Market Research Benchmark for Large Language Models

Sean Niklas Semmler

October 20, 2024

1 Introduction

In recent years, the rapid advancement of artificial intelligence, particularly in the domain of natural language processing (NLP), has opened up new possibilities for automating and enhancing various aspects of market research. Large language models (LLMs) have exhibited noteworthy aptitudes in comprehending and producing text that is akin to that of humans, rendering them potentially efficacious instruments for examining market trends, competitive landscapes, and consumer conduct. Nevertheless, the efficacy of these models in particular market research tasks remains a topic of active investigation. The objective of this study is to benchmark the performance of five language models in the context of market research applications. The performance of these models, namely $phi-2$, $bloomz-1b7$, $stablelm-2-1\_6b$, $tinyllama-1.1b-chat$, and $opt-1.3b$, is evaluated across a range of tasks designed to mirror real-world market research scenarios. The benchmark includes questions on trend analysis, competitive landscape assessment, pricing strategies, and market size estimation, thereby enabling an assessment of the models' capabilities in various aspects of market intelligence gathering and analysis.

2 Benchmark Questions

Our benchmark focused on five key questions relevant to market research:

1. **Trend Analysis in AI Funding:** "List the top 5 AI companies globally with the highest funding in 2023."
2. **Competitive Landscape in Cloud Services:** "Identify the top 3 cloud service providers by market share in 2023 and provide their market share percentages."
3. **Trend Analysis in Data Science:** "List the top 5 trending libraries in the Data Science market as of Q4 2023, along with their primary use cases and growth rates over the past year."
4. **Competitive Pricing Analysis:** "What is the pricing range for JetBrains PyCharm? Which pricing package would you recommend a student?"
5. **Market Size Estimation:** "Using the historical growth rate of the global AI market from 2020 to 2023, estimate the market size for 2025. Provide your reasoning and state any assumptions."

The objective of these questions was to evaluate the models' capabilities in various aspects of market research, including trend identification, competitive analysis, pricing strategies, and market forecasting. The sequence of questions was designed to progress from relatively straightforward listing questions with increasing complexity in the required answer length (Question 1-3), to two questions that required the models to demonstrate their ability to reason and draw conclusions (Question 4-5).

3 Benchmark Setup

3.1 Language Models

For this benchmark, we selected five language models of varying sizes:

- **phi-2** [1]: A 2.7B parameter model developed by Microsoft, released in December 2023. It represents a recent advancement in smaller-scale language models with impressive capabilities.
- **bloomz-1b7** [2]: Part of the BLOOM family, this 1.7B parameter model was published in November 2022. It's trained on a diverse multilingual dataset, aiming for broad applicability across languages and tasks.
- **stablelm-2-1_6b** [3]: Stability AI's latest small-scale model with 1.6B parameters, released in January 2024. It's trained on a diverse dataset for general-purpose use, balancing performance and efficiency.
- **tinyllama-1.1b-chat** [4]: A compact 1.1B parameter model optimized for chat-like interactions, released in January 2024. It's designed to emulate larger language models while maintaining efficiency.
- **opt-1.3b** [5] [1]: Part of Facebook's OPT (Open Pre-trained Transformer) series, this 1.3B parameter model was released in May 2022. It represents an early effort in open-source language models.

The models were selected for comparison of the capabilities of various artificial intelligence (AI) companies and open-source models. Unfortunately, access to a Google model was not readily available. To assess potential progress over time, a range of publication dates was employed. Additionally, models with comparable sizes (between 1 and 3 billion parameters due to hardware constraints) were chosen to facilitate comparison of performance across different designs and training approaches.

3.2 Evaluation Metrics

To comprehensively assess the performance of the language models in our market research tasks, we employed the following evaluation metrics:

[1] https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-model
[2] https://huggingface.co/bigscience/bloomz-1b7
[3] https://huggingface.co/stabilityai/stablelm-2-1_6b
[4] https://github.com/jzhang38/TinyLlama
[5] https://huggingface.co/facebook/opt-1.3b

Accuracy (Acc): Measures the correctness of the information provided by the model. A high accuracy score indicates that the model's responses align closely with verified facts and data.
Relevance (Rel): Assesses how well the model's response addresses the specific question asked. High relevance scores suggest that the model understands the query and provides pertinent information.
Completeness (Comp): Evaluates the thoroughness of the model's response. A complete answer covers all aspects of the question without significant omissions.
Coherence (Coh): This measure assesses the model's capacity to generate responses that are fluent, natural, and consistent with human language.
Reasoning (Reas): The evaluation of the model's capacity to draw logical conclusions, make inferences and provide well-structured arguments is of particular importance in the context of questions that require analysis or recommendations. Furthermore, the evaluation of the linguistic complexity of the answer is also a key aspect of the assessment.
Response Time (RT): Measures the time taken by the model to generate its response. This metric is crucial for understanding the model's efficiency, especially in time-sensitive market research scenarios.

Each metric was scored on a scale of 1 to 5, with 5 being the best performance, except for Response Time, which is measured in seconds. These metrics were chosen to provide a holistic view of each model's performance, balancing the quality of information provided (Accuracy, Relevance, Completeness) with the reliability of the output (Coherence), the depth of analysis (Reasoning), and the efficiency of the model (Response Time). This multi-faceted evaluation approach allows for a nuanced understanding of each model's strengths and weaknesses in the context of market research tasks.

3.3 Testing Environment

The benchmark was conducted on an Intel® Core™ i7-9700K processor with 8 cores, 16GB RAM, and a 64-bit architecture. The experiment was conducted using a NVIDIA GeForce RTX 2060 SUPER graphics card with 8 GB of VRAM and the CUDA 12.1 software development kit. The operating system in use was Windows 10 Home, version 22H2. Python version 3.12 was utilised, and time measurement was performed using the Python time.time() function. All language models were loaded from the Hugging Face Model Hub using the Transformers library. This approach ensured consistent access to the most up-to-date versions of each model at the time of the experiment.

4 Results

4.1 Performance Comparison

Acc: Accuracy, Rel: Relevance, Comp: Completeness, Coh: Coherence, Reas: Reasoning, RT: Average Runtime

4.2 Analysis of Results

Our benchmark evaluated five language models ($phi-2$, $bloomz-1b7$, $stablelm-2-1\_6b$, $tinyllama-1.1b-chat$, and $opt-1.3b$) across five market research questions, assessing

Table 1: Evaluation Matrix for Question 1: Top 5 AI Companies

Model	Acc	Rel	Comp	Coh	Reas	RT (s)
phi-2	4	5	4	4	3	9.23
bloomz-1b7	3	4	3	4	1	1.49
stablelm-2-1_6b	3	4	4	4	2	13.16
tinyllama-1.1b-chat	2	2	4	2	2	24.43
opt-1.3b	1	1	1	1	1	104.94

Table 2: Evaluation Matrix for Question 2: Top 3 Cloud Service Providers

Model	Acc	Rel	Comp	Coh	Reas	RT (s)
phi-2	4	5	5	4	5	9.65
bloomz-1b7	3	4	3	4	5	0.96
stablelm-2-1_6b	2	2	2	2	1	112.65
tinyllama-1.1b-chat	4	5	5	5	5	25.87
opt-1.3b	2	3	2	2	2	101.71

their performance based on accuracy, relevance, completeness, coherence, reasoning, and response time. The $phi-2$ model emerged as the standout performer, demonstrating consistent excellence across all questions. It particularly excelled in trend analysis and competitive landscape assessments, showcasing high accuracy, relevance, and completeness in its responses. However, this comprehensive performance came at the cost of slower response times, especially for more complex queries. In contrast, $bloomz-1b7$ proved to be the speed champion, consistently delivering the fastest responses across all questions. This makes it an attractive option for time-sensitive applications. However, its performance on other metrics was inconsistent, with notable struggles in addressing the data science libraries question. The $stablelm-2-1\_6b$ model showed moderate performance across most metrics. While it performed admirably in the market size estimation question, it faltered when addressing the cloud service providers query. Like phi-2, it exhibited slower response times, particularly for complex questions. $Tinyllama-1.1b-chat$ demonstrated exceptional performance in specific domains, particularly excelling in the cloud service providers and data science libraries questions. It maintained solid performance across other questions with reasonable response times, indicating its potential as a balanced option for certain market research tasks. Lastly, $opt-1.3b$ consistently underperformed, showing the weakest overall results across most questions and metrics, coupled with the slowest response times. Our question set provided valuable insights into the models' capabilities. The AI companies funding question revealed varying levels of up-to-date knowledge among the models. The cloud service providers question highlighted their ability to provide specific market share data. The data science libraries question tested their awareness of current trends and capacity for detailed information. The PyCharm pricing question assessed their capability to provide accurate product information and make recommendations. Finally, the market size estimation question evaluated their analytical and reasoning capabilities.

Table 3: Evaluation Matrix for Question 3: Top 5 Trending Libraries in Data Science

Model	Acc	Rel	Comp	Coh	Reas	RT (s)
phi-2	4	5	5	4	5	45.21
bloomz-1b7	1	1	1	1	1	0.41
stablelm-2-1_6b	3	4	5	3	3	89.47
tinyllama-1.1b-chat	4	5	5	4	5	62.07
opt-1.3b	2	2	3	3	2	101.81

Table 4: Evaluation Matrix for Question 4: PyCharm Pricing

Model	Acc	Rel	Comp	Coh	Reas	RT (s)
phi-2	3	4	3	3	2	91.56
bloomz-1b7	2	2	2	4	1	0.27
stablelm-2-1_6b	3	3	3	3	2	113.03
tinyllama-1.1b-chat	4	4	4	3	5	29.64
opt-1.3b	4	4	4	2	2	100.78

5 Conclusion

Based on our comprehensive benchmark, we can draw several key conclusions about the performance and potential applications of these language models in market research tasks. $Phi-2$ stands out as the most versatile and reliable model, making it an excellent choice for a wide range of market research tasks where accuracy and completeness take precedence over speed. Its consistent performance across various question types demonstrates its potential as a go-to solution for in-depth market analysis. $Bloomz-1b7$, with its unparalleled speed, presents itself as an ideal candidate for applications requiring rapid responses. This makes it particularly suitable for preliminary research or time-sensitive queries. However, its inconsistent accuracy suggests that its outputs may need verification for critical tasks, positioning it as a valuable tool for initial, fast-paced market explorations. $Tinyllama-1.1b-chat$ shows promise as a lightweight alternative to phi-2. Its strong performance in specific domains, coupled with reasonable response times, makes it a valuable option for targeted market research tasks. This model could be particularly useful in scenarios where a balance between accuracy and speed is crucial. $Stablelm-2-1\_6b$ and $opt-1.3b$, while showing potential in certain areas, may require further fine-tuning or updates to compete effectively in market research applications. Their current performance suggests that they might be better suited for specific, niche tasks rather than general market research queries. For future market research applications, we recommend a strategic approach to model selection. $Phi-2$ should be the primary choice for tasks demanding high accuracy and comprehensive responses. $Bloomz-1b7$ is ideal for quick, preliminary research or time-sensitive queries where rapid insights are crucial. $Tinyllama-1.1b-chat$ offers a compelling middle ground, serving as a lightweight alternative to $phi-2$ for scenarios where both accuracy and efficiency are important. In conclusion, this benchmark highlights the diverse strengths of different language models in the context of market research. By carefully selecting the appropriate model based on the specific requirements of each task – be it depth of analysis, speed of response, or domain-specific knowledge – researchers and analysts can significantly

Table 5: Evaluation Matrix for Question 5: Global AI Market Size Estimation

Model	Acc	Rel	Comp	Coh	Reas	RT (s)
phi-2	4	5	5	3	4	93.50
bloomz-1b7	2	2	1	1	1	0.56
stablelm-2-1_6b	3	4	5	4	5	30.05
tinyllama-1.1b-chat	3	4	5	3	4	58.62
opt-1.3b	3	3	4	3	3	102.25

Table 6: Overall Performance Matrix

Model	Acc	Rel	Comp	Coh	Reas	RT (s)
phi-2	4	5	4	4	4	49.83
bloomz-1b7	2	3	2	2	1	0.74
stablelm-2-1_6b	3	4	3	3	2	71.67
tinyllama-1.1b-chat	3	4	5	3	4	40.13
opt-1.3b	3	3	3	2	2	102.30

enhance the efficiency and effectiveness of their market research processes.

References

[1] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.