter size in LLMs has seen significant growth over the years and is expected to continue for the foreseeable future (see Figure 1). This trend reflects the shift towards more complex and data-intensive models to achieve superior performance across diverse natural language processing (NLP) tasks. The increasing scale of LLMs is driven by the need for enhanced generalization, multi-modal capabilities, and improved transfer learning effectiveness. Multi-modal capabilities enable a model to comprehend various types of data, while transfer learning measures its ability to apply learned knowledge across different tasks or domains. The ongoing trend towards larger parameter sizes in LLMs underscores the continuous push towards more powerful and versatile models for advanced language understanding and generation tasks.

3. METHODOLOGY FOR COMPARISON AND EVALUATION

In the following section we will outline how we have evaluated the LLMs with respect to their ability in addressing specific tasks, within the context of their performance with COBOL code comprehension. Our evaluation process includes a dataset containing COBOL programs that are part of a COBOL application, which forms the basis for assessing the capabilities of these LLMs across different query types.

3.1 Benchmark dataset and evaluation framework

We evaluated the LLMs against a variety of tasks involving code comprehension, utilizing a diverse range of COBOL code snippets, from straightforward functions to intricate program structures, mirroring real-world scenarios commonly encountered in software development and maintenance. This evaluation was conducted using the same knowledge base for each of the models we tested.

Our evaluation framework incorporates four distinct classes of query types:

1. **Basic queries:** Evaluate the LLMs' understanding of fundamental programming concepts and COBOL code

navigation skills, such as how a function works from a technical standpoint, or where in a large piece of code a particular capability is executed.

2. **Aggregation queries:** Evaluate the LLMs' proficiency in aggregating information from various sections of the codebase, such as generating a comprehensive data dictionary. The data dictionary serves as an example of how well the model can aggregate information effectively. These queries assess the model's ability to extract and organize relevant data elements across different sections of the codebase.
3. **Reverse engineering queries:** Assess the LLMs' ability to comprehend COBOL syntax and turn it into human-interpretable forms, such as user stories, acceptance criteria, or test cases. This evaluation focuses on assessing how effectively the LLMs interpret code semantics and transform technical details into formats that are easily understandable by humans.
4. **Code improvement queries:** Evaluate the model's capability to interpret human input and suggest modifications to the code. For instance, examples include tasks like adding new data validation routines or soliciting insights on areas where code can be strengthened in response to production incidents. This evaluation focuses on assessing the models' capability to provide actionable insights for enhancing code quality and performance while preserving the integrity of the original codebase.

3.2 Benchmarking process

Each query type was submitted to the LLMs, and their responses were compared against human correct answer, i.e., answers provided by expert COBOL developers, which served as the gold standard of knowledge.

Our analysis centered on three key areas:

- Comparative analysis of open-source LLMs' performance across the query types described above
- Identification of strengths and weaknesses of each model for specific task solving
- Factors influencing model performance.

¹⁷ <https://tinyurl.com/6ekdke4a>

4. RESULTS AND CONCLUSIONS

4.1 Key findings from the comparison and evaluation

We have computed the similarity for each query to visualize and compare the performance of each model in solving specific tasks (see Table 5 below).¹⁸ We used a zero-shot approach, meaning each model was evaluated on its first attempt at answering the query. Each value in the matrix represents the cosine similarity score between the LLM responses from the model and the Human Correct Benchmark Answers, with green indicating perfect similarity and purple indicating no similarity.¹⁹ Importantly, similarity values range from 0 to 1 but do not represent accuracy percentages; rather, they indicate the degree of similarity between responses, with higher values indicating greater similarity.

The table below compares the similarity between the responses generated by different models for the four distinct query types described in Section 3.1. compared to the human provided correct answers. While the heatmap indicates some visual variability in performance across models, the overall differences in cosine similarity scores are relatively small, suggesting that most models perform at a high level in aligning with human references.

From this table, we can infer several insights regarding model performance that can guide businesses in selecting the right model for their specific use cases:

Consistency: Models that show relatively high scores across multiple query types tend to offer more consistent performance. For example, ChatGPT4o consistently delivers strong similarity scores across various tasks, making it a reliable option for broad, versatile use. Similarly, Llama3.1_405B and Gemini-pro-1.5 demonstrate steady performance, indicating adaptability across diverse queries.

Specialization: Some models excel in specific areas, making them ideal for focused use cases. For instance, Claude-Sonnet_3.5 ranks highly in the Code Update Query, highlighting its proficiency in code generation tasks. Grok-2 performs exceptionally well in the Aggregation Query, suggesting it may be the best fit for scenarios requiring data aggregation. It's important for customers to evaluate their primary use case towards making a strategic choice pertaining which LLM to select.

Table 5: Similarity between the responses generated by different models and the human-provided correct answers for various query types

MODEL	BASIC QUERY	AGGREGATION QUERY	REVERSE ENGINEERING QUERY	CODE UPDATE QUERY	AVERAGE
ChatGPT4o	High	High	High	High	High
Llama3.1_405B	High	High	High	High	High
Gemini_1.5_pro	High	High	High	High	High
Claude-Sonnet_3.5	High	Low	High	Very High	High
Llama3.1_70B	High	High	High	High	High
Grok_2	High	High	High	High	High
Wizard-LM8x22	High	Low	High	High	High
DeepSeek_v2.5	High	High	High	High	High
ChatGPT4-o1-preview	High	High	High	High	High
Mixtral_8x22B	High	High	High	High	High
Mistral_Large_v2	High	High	Low	High	High
Llama-3.2-1b-instruct	High	High	High	High	High

Perfect similarity
No similarity

¹⁸ To compute the cosine similarity between text blocks, a combined approach was used, integrating both term-frequency-based and semantic-level similarity measures.

¹⁹ <https://tinyurl.com/53jzcvss>

“

Navigating the landscape of Large Language Models demands strategic selection, where the right choice becomes the critical bridge between raw computational potential and transformative organizational intelligence. ”

Variability: Models with a wider range of similarity scores may indicate specialized strengths but could also reflect inconsistent performance. For example, WizardLM-2-8x22B shows variation across queries, excelling in some areas but performing lower in tasks like Aggregation Query, which could indicate a need to match the model with its strengths for optimal results.

Outliers: Models with lower similarity scores in certain queries highlight areas where they may struggle. However, rather than a weakness, this can be an opportunity for businesses to focus on the tasks where these models excel. For example, while Mistral-Large-v2 and Mixtral-8x22B show lower scores in Reverse Engineering, they may still be excellent choices for specific, targeted tasks if aligned with the business's key needs. Moreover, fine-tuning can significantly enhance the capabilities of even smaller models, as seen with some of the latest advances such as Llama-3.2-1b-instruct. Although this model shows lower performance across most query types, it can still be highly effective when used strategically.

4.2 Implications for companies considering the adoption of open-source LLMs for internal data processing

Performance evaluation: It's crucial for companies to conduct a comprehensive assessment of proprietary and open-source LLMs across tasks relevant to their use cases. We have selected cosine similarity as a valuable metric for comparison since it provides quantitative unbiased results.

Specialization consideration: Companies should actively design queries to comprehensively test open-source LLMs based on their specific use cases. This practical approach helps identify which models are most suitable for fulfilling their data processing needs and achieving objectives effectively.

Variability awareness: Companies should assess performance variability across different tasks or queries when adopting open-source LLMs. This includes thorough testing and evaluation of the models' capabilities across various scenarios. This assessment enables them to tailor customization or fine-tune efforts effectively, ensuring optimal performance alignment with their specific use cases.

Cost-benefit analysis: While open-source LLMs offer cost advantages compared to proprietary models, companies must weigh these benefits against potential trade-offs in performance and variability. One practical approach is to create a structured template that evaluates factors such as initial setup costs, ongoing maintenance expenses, potential productivity gains, and the expected impact on data processing efficiency.

4.3 Final thoughts on the significance of selecting the right model for specific use cases

In this article we have described a robust and repeatable method to evaluate Large Language Models across various knowledge domains, facilitating meaningful comparisons between different models.

Our strategy incorporates a flexible approach, enabling us to evaluate LLMs for any given use case. This adaptability allows us to evaluate the models efficiently within meaningful contexts as they get released, keeping pace with the latest advancements.

The framework that we have established is reusable and can be applied to many different use-cases. This structured approach systematically evaluates the costs and benefits associated with each LLM, providing stakeholders with clear insights into the value proposition and helping make informed choices that align with organizational objectives and resource constraints.

Selecting the right model for a specific use case is crucial, as it significantly impacts various aspects such as cost, footprint, and the ability to satisfy regulators. The choice of model also directly influences the performance and effectiveness of the intended use case, ensuring optimal resource allocation and compliance with regulatory standards.

5. GLOSSARY

Prompt: In the context of AI, a prompt is a text input given to a language model, which then generates an output based on the input provided.

Fine-tuning: A process in machine learning where a pre-trained model is further adjusted or 'tuned' on a new, often smaller, and more specific dataset.

Context length: The 'context length' denotes the maximum number of tokens (e.g., words, characters) an AI model can process or analyze at any given time.

Parameters: Parameters are the number of learnable parameters like weights and biases in a neural network.

Large Language Model (LLM): An LLM is a type of neural network. LLMs are typically built using neural network architectures, such as transformer models.

Transformer: The transformer model is a type of neural network architecture introduced by the landmark research paper by Google, "Attention Is All You Need", authored by eight scientists in 2017. This architecture was revolutionary for its use of self-attention mechanisms.

Token: A token refers to the smallest unit of data, usually a subword, that can be processed by the LLM.

© 2024 The Capital Markets Company (UK) Limited. All rights reserved.

This document was produced for information purposes only and is for the exclusive use of the recipient.

This publication has been prepared for general guidance purposes, and is indicative and subject to change. It does not constitute professional advice. You should not act upon the information contained in this publication without obtaining specific professional advice. No representation or warranty (whether express or implied) is given as to the accuracy or completeness of the information contained in this publication and The Capital Markets Company BVBA and its affiliated companies globally (collectively "Capco") does not, to the extent permissible by law, assume any liability or duty of care for any consequences of the acts or omissions of those relying on information contained in this publication, or for any decision taken based upon it.

ABOUT CAPCO

Capco, a Wipro company, is a global management and technology consultancy specializing in driving transformation in the energy and financial services industries. Capco operates at the intersection of business and technology by combining innovative thinking with unrivalled industry knowledge to fast-track digital initiatives for banking and payments, capital markets, wealth and asset management, insurance, and the energy sector. Capco's cutting-edge ingenuity is brought to life through its award-winning Be Yourself At Work culture and diverse talent.

To learn more, visit www.capco.com or follow us on LinkedIn, Instagram, Facebook, and YouTube.

WORLDWIDE OFFICES

APAC

Bengaluru – Electronic City
Bengaluru – Sarjapur Road
Bangkok
Chennai
Gurugram
Hong Kong
Hyderabad
Kuala Lumpur
Mumbai
Pune
Singapore

MIDDLE EAST

Dubai

EUROPE

Berlin
Bratislava
Brussels
Dusseldorf
Edinburgh
Frankfurt
Geneva
Glasgow
London
Milan
Paris
Vienna
Warsaw
Zurich

NORTH AMERICA

Charlotte
Chicago
Dallas
Houston
New York
Orlando
Toronto

SOUTH AMERICA

São Paulo

THIS UNIQUE IMAGE WAS GENERATED USING MID-JOURNEY, STABLE DIFFUSION AND ADOBE FIREFLY

WWW.CAPCO.COM



CAPCO
a wipro company