Original Research Article

Evaluating Generative AI for Startups: A Benchmarking Study of Large Language Models

Shaina Gupta

The Overlake School, 20301 NE 108th St, Redmond, WA 98053, United Stetes

ABSTRACT

Startups are critical for global economic development, and they are often constrained by resources and affiliations, limiting their growth, unlike larger companies. Concerningly, nearly 70% of startups fail 2-5 years after their launch. The way to eliminate this discrepancy is to leverage commercial, readily-available AI tools to assist with tedious tasks, so human capital and financial resources could be better spent on business development. This paper intends to use prompt engineering principles and evaluation rubrics that assess appropriate AI tools for different necessities of startups to answer the following question: How do different large language models (LLMs) vary in their prompt responses in AI-driven solutions for startups, and what implications does this have for selecting generative AI tools in small business contexts? An exclusive public dataset of prompts and commercial-LLM responses is being released that can be used by startups to evaluate the effectiveness of integrating AI tools for specific business activities. This dataset can be leveraged by startups of all types, to baseline the selection of AI tools, allowing them to allocate resources to more meaningful aspects of a business. This is being done for three business cases, including web design, market research, and business support. Each of these business cases have several prompts which are evaluated with 2-3 different LLMs to determine the optimal LLMs for different use cases. The key findings were that for certain use cases like web design, general usage LLMs like ChatGPT 5.0 produced optimal results, but in contrast, for other use cases like market research and business analysis, the specialized LLMs that provided lots of research performed better, like Claude. Therefore, the quality of results based on LLMs is on a case by case basis, but it can be extrapolated to the majority of prompts under the jurisdictions of those three use cases.

Keywords: Startups; Large Language Models; Prompt Engineering; AI Evaluation; Generative AI; Prompt Dataset

INTRODUCTION

Startups make up a smaller share of U.S. businesses than often assumed — about 1 million startups currently operate in the United States, employing roughly 5 million people (about 4% of the private-sector workforce) (1, 2). Additionally, in terms of global funding, there are statistics on the $368 billion global / $221 billion U.S. split (3). Startups face unique hardships in their early phases, including fewer employees and limited resources. The *Wall Street Journal* reported that the average new U.S. business launched during the pandemic with about 4.6 employees, down from 5.3 pre-pandemic (2, 4). Founding teams typically consist of

Corresponding author: Shaina Gupta, E-mail: us.shaina@gmail.com.

Accepted October 29, 2025
https://doi.org/10.70251/HYJR2348.36223233

2–4 people (5, 3). Survival rates are equally challenging — multiple sources confirm that about 90% of startups fail, with 10% failing in the first year and about 70% failing between years 2 and 5 (6, 7, 8). These numbers and proportions emphasize the significance of startups within the US and thus the importance of increasing the success rates of startups. As AI is being leveraged by businesses of all kinds, it is imperative that startups learn how to optimize AI for their greater benefit, in order to allow them a better chance to grow (12).

One of the defining challenges for startups is operating under severe resource constraints including limited capital, small teams, and reduced access to infrastructure compared to established firms. Early-stage companies often rely on bootstrapping or small seed investments, which forces them to prioritize short-term survival over long-term growth (10, 13). This lack of capital limits their ability to hire experienced talent, invest in advanced technology, or scale marketing campaigns at a competitive pace (7). The lack of team members means founders and early hires wear multiple hats: handling operations, sales, and product development simultaneously (4, 13).

Even when funding is secured, startups face cash flow irregularities due to irregular revenue streams and the need to reinvest heavily in product development (2, 13). Infrastructure constraints, such as lack of established supply chains, customer networks, or legal support, further slow their ability to compete against larger companies (4, 14). In tech-focused sectors, limited resources also mean they often must outsource key functions like manufacturing or software development, which can increase dependency and operational risk (3, 7).

LLM benchmarking serves to be a novel contribution to startups because it allows startups to see exactly how LLMs perform on the tasks that matter to their product, whether it is coding or customer support (9, 10). Startups are no longer forced to "trust the marketing", but rather can see data to prove which LLM is most optimal for their needs (9, 15). Additionally, benchmarking serves to help startups save money, because some models may give similar results with a less expensive version compared to others (10, 12). Using readily available commercial AI tools allows startups to allocate their resources more appropriately, such as toward product development (9-12). However, LLM benchmarking remains an emerging area, as many companies have yet to recognize its full potential or dedicate sufficient time to it due to other competing priorities, leading to a gap in the depth of LLM benchmarking.Thus, the purpose of this paper is to provide tools for LLM benchmarking that can be used by startups to learn more about which LLMs are the most optimal for their needs.

All AI tools assessed will be either free or offered at minimal cost to the company to accommodate limited budgets (19). Different AI tools will be evaluated for startup areas that can be improved by artificial intelligence, including web design, business support, and market research (11, 12, 14). Prompts for each of these cases will be assessed through rubrics emphasizing the effectiveness of the different tools (23). On a larger scale, the study will evaluate the effectiveness of various AI models within the context of startups (8, 15). By publicly releasing a dataset designed with prompt engineering principles that can be used by startups to evaluate the integration of certain AI tools into specific business activities, the project provides a free resource and manual for achieving optimal results (8, 9).

LITERATURE REVIEW

Large language models

LLMs (large language models) are AI models that have been trained on vast and diverse text corpus, enabling them to generate and process human-like language (1, 22). They are highly adept co-creators when paired with multimodal capabilities like generating code or images. Startups are particularly well-poised to benefit from LLMs compared to other organizations, due in part to their versatility as tools and due in part to very favorable cost-to-value ratio, which enables them to allow early stage teams to expedite product design, marketing and product automation without a full technical organization (8, 11, 12). There are various LLMs that are specialized or general purpose models. General purpose models like ChatGPT 5.0 are versed in all and can be used for more than just one use case. It can be optimized for at least two of the three aforementioned cases. In the circumstance of a specialized model, it specializes for one case, and provides the most optimal results in those circumstances. One instance of this that will be discussed is Windsurf, a platform specifically used for vibe coding and usage of SWE and Claude, types of LLM models, to prompt AI to change code or provide the steps to change code.

The unique requirements of the startups

As Investopedia puts it, "The term 'startup' refers to a company in the early stages of its operations" (2). Startups differ from established businesses, not

just in size but also in how they have to operate under uncertainty. Startups must constantly improve their product and adapt their strategies to gain customers and attract investment. Since startups are essentially early stage businesses, they often have lesser resources and need to design from scratch, from their unique branding and website, to marketing and often develop a new product or a new version of something (5, 24). General businesses often inherit and operate with established products with more stable demand for products (4).

Since startups develop both their products and their strategies from scratch, decision making is much faster and more experimental. While this agility is beneficial for innovation, it undoubtedly increases the risk of burnout and inefficiency without proper systems in place for support.

For startups, AI and LLMs are critical to improvement because they can automate work like market research, content generation, and customer support without requiring a large amount of human resources (9-12). General businesses have less of a need to hustle for money since they are more stable in their ROIs and have lower dependence on investors (2, 13).

Related work

Existing studies shown in Figure 1 below like the one by Impink from Boston University and another by Martins clarify how startups are utilizing Artificial Intelligence tools in areas such as automation understandings (11), customer-service (11), product development (11), and design (3), however, most studies are descriptive and observational. Some describe the classifications and advantages of generative AI in an enterprise context but do not systematically evaluate and compare LLM performance in startup use cases (10). A taxonomy of related work is developed in Table 1.

This work extends preexisting research by introducing a scientific and empirical method evaluating how LLMs (e.g., GPT-5.0, Claude, Perplexity) react to prompts pertinent to start-ups in a variety of industries and roles (16, 17, 20). The core unique contributions of this research are twofold. First, a set of prompts and a response evaluation matrix is developed. This prompt dataset is developed based on prompt engineering principles, including zero-shot, few-shot, and role-based prompts in various temperature conditions (and therefore varying specificity).

The evaluation matrix established does not merely clarify if LLMs can help; it evaluates how well they help with specific consideration to the type of business function, specificity of the prompt, and the specific properties and optimizations of the AI model being used. Additionally, an evaluation matrix that can be

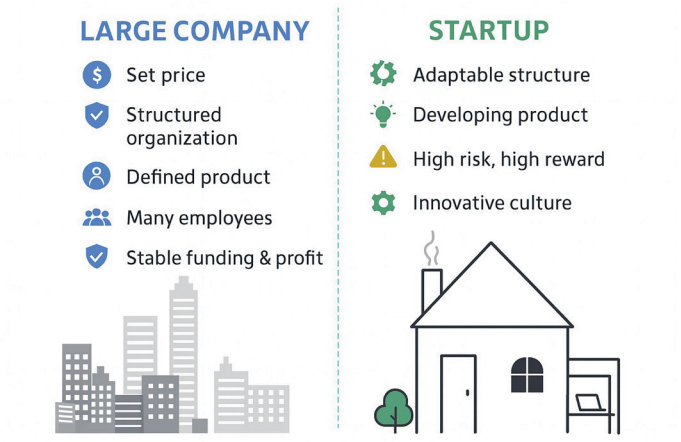


Figure 1. Comparative Overview of Large Companies and Startups, Highlighting culture, *disadvantages, and advantages of startups vs larger companies.*

Table 1. Related works highlighting the comprehensive coverage of this research in comparison to other related works

	Business use cases			AI-use in startups	Prompt engineering-based verification	Dataset release
	Business Support	Web design	Market research			
Lively23 [22]	X	✓	X	X	X	X
Martins24 [24]	X	X	X	✓	X	X
Ahlgren25 [9]	✓	X	X	✓	✓	X
Impink24 [3]	X	X	X	✓	X	X
Linkon23 [10]	✓	X	X	X	X	X
Our work	✓	✓	✓	✓	✓	✓

used to evaluate the effectiveness of specific LLM tools in startup use cases, and broadly in any context was developed (3, 10).

Secondly, this work will create a publicly accessible dataset of prompt responses that will build evidence to guide startup leaders, researchers, and developers in selecting AI tools according to real needs. This work adds evidence, direct and actionable comparison amid many participants who are only generally contributing to an open conversation. This work will help decision making for startup stakeholders using structured experimentation and the direct findings generated from generative AI model output (2, 4, 7).

METHODS AND MATERIALS

Problem

In the status quo, some of the most common concerns of startups include funding, specifically because of a lack of successful marketing (2, 13). Additionally, not having enough resources like the right people to help with the more intricate business details such as animation, often leading to smaller teams where each individual puts in lots of work, leading to burnout (4, 5). Additionally, retaining customers can be hard especially in competitive business environments (6, 7). Figure 1 below contrasts the stark differences in the nature of a startup and a larger business, ultimately highlighting the adaptability and lack of structure in startups, which contrasts the structured organization and great amounts of resources (3, 14) It also emphasizes the advantages and disadvantages of startup culture versus large company culture.

Thus the question was posed: How do different large language models (LLMs) vary in their prompt responses in AI-driven solutions for startups, and what implications does this have for selecting generative AI tools in small business contexts? This research will critically show how certain prompts react in different LLMs and ultimately provide a jumping-off point for small businesses to leverage AI for their benefit.

Business use cases

Startups require strong foundations to succeed in the demanding status quo. The most critical factors include the following; ***web design and development*** which highlights having a strong online presence, as people have transitioned to the online world from the physical one; ***market research*** which allows startups to understand their competition and make educated decisions to enforce a stronger business and at a higher-level ensure success; ***business analysis/operations support*** which is essential for the smooth running and overall success of the business. Each of these use cases are different in their own way and focus on different parts of a business which are critical to a smooth business foundation. The different aspects of a business foundation allow it to flourish as well as be more compelling to users and clients, thereby making it more usable, successful, and ultimately increasing the chances of survival.

Web design and development

As the world transitions to an online setting with increasing interest in online platforms and social media, having a compelling and interesting webpage for any product or startup is important for countless reasons (24). One reason is that businesses with webpages essentially offer a place to learn more about the product and gain traction. However, the design of the platform and the aesthetics also impact the user's choice of product. For example, if Business A has a better online platform, i.e., a compelling web page with good web design, it will likely become more popular than Business B which has an average online platform and an average web design, assuming similar product and similar customer focus/target (14). Thus, it is clear that the success of the business is heavily reliant on the web design, making it critical to assess the web design and use prompt generation to come up with responses that create the most optimal web design that can be extrapolated to create stronger businesses (12, 14). Additionally, most established companies already have IT departments overseeing this field, and have some form of website. In fact, they likely have a well-established website that gains lots of traction, unlike startups (4). Startups begin from scratch with no website and limited human resources, that is no or minimal personnel solely dedicated to helping with establishing a website (2, 13).

Market research

Another critical pillar of startup success is market research, which provides the evidence base for decision-making (2, 13). Startups operate under high uncertainty, and without reliable insights into customer needs, competitor actions, and broader market dynamics, they risk making blind investments or pursuing strategies misaligned with actual demand (6, 7). Market research equips startups to answer questions about the ideal

customer base, biggest solvable pain points, competitor positioning, and how receptive markets are to certain tasks (3). Effective market research allows startups to mitigate risk, uncover opportunities, and refine product-market fit (2, 4). For example, a new food-tech startup may discover through research that its target demographic values sustainability over price, leading to strategic adjustments in packaging, pricing, and messaging.

Big companies already contain established products and customer bases, with existing suppliers and reasonable pricing, but startups need to figure this out from the beginning till the last bit, with minimal help from human resources (4). By using LLMs for market research use cases, startups can quickly generate survey questions, identify key market trends, or draft competitive analyses tailored to their industry (10-12). This not only saves time but also expands the scope of insights available, giving startups a systematic way to integrate external knowledge into strategic planning (8, 9). Ultimately, strong market research forms the backbone of educated decisions and increases the likelihood of sustainable success (2, 13).

Operations support

While web design and market research focus on how a startup presents itself externally, the role operations support centers on introspection and internal optimization. A startup's excellent product and compelling design may fail if it cannot manage its operations, allocate resources efficiently, or identify weaknesses before they become critical issues (2, 13). Operations support ensures that startups are not only market-ready but also internally resilient (4). The operations support function provides structured frameworks for evaluating performance across key areas and alignment between strategy and execution (13). By systematically analyzing factors, startups gain a clearer picture of where they are excelling and where they need corrective action. For instance, an analyst might find that while customer acquisition is strong, retention rates are lagging due to insufficient onboarding support, a fixable weakness that could significantly improve growth (6, 7). Additionally, operations support can help the startup with what it needs in the background to succeed like lawyers, human resources and management, along with admin and marketing, most of which can be offloaded onto AI to alleviate some of the pressure on these startups in the initial phases (10, 12).

In larger companies, there are entire departments dedicated to operations support, but in startups, it is handled by maybe 1–2 people who are also handling other business aspects as well (13, 25). LLMs in this context can help simulate internal evaluations, generate diagnostic questions, and propose improvement strategies (8, 9). By applying LLM-driven business analysis, startups can uncover hidden inefficiencies, design KPIs, and identify action plans more quickly than through traditional manual review alone (8, 12, 15). This strengthens their ability to adapt to challenges and maximize their limited resources, ultimately improving survival odds (10, 15).

Dataset creation

A dataset was created to make it easier for startups to evaluate LLMs for different use cases. The primary thing the dataset emphasizes is the benefit of using certain LLMs over others, but it also provides helpful user prompts/serves as a prompt library that can be used for different startup cases. They wouldn't need to spend time, money, or energy on writing prompts to evaluate different AI tools, instead they can use the prompts developed in the dataset based on prompt engineering principles, and evaluate as many LLMs they would like to, so that they can select the ones they find useful (9). It also shows common prompting patterns that work well. This dataset can guide tool selection and serve as a benchmark for different LLMs in different use cases, and essentially help with choosing the most optimal LLM.

In order to determine the most optimal LLMs for each of these use cases, three different LLMs were assessed for each of the three cases with 30 different prompts. Each of the 75 total prompts across all use cases was optimized through the use of Perplexity (17) and ChatGPT 5.0 (16) to ensure consistency amongst prompts, where each prompt was slightly different and multiple topics were covered, each with different supporting details. Additionally for each use case, the 25 established prompts each used different concepts of prompt engineering to see if there was significant difference in the effectiveness of the prompts (9).

Each prompt is scored using an AI rubric scorer, Agentic AI (23). This scorecard gave a score out of 75 for each prompt, where a max of five points could be given for each of the 15 criteria seen on the rubric, which include clarity, role usage, desired output format, length, and audience specification to list a few. From there, once the prompts scored a minimum of 50/75 on the score card, they were run in the three LLMs for their

business case, one of which was always ChatGPT 5.0 (16). The number 50/75 was chosen because a threshold was needed, and getting a minimum 80 percent of all possible points seemed like a reasonable goal for each prompt.

Prompt engineering techniques used

The main prompt engineering techniques used include the following:

- Few shot responses essentially guide the LLM and produce a more predictable response because they provide the model with several examples to guide its responses (9).
- Zero shot responses are less predictable in that there are no examples or bounds given, along with no explicit instructions on how to complete the task or answer the question (9).
- Low temp responses are essentially predictable and factually consistent because they generate more focused, deterministic responses by lowering randomness (23).
- High temp responses essentially produce more diverse, creative responses by increasing randomness (23).
- Few shots use the concept of low temp, while zero shots use the concept of high temp (9).
- Role prompting, the last of the five, essentially involves the AI adopting a specific persona, role, or expertise to help focus their responses, shape style and perspective and guide them. This puts responses into the desired context and makes them more relevant (9, 12).

Prompt generation

In order to understand the optimal LLMs for certain business cases, it is critical that there are some baseline prompts that can be used to evaluate the effectiveness of the LLMs (9). This means having prompts that heed by one of the three business cases and has a certain level of quality to ensure commonalities between the prompt qualities. For this purpose, an online prompt scorer was utilized that scores the quality of the prompt given on a scale of 1 to 5 for 15 criteria (23). A score closer to 75 emphasizes the sheer quality of the prompt against the 15 criterion. For these purposes, a prompt needs to be anywhere between 50 and 75 to keep a certain level of similarity between the both (23). Once 30 prompts for each of the three cases are generated, each with varying levels of prompting characteristics like few shots and temperature, and each prompt satisfies the 50–75 rule, the prompts will then be run in the corresponding LLMs. Figure 2 highlights the process of prompt engineering and response development, highlighting the steps that were taken to minimize human bias within the research (9, 23). It also shows the process after the prompt reaches the score threshold, which will be detailed further.

Dataset structure and dissemination

The data is stored as a csv file that could be downloaded at https://docs.google.com/spreadsheets/d/e/2PACX-1vQq5C-1Z6UTP1ziJY7YPXLMPClhG8J90SCzEl6Bsckc_ry3QzL93KUs35Bta_iPWw-bflrbe4-hIwa_/pub?output=csv. Once opened, each of the different columns would be visible, including the prompt and its grading, the 225 responses across the 75 total prompts, and the scores of each of the 225 responses. Each of the prompts is placed in the proper column based on the LLM the response is being evaluated in which depends on use case. Table 2 below shows an example of what the dataset would look like with one prompt for each use case. See example below.

Experimental methodology

For these purposes, ChatGPT 5.0 was used for all three use cases since it is quite generic: Claude and Perplexity for market research: Jasper and Writer for business

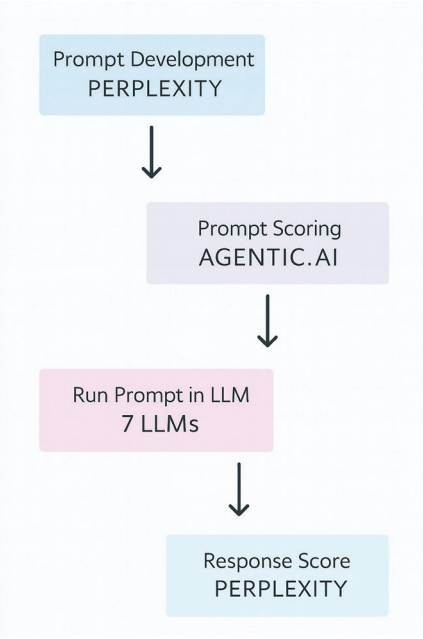


Figure 2. Prompt Development and Scoring Workflow for LLM *Evaluation.*

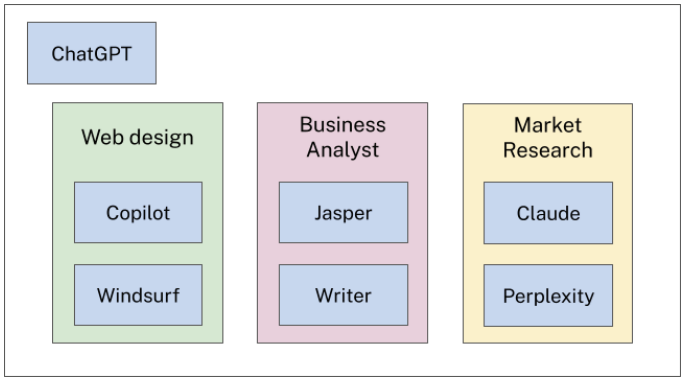
support; and Copilot and Windsurf for web design. Table 3 gives a visual representation of the three use cases and the LLMs used for each one. Note that all three use cases lie within ChatGPT 5.0 because it is universally shared amongst all three use cases, though different prompts are run through ChatGPT 5.0 for each use case.

Response evaluation

Once responses are generated, each response will be evaluated for quality on a rubric developed, which specifically categorizes five important criteria which are then scored. Then, the responses from the three LLMs for each response will be compared to determine the most optimal LLM for that prompt. By repeating that step countless times for the rest of the prompts, the pattern of the most optimal LLMs will be evident. It is critical to note that there is a certain level of bias since the rubric was created by humans and the prompt responses were also evaluated by humans. Thus, it was critical that LLM models were being used to evaluate LLM responses to minimize scoring bias from humans (16, 17).

Each of the responses was then evaluated against a rubric developed for certain criteria. Once the 25 responses from each category were evaluated against each of their three LLMs, the scores of the 25 responses in each of the three LLMs were averaged to determine the LLM with the highest average score, which would be considered the best LLM for that use case. In order to ensure consistency amongst grading, Perplexity was used to grade the responses since humans have inherent bias (17). In Table 4, a rubric was provided as scoring guidelines (see below) and used a prompt format (see below) to score all prompts and decrease bias.

Table 3. The Different LLMs Evaluated Within Each of the Three Use Cases, With One Common LLM of ChatGPT



Response Scoring **Prompt:**
My question is <prompt>.
Can you score the response below. Be extremely critical.
<response>
My scoring rubric is;
<rubric>

RESULTS

After developing and testing all the prompts, receiving the responses, and scoring them, the optimal LLMs based on the provided data was able to be determined. Figure 3 is a heatmap which essentially shows for each use case which LLMs reach each criterion the best. The color that corresponds to the highest score of 3, based on this rubric for each criteria, indicates that the particular LLM is most optimal under that criterion.

Figure 4 essentially highlights that the LLM from a use case with the highest average score amongst all criteria is known as the most optimal LLM for that particular use case. Essentially, Figure 4 displays all the averages for the 7 LLMs, but divides it into the three use cases so conclusions based on the LLMs are only being drawn with the ones under the same use case.

DISCUSSION

As a whole, the results were analyzed firstly from the angle of the different use cases before making some generalizations about all the LLMs.

Table 2. Dataset Structure With Prompts and Responses from Respective LLMs

prompt	category	prompt_engi	score	res_chatgpt	score_res_ch	res_perplexit	score_res_pe	res_claude	score_res_cl	res_jasper	score_res_ja	res_writer	score_res_wi	res_windsurf	score_res_wi	res_copilot	score_res_copilot
Act as a busir	business ana	Role playing	52	Perfect,	2:2:3:2:2					Starting a	2:3:2:2:2	Evaluating	2:3:3:2:3				
List and brief	market resea	zero shot	53	Below is a,	3:3:2:2:2	Here is a,	3:3:2:2:2	Market	2:3:3:3:2								
In the same	web design	few shot	57	A tiered price	3:3:3:3:2									The pricing p	3:3:3:3:3	A tiered price	3:3:2:3:2

Note that the cell values are shortened due to formatting.

Table 4. Prompt Scoring Rubric with 5 Criteria and 3 Possible Dimensions for Each One, With Questions on the Side of each Criteria to Help in Evaluation

High-level description	**High school-level:** not too specific, off topic, grammar mistakes, not very content heavy	**College-Level:** What a high schooler would've produced, informed, nothing wrong with it, just not exceptional	**PhD- Level Expert:** very specific, content-heavy, on-topic, some great ideas/angles/areas of concern that a regular person might not even think of (exceptional)
Dimension	1 (Poor)	2 (OK)	3 (Excellent)
Relevance: How much the response relates to the topic	Slightly off-topic	Partially on-topic	Directly answers the prompt
Accuracy: How true the information is in the response	Many errors (>2)	Some minor errors (1 or 2)	Factually correct (0)
Clarity: How easy to follow is the response (structure of response, summary)	Hard to follow	Structured, but lacking	Structured in detail, but also summarized
Completeness: Depth covered	Misses key points	Covers most points	Covers everything thoroughly
Creativity: Novel ideas, angles considered	Very generic	A bit original	Fresh and engaging

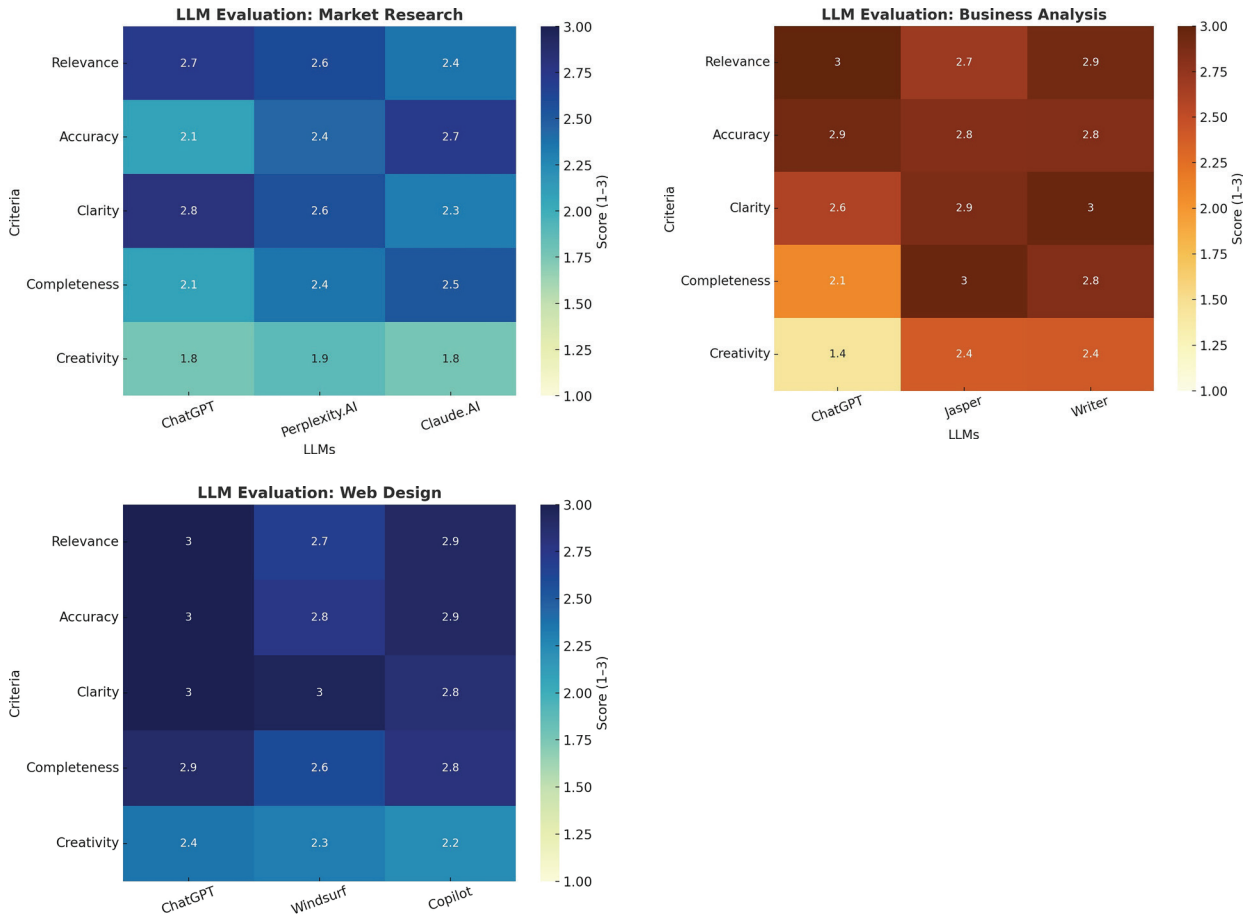


Figure 3. Heatmap of 3 LLMs with Their Average Score across 5 Criteria in Market Research, *Business Analysis, and Web Design.*

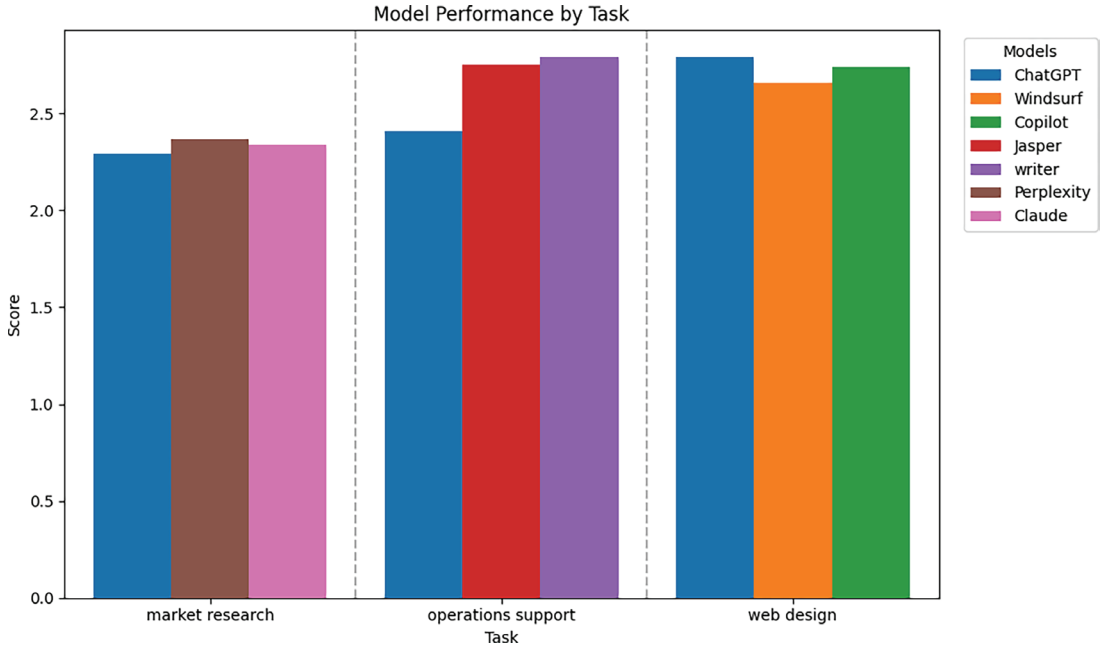


Figure 4. Bar Graph of Average Scores Across All 5 Criteria for the 7 Evaluated LLMs, where *ChatGPT is Common in All 3.*

For market research, it was noted that for relevance (2.7) and clarity (2.8), ChatGPT 5.0 was the best LLM (18) but for accuracy (2.7) and completeness (2.5), Claude.AI was best (20). For creativity (1.9), the best LLM was Perplexity.AI (17). From this, it is clear that for different criteria, different LLMs were better. However, in order to evaluate the effectiveness of the LLM, it was critical to look at the average performance indicated by all five criteria. Figure 4 shows this, and it is clear that all three LLMs are very similar, with a slight edge to Perplexity.AI which has an average score of roughly 2.35, which is very research heavy (17).

For business analysis, ChatGPT 5.0 was best for relevance (3) and accuracy (2.9) (18), while Writer.AI was best in clarity (3) (27) and tied with Jasper.AI for creativity (2.4) (21). Jasper.AI was best for completeness (3) (21). As a whole, the best LLM for business analysis/operations support based on Figure 4 is Writer.AI (26) by a slight difference with Jasper.AI (21), which is a close second. ChatGPT 5.0 is significantly lesser than both by roughly 0.2 - 0.3 (16). Writer.AI's overall score was roughly 2.8 (27).

For web design, ChatGPT 5.0 was best for relevance (3) and accuracy (3), completeness (2.9), and creativity (2.4) (16), while clarity (3) was best in both ChatGPT 5.0 (16) and WindSurf (27). Copilot was not optimal for any criterion (20), while ChatGPT 5.0 was the best LLM for web design with a score of around 2.8 as well (16).

The best LLMs for each use case were Perplexity.AI, Writer.AI, and ChatGPT 5.0, in the respective orders of the business cases (16, 17, 26). Perplexity.AI, though a general usage LLM, is typically more research based and could thereby be classified as more research heavy (17). Writer.AI is not a general usage LLM and is intentionally made in a way to create long responses with lots of information, similar to an outline or a business plan (26). ChatGPT 5.0 is more generic, and is used by many for various business cases (16, 25). Therefore, it is clear that depending on the case, the preference of general usage LLMs or specific ones differ, but as a whole, the more research and writing heavy LLMs which could be classified as specialized produce more optimal results (16, 17, 26).

In terms of pricing for business, ChatGPT 5.0 is the least expensive LLM from $25 to $60 (16), while Writer.AI is roughly $29–$39 for initial starting price (26). Perplexity.AI is the most expensive at roughly $40 a month or $400 yearly on a business plan (17). From this, it is clear that general use LLMs are cheaper than specialized ones. Overall, the choice of what LLM to use should be dependent on the need for responses, the price, and other factors, but as a whole, the prices seem reasonable for a startup (2, 13).

While this data was collected in a way to mitigate bias, there are limitations. One such limitation is the fact that only 25 prompts were analyzed for each of the three cases, which may not be a large enough prompt count to reduce the effect of other confounding variables in this experiment. The prompts themselves could only cover a finite range of topics meaning the effects of the prompts, temperatures, and roles that were never considered remain unknown. Additionally, the LLMs that did the evaluating could have some sort of unknown bias which could also confound the results.

For those reasons, future studies should do this experiment with more prompts and use cases, and should have multiple phases of prompt evaluation and response evaluation to lessen the impact of any bias. By doing so, future studies may be able to make a more concrete statement on the most optimal LLM that can be applicable to any startup (3, 12).

In the cases of market research and business support, the results indicated that specialized LLMs were better than general-usage LLMs. Specialized models are trained or fine-tuned on domain-specific knowledge and data. Market research and business support depends heavily on such knowledge. Thus, specialized models perform better on them. Web design is more general, and not domain specific. Therefore, ChatGPT provides comparable and sometimes even better performance than specialized LLMs. Even so, we observe that all three LLMs generally provide extremely good results on web design use cases, as evidenced by their higher scores in Figure 4. Our findings provide guidance for startups making the selection between specialized and general-purpose LLMs for specific use cases, turning LLM choice into a meaningful way to invest into startups rather than a technical afterthought.

CONCLUSION

To conclude, the findings included that for three business cases including Business Analysis, Web Design, and Market Research, certain LLMs proved to work better than others.These results challenge the assumption that one general-purpose model can effectively handle every type of prompt. Instead, they show that specialized LLMs, especially those built for research and analytical writing, can yield higher-quality outputs for startups seeking targeted solutions. At the beginning, the following question was posed: How do different large language models (LLMs) vary in their prompt responses in AI-driven solutions for startups, and what implications does this have for selecting generative AI tools in small business contexts? Through this study it is clear that startups should treat LLM model selection as a strategic decision rather than a technical afterthought. Choosing the right LLM can directly influence the efficiency and output quality, since the resources can be allocated more effectively towards more impactful tools.

Beyond summarizing the findings, this dataset serves as a foundation for future benchmarking and practical application, and contributes to a growing need for transparency and accessibility in AI evaluation. For startups, it offers an accessible and trustworthy tool to guide smarter AI adoption and budgeting decisions without relying solely on vendor claims. As a whole, this dataset serves to be a starting point for future researchers to include their benchmarking of different LLMs with other use cases considered, to make this dataset more universally usable and ensure that it can speak to a wider scope, making it a more viable tool for startups.

FUNDING SOURCES

The author declares no funding sources.

CONFLICT OF INTERESTS

The author declares no conflicts of interest related to this work.

REFERENCES

1. ClearlyPayments. *The number of businesses in the USA and statistics for 2024.* October 7, 2024. Available from: https://www.clearlypayments.com/blog/the-number-of-businesses-in-the-usa-and-statistics-for-2024/ (accessed on 2025-05-16).
2. CB Insights. *Why startups fail: Top 12 reasons.* November 30, 2022. Available from: https://www.cbinsights.com/research/report/startup-failure-reasons-top/ (accessed on 2025-05-10).
3. Boston University Technology & Policy Research Initiative. *Understanding how startups use AI.* Available from: https://sites.bu.edu/tpri/files/2025/01/Understanding-How-Startups-Use-AI_11-21-24.pdf (accessed on 2025-01-12).
4. Wall Street Journal (via LinkedIn). *US startups surge in new businesses amid pandemic.* September 2, 2024. Available from: https://www.linkedin.com/posts/changmarvin_rise-of-the-pint-size-startup-is-reshaping-

activity-7236728202067329024-84zu (accessed on 2025-05-15).
5. Jasper AI. Jasper (Large Language Model). Available from: https://www.jasper.ai/ (accessed on 2025-09-12).
6. Embroker. *Startup failure rate statistics (2025).* Exploding Topics, January 6, 2022. Available from: https://explodingtopics.com/blog/startup-failure-stats (accessed on 2025-03-03).
7. Embroker. *Startup failure rate (Full report).* Available from: https://www.embroker.com/blog/startup-failure-rate/ (accessed on 2025-03-03).
8. De Michele, Davide, et al. Automated Business Process Analysis: An LLM-Based Approach to Value Assessment. *arXiv preprint* arXiv:2504.06600, 2025. Available from: https://arxiv.org/pdf/2504.06600 (accessed on 2025-9-13).
9. Ahlgren TL, Sunde HF, Kemell K-K, and Nguyen-Duc A. Assisting early-stage software startups with LLMs: Effective prompt engineering and system instruction design. *Information and Software Technology*. 2025; 187: 107832. https://doi.org/10.1016/j.infsof.2025.107832
10. Ali L. and Chowdhury. Advancements and applications of generative AI and LLMs on business management. *Journal of Computer Science and Technology Studies*. 2023.
11. Kauffman Foundation. *Access to capital for entrepreneurs: Removing barriers (2023).* April 29, 2024. Available from: https://www.kauffman.org/reports/access-to-capital-removing-barriers-entrepreneurs-2023/ (accessed on 2025-01-20).
12. Turing.com. *How Large Language Models Are Changing the Face of Business Analytics.* Available from: https://www.turing.com/resources/how-llms-are-changing-the-face-of-business-analytics?utm_source=cha tgpt.com (accessed on 2025-9-13).
13. Fine Chris H. Operations for Entrepreneurs: Can Operations Management Make a Difference in New Ventures? *Production and Operations Management*. 2022. Available from: https://onlinelibrary.wiley.com/doi/10.1111/poms.13851(accessed on 2025-9-13).
14. Harvard Business Review. *Research: In recessions, employees avoid jobs with startups.* September 28, 2023. Available from: https://hbr.org/2023/09/research-in-recessions-employees-avoid-jobs-with-startups (accessed on 2025-05-19).
15. Chen Xi, et al. Evaluating Large Language Models on Business Process Modeling: Framework, Benchmark, and Self-Improvement Analysis. *arXiv preprint* arXiv:2412.00023, 2024. Available from: https://arxiv.org/pdf/2412.00023 (accessed on 2025-9-13).
16. M13. *5 roles you never knew you needed.* January 19, 2022. Available from: https://www.m13.co/article/5-roles-you-never-knew-you-needed (accessed on 2025-06-01).
17. Microsoft. Copilot (Large Language Model). Available from: https://copilot.microsoft.com/ (accessed on 2025-09-12).
18. Patel N. GoDaddy CEO Aman Bhutani on the Enduring Power of the Website. *The Verge*. November 25, 2024. Available from: https://www.theverge.com/24305364/godaddy-aman-bhutani-website-open-web-ai-decoder-podcast-interview (accessed on 2025-09-13).
19. Perplexity AI. Perplexity AI (Large Language Model). Available from: https://www.perplexity.ai/ (accessed on 2025-09-12).
20. Anthropic. Claude 3.5 Sonnet (Large Language Model). Available from: https://claude.ai/ (accessed on 2025-09-12).
21. Failory. *Startup failure rate: How many startups fail and why in 2025?* February 23, 2022. Available from: https://www.failory.com/blog/startup-failure-rate (accessed on 2025-04-08).
22. Lively J. Integrating AI-generative tools in web design education. Lindenwood University Faculty Research Papers. 2023. Paper 482.
23. Agentic Workers. Prompt Scorecard. Agentic Workers. Available from: https://www.agenticworkers.com/prompt-scorecard (accessed on 2025-09-13).
24. Martins Maicon Roberto. Startup Guide to AI: Integrating Technology for Business Success. *International Journal of Scientific Research and Management (IJSRM).* June 2024: 12 (06): 1264–74. Available from: https://ijsrm.net/index.php/ijsrm/article/view/5380 (accessed on 2025-09-13). https://doi.org/10.18535/ijsrm/v12i06.ec01
25. OpenAI. GPT-5.0 (ChatGPT) (Large Language Model). Available from: https://chat.openai.com/ (accessed on 2025-09-12).
26. Writer.com. Writer (Large Language Model). Available from: https://writer.com/ (accessed on 2025-09-12).
27. Windsurf AI. Windsurf (Large Language Model). Available from: https://windsurf.ai/ (accessed on 2025-09-12).