Generative Artificial Intelligence and Evaluating Strategic Decisions

Anil R. Doshi
University College London
UCL School of Management
anil.doshi@ucl.ac.uk

J. Jason Bell
University of Oxford
Saïd Business School
jason.bell@sbs.ox.ac.uk

Emil Mirzayev
University College London
UCL School of Management
e.mirzayev@ucl.ac.uk

Bart S. Vanneste
University College London
UCL School of Management
b.vanneste@ucl.ac.uk

29 July 2024

Acknowledgement: We are grateful to the strategy professors for their help in evaluating business models. We thank Sen Chai, Vivianna Fang He, Isabel Fernandez-Mateo, and Dan Sands. We also thank conference participants at the AI and Strategy Consortium, DRUID 24, HBS Idea-X, MAD Conference, Oxford Human-Algorithm Interaction Workshop, Strategy Science Conference, and the Sumantra Ghoshal Conference for their feedback. All authors contributed equally.

ABSTRACT

Strategic decisions are uncertain and often irreversible. Hence, predicting the value of alternatives is important for strategic decision making. We investigate the use of generative artificial intelligence (AI) in evaluating strategic alternatives using business models generated by AI (study 1) or submitted to a competition (study 2). Each study uses a sample of 60 business models and examines agreement in business model rankings made by large language models (LLMs) and those by human experts. We consider multiple LLMs, assumed LLM roles, and prompts. We find that generative AI often produces evaluations that are inconsistent and biased. However, when aggregating evaluations, AI rankings tend to resemble those of human experts. This study highlights the value of generative AI in strategic decision making by providing predictions.

Keywords: generative artificial intelligence, artificial intelligence (AI), large language models (LLMs), strategic decision making, strategic decisions, business models

1. INTRODUCTION

Strategic foresight—the ability to accurately predict the consequences of a strategic decision—is at the core of important strategy theories (Csaszar & Laureiro-Martinez, 2018; Gavetti & Menon, 2016; Kapoor & Wilde, 2023; Peterson & Wu, 2021). For example, when choosing a business model, superior performance may come from recognizing the attractiveness of a new industry in theories of competitive positioning (Porter, 1980) or from anticipating the value of a resource in the resource-based view (Barney, 1986). Strategic decisions involve uncertainty and often require commitment, making them irreversible or costly to undo (Ghemawat, 1991; Leiblein et al., 2018). Hence, predictions about the relative or absolute value of alternatives are important for strategic decision making. A key line of inquiry has been how individual evaluators and aggregations of their predictions affect the evaluation of a strategic decision (Csaszar & Eggers, 2013; Joseph & Gaba, 2020; Knudsen & Levinthal, 2007; Piezunka & Schilke, 2023).

Developments in artificial intelligence (AI) offer a new set of evaluators and potential aggregations to support evaluating strategic decisions, such as selecting a business model, choosing a firm to acquire, and recrafting an organization's design (Balasubramanian et al., 2022; Choudhary et al., 2023; Gaessler & Piezunka, 2023; Zohrehvand et al., 2024). In particular, generative AI consists of models that can produce high quality output, including text, images, and audio (Murphy, 2023). For instance, a large language model (LLM) generates human-like text based on neural networks with hundreds of billions of parameters (Chang et al., 2023). They are trained on vast corpora including web pages, news articles, and books. Once trained, an LLM requires little to no additional data to perform a new task (Brown et al., 2020). Instead, they can be guided using instructions, or "prompts" (Liu et al., 2023). Prompts enable a user to direct the model's response to a specific task, without requiring task-specific training

(Kojima et al., 2022). This flexibility can be useful when data for the task are limited or circumstances are unique, as is often the case for strategic decision making (Choi & Levinthal, 2023).

We extend research on human evaluators (e.g., Csaszar & Eggers, 2013; Knudsen & Levinthal, 2007), by investigating how artificial evaluators and aggregations of their predictions affect the evaluation of strategic decisions. We build on the insight that aggregating many imperfect predictions can improve the overall prediction by offsetting errors (He et al., 2022; Krogh & Vedelsby, 1994; Lichtendahl et al., 2013; Mollick & Nanda, 2016; Page, 2008). Aggregation of predictions has attracted interest (e.g., Csaszar & Eggers, 2013; Piezunka & Schilke, 2023), because the benefits from aggregation become significant precisely when individual predictions are challenging (Geman et al., 1992; Surowiecki, 2005), as is often the case for strategic decisions. We decompose the benefit derived from aggregating AI predictions into two different effects. We assess the benefit of diversity for a given level of scale (a diversity effect), by aggregating predictions from multiple LLMs (e.g., from Google, Meta, and OpenAI), roles (e.g., employee, investor, industry expert), and prompts, holding constant the number of predictions per AI evaluator. We then assess the benefit of scale for a given level of diversity (a scaling effect), by aggregating all predictions from the different LLMs, roles, and prompts.

In two studies, we focus on the strategic decision of choosing a business model (Casadesus-Masanell & Ricart, 2010; Guzman, Oh, & Sen, 2020; 2023; Kotha et al., 2023) that describes a firm's customers, products or services, and main activities (Markides, 2000; Massa et al., 2017). In the first study with a sample of 60 AI-generated business models, we examine the extent to which the business model rankings by generative AI agree with those of human experts. The rankings are derived from pairwise evaluations of business models. Our analysis yields three main results. First, selecting the business model that is more likely to succeed in a pairwise evaluation is challenging for generative AI. Its pairwise evaluations are often inconsistent (i.e., the order of business models affects choice), because they show bias (i.e., a systematic preference for the first or second business model). However, our second result demonstrates that aggregating pairwise evaluations can circumvent those challenges. Rankings based on many AI evaluations tend to agree with those of human experts (when aggregating all: correlations of 0.675 (Pearson) and 0.463 (Spearman), and choosing the same best business model in 5 out of 10 industries and the same worst business model in 6 of them). AI rankings agree with human experts more than human non-experts do. Third, we decompose the benefit derived from aggregating AI evaluations. The scaling effect (aggregating many evaluations for a given level of diversity) is more pronounced than the diversity effect (aggregating from multiple LLMs, roles, and prompts for a given level of scale).

In the second study with a sample of 60 business models submitted to a business model competition by entrepreneurs competing for a share of the USD one million in prize money, we assess the extent to which the findings from study 1 generalize to business models of actual startups. The AI rankings are generated in the same way as in study 1. The human expert rankings are based on scores awarded by judges of the business model competition. We find that the study 1 results mostly generalize, qualitatively and quantitatively, to the startups' business models in study 2.

This study makes two contributions to the literature on strategic decision making (e.g., Eisenhardt & Zbaracki, 1992; Peterson & Wu, 2021; Piezunka & Schilke, 2023). First, the focus has naturally been on human evaluators, because they have traditionally evaluated strategic decisions. We explore the role of artificial evaluators that rely on generative AI, and LLMs specifically. As LLMs become more capable and widespread within firms, they offer the potential to help build strategic foresight. Second, the literature has investigated aggregations of evaluations, because they are most beneficial when individual predictions are difficult (Geman et al., 1992). When aggregating predictions from human evaluators, a challenge is deciding whose predictions to aggregate. A goal is to obtain predictions that differ from each other so that errors cancel out (Page, 2008). Approaches include selecting on evaluators' skill level, cognitive style, and demographics (Almaatouq et al., 2024; Csaszar & Eggers, 2013; De Oliveira & Nisbet, 2018; Knudsen & Levinthal, 2007). When aggregating predictions from artificial evaluators, that challenge is the same but the solutions differ. We explore selecting on evaluators' LLM, role, and prompt. We analyze the benefits of these selections by considering both a diversity effect (aggregating more LLMs, roles, and prompts) and a scaling effect (aggregating more evaluations). Taken together, this study demonstrates the potential role of artificial evaluators for strategic decision making by providing predictions.

2. BACKGROUND

2.1 Strategic decisions

Strategic decisions have been described as important, difficult to undo, and involving uncertainty (Eisenhardt & Zbaracki, 1992; Elbanna & Child, 2007; Van den Steen, 2018). They are important because they significantly affect the success or failure of organizations (Porter, 1980) and guide and constrain subsequent decisions (Casadesus-Masanell & Ricart, 2010; Mintzberg et al., 1976). These decisions are difficult to undo due to the path dependence they create, given their interdependence with subsequent decisions (Leiblein et al., 2018; Page, 2008), and because they often involve committing scarce resources (Ghemawat, 1991). Additionally, they involve uncertainty because relevant future states are difficult to anticipate and their occurrence probabilities hard to quantify (Arend, 2024; Levinthal, 2011). Thus, strategic decisions are crucial due to their importance but are challenging to get right because of the inherent uncertainty. Additionally, their irreversibility means that making these decisions involves making predictions.

2.2. Human evaluators

Past research has investigated how individuals and their aggregations can contribute to strategic foresight, the ability to accurately predict the consequences of a strategic decision (Gavetti & Menon, 2016). At the individual level, the focus has been on individuals' cognition (Gavetti & Levinthal, 2000; Helfat & Peteraf, 2015), including their mental representations, experiences, and biases (Bardolet et al., 2011; Csaszar & Laureiro-Martinez, 2018; Gary & Wood, 2011; Kapoor & Wilde, 2023; Peterson & Wu, 2021). At the aggregation level, a key consideration is the organizational structures used to combine individual predictions (Csaszar & Eggers, 2013; Joseph & Gaba, 2020; Knudsen & Levinthal, 2007; Piezunka & Schilke, 2023).

Aggregating many imperfect predictions can improve the overall prediction (Lichtendahl et al., 2013; Mollick & Nanda, 2016), where improvement is typically assessed as a reduction in prediction error (i.e., the difference between a prediction and the actual outcome). Aggregating can occur in many ways, including through averaging or majority voting (Csaszar & Eggers, 2013). The benefit of aggregation—also called wisdom of the crowds (Surowiecki, 2005)—has long been of interest (Condorcet, 1785; Galton, 1907).

Aggregating multiple predictions yields a better prediction if their errors at least partially offset. For example, if the task is predicting the value of a strategic alternative (i.e., a continuous outcome or a "regression" task), positive prediction errors offset negative prediction errors (Larrick & Soll, 2006). Alternatively, if the task is predicting which of two strategic alternatives is better (i.e., a discrete outcome or “classification” task), sufficient correct predictions offset incorrect predictions through majority voting (Hansen & Salamon, 1990). Offsetting implies that aggregation performs better with greater diversity and number of predictions (Krogh & Vedelsby, 1994; Page, 2008). Moreover, aggregating is especially beneficial in contexts where a single prediction is more likely to be error-prone (Geman et al., 1992; Ueda & Nakano, 1996), such as strategic decision making.

2.3. Artificial evaluators

Recent advances in generative AI, and in particular LLMs, offer the possibility of artificial evaluators.

2.3.1. Large language models

A language model is a model that predicts the next word given a sequence of words (Murphy, 2023). When language models become large, abilities emerge that are absent in smaller models, including answering previously unseen questions, performing arithmetic, and reasoning over multiple steps (Wei et al., 2022a). This advancement marks a significant leap in the field of natural language processing (NLP).

LLMs are based on deep learning, or neural networks. The term “large” in the name signifies the extensive number of parameters used by these models, which can exceed one trillion. A key element of these models is the Transformer architecture (Vaswani et al., 2017), which improves data processing efficiency by parallelizing computations. It also provides an effective way to discern patterns in the data, facilitating the models’ ability to understand context and to generate contextually relevant text. Learning good values of many parameters requires substantial computational power and data. Typically, model training is done on thousands of powerful graphics processing units (GPUs) designed for executing parallel computations.

Notwithstanding such high levels of computing power, training an LLM can take multiple months (Naveed, 2023). A key source of data for training an LLM is the corpus of billions of publicly available websites (e.g., as captured in the Common Crawl).[1] LLMs are trained using a self-supervised learning approach, where the model learns to predict parts of the input data from other parts of the data, without explicit external labels. This approach enables the model to learn a comprehensive language representation, which can be queried with prompts, instead of requiring further training for new tasks.

2.3.2. LLM predictions

LLMs possess the ability to make predictions whereby an output is assigned to a new input (Chang et al., 2023). A classic prediction example is text classification, a common task in NLP. Applications include sentiment analysis, where an LLM categorizes text as positive, negative, or neutral based on its tone, and email filtering, where emails are sorted into categories like spam, travel, or invoices. A more recent prediction example is question answering, where an LLM is provided a question and predicts the answer, by either selecting the correct answer from a list or generating its own answer (Kojima et al., 2022). An important overall evaluation metric for LLMs is the extent to which they can correctly answer questions (Hendrycks et al., 2021). Furthermore, when provided with a pair of different responses to a single question, LLMs can be asked to predict which better addresses the user's question. Here the output is the response preference and the input is the question with the pair of responses. The LLM judgment can then be compared to a human's evaluation of the same responses (Zheng et al., 2023).

Our setup mirrors this approach: we present an LLM with pairs of business models. Its task is to analyze and determine which of the two business models is more likely to succeed. In this context, the LLM's evaluation of business models is effectively a prediction task because it

[1] https://commoncrawl.org/

assigns a label (i.e., “preferred” to one business model) to a new input (i.e., a pair of business models).

2.3.3. Aggregating LLM predictions

The wisdom of the crowd, or the benefit of aggregating predictions, depends on two mechanisms: the crowd’s diversity and scale (Keuschnigg & Ganser, 2017; Page, 2008; Surowiecki, 2005). The positive impact of each on aggregation can be mathematically derived (Geman et al., 1992; Jiang et al., 2017; Krogh & Vedelsby, 1994; Ueda & Nakano, 1996; Wood et al., 2023). First, diversity indicates that the predictions differ from each other. Through aggregating diverse predictions, incorrect predictions can be offset. Optimistic predictions cancel out pessimistic predictions (for a continuous outcome) or correct predictions outweigh incorrect ones (for a discrete outcome). If predictions are not diverse, then the aggregated prediction will resemble a typical individual prediction and the crowd cannot provide much benefit. Second, scale refers to the number of predictions contributing to the aggregation. By aggregating many predictions, offsetting inaccurate predictions becomes more probable. For a continuous outcome, selecting only a few predictions might result in a set of mostly optimistic predictions or of mostly pessimistic predictions, offering limited aggregation benefits in either case. However, selecting many predictions is more likely to include both optimistic and pessimistic predictions, enhancing the aggregation benefits (Batchelor & Dua, 1995). For a discrete outcome (with individual predictions correct at least half the time), selecting a few predictions occasionally leads to the wrong result. However, when aggregating many predictions, the correct predictions are more likely to outweigh the incorrect ones, making the correct result increasingly likely (Dietterich, 2000).

To benefit from aggregating LLM predictions, these mathematical principles imply that artificial evaluators must also ensure diversity and scale. Achieving scale is viable using generative AI because the technology's flexibility and scalability (Kojima et al., 2022) allow for the generation of many predictions. Achieving diversity is less straightforward. Predictions need not only be diverse but also accurate on the whole. Without knowing the correct prediction, it is challenging to know which diverse predictions are beneficial. Instead, a common approach is to source from diverse evaluators (Page, 2008). For example, past research on human evaluators has considered the implications of differences in their skill level, cognitive style, and demographics (Almaatouq et al., 2024; Csaszar & Eggers, 2013; De Oliveira & Nisbet, 2018; Knudsen & Levinthal, 2007).

Likewise, we investigate the implications of differences in artificial evaluators' LLMs, roles, and prompts. First, LLMs can generate different predictions for identical tasks, partly because they employ different neural network architectures and are trained on partially different data. Even if their overall accuracy is comparable, LLMs may make different mistakes (Hendrycks et al., 2021; Wei et al., 2022a). Second, just as humans' perspective or role influences their predictions (Page, 2008), LLMs' assigned role influences their output (Boussioux et al., 2024; Deshpande et al., 2023; Xu et al., 2023). For example, an LLM performed better on multiple choice questions in different fields (e.g., biology, econometrics, or international law) when impersonating a domain expert than a non–domain expert (Salewski et al., 2023). Third, different instructions or prompting approaches may influence the prediction an LLM makes (Wei et al., 2022b). LLMs use attention mechanisms to focus on different parts of the input when generating responses (Vaswani et al., 2017). Different prompts can shift the model's attention to various aspects of the input, influencing the final output. Additionally, complex problems can be approached in different ways, and the design of prompts plays a crucial role in this process (Wei et al., 2022b).

3. STUDY 1: AI-GENERATED BUSINESS MODELS

Study 1 used AI-generated business models for reasons of internal validity. We aimed for evaluators to focus on assessing the prospective success of business models rather than being influenced by correlates like presentation style (Tsay, 2021). Using AI-generated business models ensures consistency across models (e.g., same components, cadence, style, length). Furthermore, the business models were short to prevent fatigue among the human evaluators. The study's pre-registration is available at: https://aspredicted.org/TH5_LDK.

3.1 Methods

3.1.1. Generating business models

We used GPT-4 from OpenAI (version gpt-4-0613) accessed via an application programming interface (API) to generate 60 startup business models (= 10 industries × 2 prompts × 3 probabilities of succeeding). First, from the Global Industry Classification Standard (GICS),[2] we selected 10 industries: commercial printing, passenger ground transportation, education services, apparel retail, food retail, brewers, health care equipment, consumer finance, application software, and movies & entertainment. These industries have a mix of target markets (business versus consumer) and outputs (products versus services). Second, we used two prompts (provided in the Online Appendix). The base prompt asked for a business model in a given industry with a description of the customers, products or services, and main activities in one query (Markides, 2000; Massa et al., 2017). The chain-of-thought prompt asked first for only the customers, second for the products or services, and last for the main activities. Chain-of-thought prompting is a popular technique that breaks down complex tasks into simpler, sequential tasks

[2] https://www.msci.com/our-solutions/indexes/gics

(Wei et al., 2022b). Third, we asked for business models that had low, medium, or high probability of succeeding in the prompt, because no objective ex-ante measure of the business models' viability exists. It is unlikely that each business model aligns with its category, but we expected that on average the business models from the different categories will differ in viability, offering an opportunity to test the LLMs' predictive capabilities for strategic decisions.

We sought business models with between 75 and 125 words to balance the level of detail of the business model and time required for human evaluation. To promote creativity, we used a temperature of 0.7, a setting that controls the randomness of the LLM's response. For each combination of industry, prompt, and probability of succeeding, the LLM generated three business models. One that met the word count was randomly selected (for examples, see Table 1). Mean and standard deviation word count of the 60 selected business models are 83.3 and 8.1, respectively.

[[INSERT TABLE 1 ABOUT HERE]]

3.1.2. Evaluating business models

Business models were assessed through pairwise evaluations, with business models in each pair from the same industry. Evaluations were performed separately by generative AI, human experts, and human non-experts. An evaluator was asked to indicate which business model of a pair was more likely to succeed. Each generative AI evaluator assessed all pairs. Human experts and separately non-experts covered all pairs collectively, but each individual only assessed a subset to mitigate survey fatigue. A human expert evaluated 10 randomly selected pairs (one pair from each industry). A human non-expert evaluated three randomly selected pairs (each from a different industry). For both human experts and non-experts the sample sizes were chosen to include each of the 60 business models in 10 pairs, on average. For random sampling of pairs, we used a discrete choice experiment design (McFadden, 1986) in which business models are randomly sampled from the full factorial of generation characteristics (i.e., prompting approach [base, chain-of-thought], and success probability [low, medium, high]), conditional on industry.

3.1.2.1. Generative AI

We used 7 LLMs × 10 roles × 2 prompts. First, we accessed seven LLMs via API: Claude2 (from Anthropic), PaLM2 and Gemini Pro (from Google and used in its Bard service), Llama2 (an open source model from Meta), and GPT-3.5, GPT-4, and GPT-4 Turbo (from OpenAI and used in its ChatGPT service). Details of the LLMs are provided in Table 2. These models were chosen based on their API accessibility, widespread use, and performance in public leaderboards.[3] Each LLM is used with a temperature of 0, because we sought its most confident response. Second, we instructed the LLMs to take on 10 different roles, five of which are connected to the startup (the founder, an investor, an employee, a potential customer, and a potential supplier) and five of which are not (a strategy professor, an industry expert, a journalist with the Financial Times, a politician, an environmental activist). These roles were chosen so that individually they should be knowledgeable about the potential prospects of a startup, while collectively their opinions may differ. Third, we used two prompts. The base prompt asked which business model is more likely to succeed by indicating "A" or "B." The chain-of-thought prompt urged an LLM to reason before giving a final response. It asked first to compare the internal fit of the business models (i.e., how well each business model's elements fit together), second to compare their external fit (i.e., how well each business model fits with its external environment), third to explain which is more likely to succeed, and finally to simply indicate "A" or "B."

[[INSERT TABLE 2 ABOUT HERE]]

[3] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard (accessed on December 16, 2023).

Each LLM evaluated ordered pairs to allow for the possibility that the evaluation of business models A versus B differs from that of B versus A. The total number of evaluations obtained from this procedure was 42,000 (= 10 industries × 6 business models × 5 pairings per model × 7 LLMs × 10 roles × 2 prompts). After excluding 4,122 invalid evaluations, the number of usable pairwise evaluations was 37,878. We excluded all 3,000 evaluations of PaLM2 with the base prompt, because it chose the second business model in 99.9% of evaluations.[4] We retained the evaluations of PaLM2 with the chain-of-thought prompt. We also sought to exclude evaluations that did not indicate a choice. Of the remaining evaluations, in 86.4% of cases, the AI responded exclusively "A" or "B," as requested. Hence, no formatting was required. In 10.7% of evaluations the AI indicated which business model was more likely to succeed, but not in the requested format (even after asking for the correct format). For example, "My apologies! Here's my revised answer: A". Hence, formating was required and we extracted the A or B choice matching text patterns. In 2.9% of (or 1,122) evaluations, the AI did not indicate which business model was more likely to succeed. For example, "I apologize for any confusion caused." Hence, no choice was provided and we excluded these evaluations.[5] See the Online Appendix for details of evaluations of each by LLM and prompting approach.

To decompose any aggregation benefits, we combine pairwise evaluations into three types of AI evaluators (see Figure 1). First, we aggregate predictions from a single LLM, role, and prompt. Each of the resulting 130 "uniform AI evaluators" contains a maximum of 300 pairwise evaluations (bottom left side of figure). Second, we aggregate predictions from multiple LLMs, roles, and prompts, by randomly sampling without replacement a maximum of 300 pairwise evaluations. We stratified by ordered pair of two business models. Thus, each of the

[4] Including these evaluations yields similar results as those reported.
[5] As a robustness check, we included these evaluations. Because these pairwise evaluations have no winner, we code these as ties (i.e., 0.5 and 0.5 for both business models). The results are similar as those reported.

resulting 130 “mixed AI evaluators” will have approximately the same number and types of pairs (bottom middle of figure). Small differences occur due to the 2.9% of evaluations that did not yield a choice. Third, we aggregate the 37,878 predictions from all LLMs, roles, and prompts, resulting in one “comprehensive AI evaluator” (bottom right of figure). The mixed AI evaluators and comprehensive AI evaluator allow us to decompose the aggregating benefits attributable to the diversity and scaling effects, respectively.

[[INSERT FIGURE 1 ABOUT HERE]]

The aggregation is based on the typical approach of averaging (Davis-Stober et al., 2014) and yields a “win” proportion for each business model. As an illustrative and simplified example, let us focus on three business models—BM1, BM2, and BM3—and two LLMs—LLM1 and LLM2 (ignoring roles and prompts). Each LLM indicates the preferred business model for each ordered pair. This leads to six pairwise evaluations per LLM, or a total of 12 evaluations for both LLMs combined. The 12 evaluations are used to create the three types of AI evaluators. First, a uniform AI evaluator aggregates six predictions from a single LLM. Based on the six predictions, we can calculate for each business model its win proportion (i.e., number of wins / number of pairwise evaluations). For example, imagine that BM2 won three of the four evaluations it was part of, then its win proportion is 0.75. Because there are two LLMs, there are two uniform AI evaluators. Second, a mixed AI evaluator aggregates six predictions from multiple LLMs. The six pairwise evaluations are randomly drawn without replacement from stratified ordered pairs. The procedure described above is used to compute the win proportion per business model. Just as with uniform AI evaluators, there are two mixed AI evaluators. Third, the sole comprehensive AI evaluator consists of all 12 evaluations. The same procedure is used to calculate the win proportion per business model. Since all evaluations are combined, there is only one comprehensive AI evaluator.

3.1.2.2. Human experts

We recruited 100 human experts by emailing well respected strategy professors (39% women; 39% US-based; rank: 35% assistant, 36% associate, 29% full) at globally renowned institutions (e.g., 34% from top 20 Financial Times business schools[6]). We asked for their help in the evaluation task within two weeks of the email request. The response rate was 51% (no responses were excluded), resulting in 510 pairwise evaluations (= 51 respondents × 10 pairs per respondent).

3.1.2.3. Human non-experts

We recruited 150 human non-experts on Prolific, an online platform. We excluded 14 people who completed the task within one minute or after 15 minutes, or incorrectly answered an attention check question. Respondents (*n*=136, 53% women) resided in the United States, were at least 18 years old (*mean*=36.3, *s.d.*=11.5), had working experience (5% 1-2 years, 18% 3-5 years, 26% 6-10 years, 16% 11-15 years, 33% 16+ years), and received 1.50 USD in compensation.[7] The responses yielded a total of 408 pairwise evaluations (= 136 respondents × 3 pairs per respondent).

3.1.3. Variables

We assess the agreement with human experts along four outcomes. Each is based on the proportion of "wins" for a business model. A business model wins if it is deemed more likely to succeed than the other business model in a pairwise evaluation. *Pearson correlation* is the correlation for the 60 business models between the win proportion by human experts and the win

---

[6] MBA 2023 Business school ranking, see https://rankings.ft.com/.

[7] Some demographic data was unavailable because respondents chose not to disclose it.

proportion by AI (or by human non-experts). *Spearman correlation* is the rank correlation for the 60 business models based on their win proportion ranked by industry. Ties are given the average rank of the group. For example, if two business models are joint third in an industry, then their rank is 3.5. *Top choice* is the proportion of industries for which the business model ranked highest by human experts is also ranked highest by AI (or by human non-experts). To account for ties, we focused on business models with rank 1 or 1.5. Human experts had eight industries with one top ranked business model and two industries with two top ranked business models. For those two industries, a top choice was registered as long as the AI (or human non-expert) had at least one of these business models ranked highest. *Bottom choice* is as *top choice*, except for the lowest-ranked business model per industry. To account for ties, we focused on business models with rank 5.5 or 6. Human experts only had unique lowest rank business models.

For the uniform and mixed AI evaluators, each outcome is computed for the individual AI evaluators and then the average (and standard error) is presented. There is only one comprehensive AI evaluator, so the four values corresponding to the measures are its computed values.

3.2. Results

3.2.1. Agreement in outcomes

At the *level of pairwise evaluation*, predicting the most promising business model is challenging for an LLM, as evaluations are susceptible to inconsistency and bias. The left panel of Figure 2 shows the proportion of pairwise evaluations that are consistent, that is, when the evaluation of business models A and B yields the same prediction as the evaluation of B and A. Inconsistency is common, with evaluations at most 80.9% consistent (for GPT-4 Turbo with chain-of-thought prompt) and frequently much lower (e.g., 42.2% for Claude2 with base prompt). The right panel shows the proportion of pairwise evaluations in which the last option is preferred, in other words, when the evaluation of business models A versus B yields the prediction B. Because every pair is evaluated twice (i.e., A versus B and B versus A), unbiased evaluation should yield the last option in 50% of all instances. However, LLMs can exhibit bias. Some systematically favor the last option, as observed in 83.5% of the evaluations for GPT-3.5 (base), and others disfavor the last option, as seen in Gemini Pro (chain-of-thought) with only 29.6% of the cases.

[[INSERT FIGURE 2 ABOUT HERE]]

When *aggregating many pairwise evaluations*, AI evaluators tend to agree with human experts more than human non-experts do, and the extent of agreement increases with greater diversity and scale of aggregations. Figure 3 shows the agreement with human expert evaluators on the four outcomes. First, uniform AI evaluators tend to agree with human experts more than human non-experts do. *Pearson correlation*, *top choice*, and *bottom choice* values for uniform AI evaluators (human non-experts) are 0.570 (0.447), 0.327 (0.200), and 0.553 (0.400). The one exception is *Spearman correlation*, where the value is 0.405 (0.416). Second, mixed AI evaluators are on average more similar to human experts than are the uniform AI evaluators across all four measures (*Pearson correlation* 0.590, *Spearman correlation* 0.427, *top choice* 0.353, and *bottom choice* 0.568). Third, the comprehensive AI evaluator shows even greater agreement with the human experts (with values of 0.675, 0.463, 0.500, and 0.600, respectively).

[[INSERT FIGURE 3 ABOUT HERE]]

To assess the statistical significance of differences between evaluators, we obtain 95% confidence intervals. We cannot base these on the standard errors of the means for two reasons. First, for uniform and mixed AI evaluators, they yield confidence intervals for the means but not for the difference in means (as the estimates are non-independent). Second, for comprehensive

AI evaluators and human non-experts, the standard errors are unavailable because their means are based on a direct aggregation of all predictions. Instead, we employ a jackknife approach (Quenouille, 1956; see Arslan, 2023 for a recent example), a resampling technique that involves a leave-one-out procedure.[8] We iteratively exclude one evaluator at a time, which is a combination of an LLM, prompt, and role (for the AI evaluators) and a respondent (for the human non-experts). After each resampling, the four outcomes are calculated. The resulting 95% confidence intervals are provided in the Online Appendix. For each outcome, we compare the AI evaluators with the human non-experts, and with each other. The general ordering is that agreement is greatest for the comprehensive AI evaluator, then the mixed AI evaluators, then the uniform AI evaluators, and finally the human non-experts. The one exception is the Spearman correlation, for which again the comprehensive AI evaluator agrees most but the mixed and uniform AI evaluators align similarly as the human non-experts.

For understanding the economic significance, we assess the different outcome variables. *Pearson correlation* and *Spearman correlation* capture the average alignment in assessment of all business models. *Top choice* and *bottom choice* capture the average alignment for the winning and losing business models, respectively. Whereas the four outcome variables are expected to move together as a first approximation, they highlight distinct types of agreement in evaluations. The two correlation measures capture whether evaluations broadly agree, while the top and bottom choices provide agreement on the options that may matter more in a selection process.

Pearson correlation uses the win proportions. The alignment with human experts is 0.447 for human non-experts and 0.675 for the comprehensive AI evaluator. The uniform AI evaluator sits approximately in the middle of these two values. The agreement of the comprehensive AI

[8] An alternative resampling technique is bootstrapping whereby sampling occurs with replacement. This approach is less suitable because of the requirement that each uniform and mixed AI evaluator covers all business models.

evaluator represents a 51.0% and 18.4% increase over the human-non experts and the uniform AI evaluators, respectively. Spearman correlation uses the ranked win percentages. The alignment with human experts is 0.416 for human non-experts and 0.463 for the comprehensive AI evaluator. The uniform AI evaluator correlation of 0.405 falls just below the human non-experts. The increase in agreement of the comprehensive AI evaluator over the human non-experts and the uniform AI evaluators is 11.3% and 14.3%, respectively. A correlation increase of less than 0.05 suggests that overlap in rank order with that of the human experts is fairly similar for the human non-experts and the comprehensive AI evaluator. Thus, the comprehensive AI evaluator's greater alignment with human experts in actual win proportions (*Pearson correlation*) is not accompanied by a substantially greater alignment in average ranking (*Spearman correlation*).

However, it does yield substantive differences in agreement at the top and bottom of the rankings. For *top choice*, human non-experts match human experts in only 2 out of 10 industries, whereas the comprehensive AI evaluator aligns in 5 out of 10 industries. The uniform AI evaluator sits approximately in the middle of these two values. This alignment matters if selecting only the winners is the key motivation of the evaluation. It matters even more if these winners achieve outsized returns, as often seen in entrepreneurial environments (Malenko et al., 2024). For *bottom choice*, human non-experts align in 4 out of 10 industries, whereas the comprehensive AI evaluator matches in 6 out of 10 industries. The uniform AI evaluator is close to the comprehensive AI evaluator. Avoiding losers is useful in selection processes, such as an incubator seeking to allocate its limited resources. It is also critical if losing options result in significant costs. For example in entrepreneurial environments, founders incur significant costs if they pursue poor strategies (Gans, Stern, &Wu, 2019).

In the Online Appendix, we investigate whether business model characteristics similarly affected human experts and AI evaluations.

3.2.2. Unpacking the diversity and scaling effects

Successive improvements in agreement among the mixed and comprehensive AI evaluators provide evidence for the benefits of diversity and scaling, respectively. Relative to the uniform AI evaluator, the diversity effect yielded improvements of between 3% and 8% across the four measures and the scaling effect yielded additional improvements of between 6% and 45%. We unpack the diversity effect (by considering different levels of diversity across LLMs, roles, and prompts) and the scaling effect (by considering different numbers of pairwise evaluations). For the diversity effect, the main results showed the comparison between 130 uniform AI evaluators (with a single LLM, role, and prompt) and 130 mixed AI evaluators (with multiple LLMs, roles, and prompts), and all had a maximum of 300 pairwise evaluations. Here we also show the results for intermediate levels of diversity, whereby each AI evaluator is sampled along some dimension(s) and the other dimension(s) are kept constant (e.g., multiple LLMs and prompts and a single role). For the scaling effect, the main results showed the comparison between the 130 mixed AI evaluators (each with a maximum of 300 pairwise evaluations) and one comprehensive AI evaluator (that aggregated all 37,878 pairwise evaluations). Here we also show the results for intermediate levels of scale, whereby each AI evaluator is aggregated along some but not all dimensions (e.g., multiple LLMs and prompts and a single role). Aggregating along more dimensions implies the aggregation of more pairwise evaluations.

We use indicators to capture the selection dimensions (LLM, role, prompt). For example, (1, 0, 1) means that AI evaluators are sampled with multiple LLMs and prompts and a single role. A total of $8 = 2^3$ combinations are feasible, including (0, 0, 0) and (1, 1, 1). Figure 4 shows the outcomes for all combinations and Table 3 provides the tabular results. In the figure, the average of uniform AI evaluators (0, 0, 0) is shown in light blue. The diversity effect is shown with one shade darker. For example, on average mixing roles increases the Pearson correlation by 0.010 (i.e., (0, 1, 0): 0.580 vs. (0, 0, 0): 0.570). The diversity effect tends to increase with the number of dimensions along which is sampled. On average, mixing increases an outcome with 0.007 for one dimension, 0.014 for two dimensions, and 0.021 for three dimensions. For dimension type, no clear differences in the selection effect emerge when comparing a single dimension (i.e., (1, 0, 0) vs. (0, 1, 0) vs. (0, 0, 1)).

[[INSERT FIGURE 4 ABOUT HERE]]

[[INSERT TABLE 3 ABOUT HERE]]

For each combination other than (0, 0, 0), we can scale the AI evaluators. For example, mixing (0, 1, 0) yields 130 AI evaluators, each including multiple roles and a single LLM and prompt (e.g., Gemini Pro × chain-of-thought) and each with a maximum of 300 pairwise evaluations. Scaling (0, 1, 0) yields 13 AI evaluators, each still including multiple roles and a single LLM and prompt but now with a maximum of 3,000 pairwise evaluations. For example, we can scale the 10 AI evaluators of Gemini Pro × chain-of-thought (each with multiple roles and a maximum of 300 pairwise evaluations), by aggregating all pairwise evaluations yielding one Gemini Pro × chain-of-thought evaluator with at most 3,000 pairwise evaluations. Thus, mixing makes AI evaluators more diverse. Scaling makes AI evaluators bigger for a given level of diversity.

The scaling effect is shown with the darkest shade. The scaling effect is more pronounced than the diversity effect. Furthermore, the scaling effect tends to increase with the number of dimensions along which is scaled. On average, scaling increases an outcome with 0.041 for one dimension, 0.064 for two dimensions, and 0.075 for three dimensions. The AI evaluators' scale increases with the number of dimensions. For example, the maximum number of pairwise evaluations increases is 600 for (0, 0, 1) and is 37,878 for (1, 1, 1). Thus, the diversity effect tends to be stronger with more diversity and the scaling effect with more pairwise evaluations, and the scaling effect is more pronounced than the diversity effect.

4. STUDY 2: BUSINESS MODEL COMPETITION

Study 2 investigates the external validity of the study 1 findings. We use data from a business model competition hosted by an US university and held in 2016. Applicants competed for a share of USD one million in total prize money.

4.1. Methods

4.1.1. Business models

Applicants are entrepreneurs of early-stage, US-based startups. They submitted detailed textual information on their business model. Each business model description was structured using the same predefined categories, such as a problem statement, value proposition, and pathway towards growth. We obtained the same number of business models as in study 1 by randomly sampling 60 from the 71 submissions. On average, a selected business model had 2,207.7 words (s.d.=381.0).

4.1.2. Evaluators

4.1.2.1. Generative AI

We used the same approach as in study 1, with minor changes where required. Business models were evaluated pairwise. We obtained a similar number of evaluations as in study 1 by randomly assigning the business models to 10 groups. All pairwise evaluations are within the same group. Paired business models are not necessarily in the same industry.

We used 6 LLMs × 10 roles × 2 prompts. First, we used the following six LLMs (see Table 2): Gemini Pro 1.0 and Gemini Pro 1.5 (from Google), Llama3 (an open source model from Meta), Mistral Large (an open source model from Mistral), and GPT-3.5 and GPT-4o (from OpenAI).[9] This set differs from study 1 because input capacity of some LLMs was insufficient for longer business models (i.e., Llama2, PaLM2) and running costs of others were prohibitive (i.e., Claude2, GPT-4, GPT-4 Turbo). Second, we used the same ten roles. Those connected to a startup required slight rewording as the business models now related to two startups instead of one (e.g., "a founder of a startup" instead of "the founder of this startup"). Third, we used the same two prompts: base and chain-of-thought. Again, a slight rewording was required because the evaluation was no longer for a single industry.

The total number of evaluations was 36,000 (= 10 groups × 6 business models × 5 pairings per model × 6 LLMs × 10 roles × 2 prompts). After excluding 0.07% (or 25) evaluations that did not indicate a choice, the number of usable pairwise evaluations was 35,975. For additional details, see the Online Appendix.

4.1.2.2. Human experts

The number of judges was 70, selected by the competition organizers for their relevant knowledge and experience (including entrepreneurs, investors, and academics). They assessed the business models using pre-defined criteria (e.g., innovativeness, scalability) on scales from one to five. A judge evaluated 4.3 business models, on average. A business model was independently evaluated by 5 judges, with their scores summed. Consequently, each business model received a single score.

[9] The number of LLMs is the number stated in study 1's pre-registration (see the Online Appendix). We report the results from all LLMs that we ran.

4.1.3. Variables

We use the same variables: *Pearson correlation*, *Spearman correlation*, *top choice*, and *bottom choice*. These are based on the business models' win proportions for AI and single scores for human experts. Ranks are from the rankings within a group.

4.2. Results

As Figure 2 showed for study 1, Figure 5 shows that *pairwise evaluations* produced by AI are often inconsistent (left panel) and biased (right panel) for study 2. Consistency (when the evaluation of business models A and B yields the same prediction as the evaluation of B and A) ranged from 29.9% (GPT-3.5 using the chain-of-thought prompt) to 78.1% (Llama 3 using the base prompt). Without bias, we would expect the first option and second options to be chosen with equal frequency (i.e., 50%). However, there was often a bias against the second option (e.g., 16.3% with Mistral Large using the chain-of-thought prompt) or, less commonly, in favor of the second option (i.e., 70.7% with Gemini Pro 1.5 using the base prompt).

[[INSERT FIGURE 5 ABOUT HERE]]

As Figure 3 showed for study 1, Figure 6 shows that, when *aggregating many pairwise evaluations*, AI and human expert evaluators tend to agree for study 2. Furthermore, agreement increases with greater diversity and scaling. For each of the four measures, agreement with human experts improved from uniform to mixed to comprehensive AI evaluators. The agreement with human expert evaluators of the uniform AI evaluator was 0.505, 0.525, 0.349, and 0.543 for *Pearson correlation*, *Spearman correlation*, *top choice*, and *bottom choice*, respectively. The mixed AI evaluator improved agreement over the uniform AI evaluator (values were 0.548, 0.572, 0.370, and 0.549 across the four measures, respectively). The comprehensive AI evaluator improved agreement over the mixed AI evaluator (0.663, 0.720, 0.400, and 0.600, respectively).

[[INSERT FIGURE 6 ABOUT HERE]]

To assess statistical significance, we use the same jackknife approach to arrive at 95% confidence intervals (provided in the Online Appendix). As with study 1, agreement with human experts is greatest for the comprehensive AI evaluator, followed by the mixed AI evaluators, and then the uniform AI evaluators. All differences exclude zero from the confidence interval, except for the difference between uniform and mixed AI evaluators in *bottom choice*.

Regarding economic significance, the comprehensive AI evaluator tends to align with human experts for the average assessment of all business models in terms of both win proportions (*Pearson correlation* of 0.663) and ranking (*Spearman correlation* of 0.720). As for the specific assessment of best and worst business models, the comprehensive AI evaluator also fairly consistently selects the *top choice* (4 out of 10 industries) and *bottom choice* (6 out of 10). The agreement between comprehensive AI evaluators and human experts is substantially greater than that of either uniform or mixed AI evaluators with human experts. The *Pearson correlation* for the comprehensive AI evaluator is 31.3% and 21.0% higher than those of the uniform and mixed AI evaluators, respectively. Similarly, the comprehensive AI's *Spearman correlation* is 37.1% and 25.8% higher than the uniform and mixed AI evaluators, respectively. For the differences of *top choice* and *bottom choice*, the comprehensive AI evaluator increases agreement between 8.1% and 14.6% over the uniform and mixed AI evaluators. Furthermore, mixed AI evaluators provide only relatively small increases in agreement over the uniform AI evaluators with absolute gains of less than 0.05 in each of the four measures.

4.2.1. Comparing studies 1 and 2

The study 1 results mostly generalize, qualitatively and quantitatively, to study 2. First, in both studies, individual evaluations were inconsistent and biased. Compared to study 1's evaluations, those of study 2 are somewhat less consistent (0.605 [study 1] versus 0.521 [study 2]) and more biased in terms of the average absolute deviation from 0.5 (0.149 versus 0.212). Interestingly, study 1's evaluations are biased more towards the second business model, whereas those in study 2 are biased more towards the first business model. Second, comparing the agreement between the comprehensive AI evaluator and the human experts, we find similar results across the two studies for *Pearson correlation* (0.675 versus 0.663) and *bottom choice* (0.600 versus 0.600). For *Spearman correlation*, the value was lower in study 1 than in study 2 (0.463 versus 0.720). Conversely, for *top choice*, the value was higher in study 1 than in study 2 (0.500 versus 0.400). Third, comparing the contributions of the diversity and scaling effects, we find that in both studies the scaling effect is more pronounced than the diversity effect. In study 1, diversity led to higher agreement from 3.2% to 8.0% (across four outcome measures) and scaling led to higher agreement from 8.5 to 52.9%. In study 2, higher agreement from diversity and scaling ranged from 3.3% to 9.5%, and 16.0% to 37.1%, respectively.

5. DISCUSSION

In two studies—the first using AI-generated business models and the second using business models submitted to a competition—we find that individual generative AI evaluations are inconsistent and biased, whereas aggregating these evaluations results in increased agreement with human experts. The increase comes from diversity and scaling evaluations produced by LLMs, roles, and prompts. Both studies show modest increases from diversity and substantial increases from scaling. The difference in magnitudes demonstrates that, in these two contexts, gains from aggregation are achieved primarily through scaling. We speculate that the relative contribution of diversity and scaling to the gains from aggregating AI evaluations depends on the nature of the evaluation task. Hence, it is possible that in other contexts their relative contributions will differ. Two dimensions that might matter are the completeness of information provided in the evaluation (i.e., the extent to which information required to make the evaluation is provided) and the timing for when uncertainty is resolved (i.e., when the evaluation's outcome will be realized).

The study's key contribution is to highlight the value of generative AI in providing predictions for strategic decision making, a task that is critical and complex. First, the strategist is central for making strategic decisions (Van den Steen, 2018). Until now, the "strategist" was assumed to be a human. The emergence of generative AI presents intriguing possibilities for how strategic decisions are made. Though current instances of LLMs have not (yet) achieved artificial general intelligence, they show a capacity for reasoning (Chen et al., 2023) that may support or complement the human decision-maker. As generative AI develops and evolves, it may become perceived as an effective substitute for expert humans (Vanneste & Puranam, 2024). Regardless, generative AI offers the prospect of altering the locus of strategic decision-making and possibly how humans discover and implement new strategic opportunities. This study provides an initial step toward understanding the role of an artificial actor in strategic decision making.

As a second contribution, we show an approach to aggregating predictions from artificial evaluators for the evaluation of strategic decisions. Predictions from human evaluators have been aggregated based on their characteristics, including skill level, cognitive style, and demographics (Almaatouq et al., 2024; Csaszar & Eggers, 2013; De Oliveira & Nisbet, 2018; Knudsen & Levinthal, 2007). We show that predictions from artificial evaluators can be usefully aggregated by LLM, role, and prompt. Such understanding is particularly valuable in the context of evaluating strategic decisions, characterized by uncertainty, because aggregation is most beneficial when predicting is difficult.

When interpreting the results, we must keep in mind two aspects. First, establishing the actual viability of business models is challenging. To make progress, we follow prior work that has relied on crowd evaluation, whereby a group of human evaluators independently assess an idea, product, or business model (Mollick & Nanda, 2016; Terwiesch & Ulrich, 2023). We found close but not full alignment between their evaluations. In study 1, it is not known who selected the better business models in the absence of information on strategic outcomes (e.g., financial performance). Observed differences between human and AI evaluations could indicate inferior or superior AI performance. In study 2, the human experts' evaluations were the basis of an economically significant outcome for the startups in the competition. The close alignment provides a more direct assessment of AI's ability to evaluate strategic decisions. Thus, our measures pertain to agreement with experts, whose assessments on the likelihood of success may differ from eventual success or other longer term outcomes.

Second, our results provide a snapshot of current capabilities of LLMs. As LLMs continue their development, we expect that single evaluations will be more consistent (i.e., the order of business models does not influence choice) and less biased (i.e., no systematic preference for the first or second business model), but these may persist to some extent. Our intuition is that the difficulty of comparing business models contributed, at least partially, to the inconsistencies and biases. We probe this intuition directly by conducting a small test with an easier task (for details, see the Online Appendix): selecting which shape has the larger area (e.g., a circle with radius of 2 units versus a triangle with sides of 4 units). We used 3 LLMs (with low, medium, and high reported inconsistencies and biases in study 2), 2 prompts (base, chain-of-thought), and 5 roles (that match the roles that were unconnected to the startup). In line with expectations, we observed higher consistency (0.792 versus 0.546 in study 2 for the same

LLMs, where the maximum is 1) and less bias in terms of choosing the last option (0.483 versus 0.305 in study 2 for the same LLMs, where approximately 0.500 is anticipated in the absence of a bias). Hence, to the extent that strategic decisions are difficult to evaluate, we anticipate inconsistencies and biases may persist when using artificial evaluators in the domain of strategy.

We suggest a key implication for practice, where managers are considering how to integrate generative AI into their organizations. The results provide managers with an approach to using generative AI as part of the strategic decision making process. Rather than relying on a single prompt made to a single LLM, if managers were to aggregate evaluations of a decision across LLMs, prompts, or roles, we posit that the resulting evaluations will be more informative. This approach allows managers to obtain inputs for strategic decisions with relatively low investments in time or resources, which can be combined with human inputs. Such an approach could scale across organizational domains, including mergers and acquisitions (Cuypers et al., 2017) or market entry (Li et al., 2015).

We highlight the following avenues for future research. First, we compared AI evaluators and human experts. Comparing AI with human experts is most insightful if human experts' evaluations are correlated with a strategic outcome or constitute such an outcome. Future research might investigate how generative AI evaluations compare with both human evaluations and an outside outcome, when the latter is available. Second, the aggregation approach follows the typical approach of averaging (Davis-Stober et al., 2014): the win proportion of a business model (i.e., the number of "winning" evaluations over the number of total evaluations). Alternative aggregation approaches exist. For example, aggregations can weight certain evaluations more than others, including overweighting based on whether an individual is a "champion" that can override the majority (Malenko et al., 2024), has great ability (Keuschnigg

& Ganser, 2017), or possesses inside information (Chen et al., 2004). Future research could assess the extent to which alternative aggregation approaches can yield tractable results among AI evaluators. Third, we used roles to induce diversity in evaluations and averaged over all roles, rather than to find roles that performed more effectively. In an exploratory analysis, we inspected ten AI evaluators that had a single role and multiple LLM and prompts (i.e., one AI evaluator per role).[10] In each study, different roles exhibited the highest average agreement with human experts. In study 1, the top three roles were the customer, investor, and supplier, while in study 2, they were the industry expert, strategy professor, and journalist. Future research might probe how roles could be devised to enhance the resulting aggregations of strategic evaluations. Finally, strategic decisions are characterized by how they are "uncovered, framed, developed, evaluated, and implemented" (Leiblein et al., 2018, p. 559). We focused primarily on evaluating (and secondarily on developing). Future work can investigate the potential role of generative AI in each of the other areas. Relevant questions include how generative AI can frame the strategic decision in the most meaningful or impactful way and how it can assess the impact of a strategic decision's implementation.

The numerous future avenues of research are indicative of the possible transformation that generative AI might herald for strategic decision making. This transformation is made possible by the different approach that generative AI takes compared to the traditional paradigm for prediction, namely supervised learning. Whereas a supervised learning model is specific to a task and requires historical input-output pairs, generative AI requires little to no additional data to perform a new task (Brown et al., 2020). This property opens up the possibility of generative

[10] Specifically, we considered the aggregation by LLM and prompt (i.e., AI (1, 0, 1)) and with the maximum number of predictions (i.e., scaling).

AI’s role in strategic decision making, where data are often limited and the circumstances are unique.

REFERENCES

Almaatouq, A., Alsobay, M., Yin, M., & Watts, D. J. (2024). The effects of group composition and dynamics on collective performance. *Topics in Cognitive Science*, 16(2), 302-321.

Arend, R. J. (2024). Uncertainty in strategic decision making: Analysis, categorization, causation and resolution. Palgrave Macmillan Cham.

Arslan, H. A., Tereyağoğlu, N., & Yılmaz, Ö. (2023). Scoring a touchdown with variable pricing: Evidence from a quasi-experiment in the NFL ticket markets. *Management Science*, 69(8), 4435-4456.

Balasubramanian, N., Ye, Y., & Xu, M. (2022). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47(3), 448-465.

Bardolet, D., Fox, C.R. and Lovallo, D. (2011), Corporate capital allocation: a behavioral perspective. Strat. Mgmt. J., 32: 1465-1483.

Barney, J. B. (1986). Strategic factor markets: Expectations, luck, and business strategy. *Management Science*, 32(10), 1231–1241.

Batchelor, R., & Dua, P. (1995). Forecaster diversity and the benefits of combining forecasts. *Management Science*, 41(1), 68-75.

Boussioux, L., Jane, J. L., Zhang, M., Jacimovic, V., & Lakhani, K. (2024). The crowdless future? How generative AI is shaping the future of human crowdsourcing. *Harvard Business School Technology & Operations Mgt. Unit. Working Paper, (24-005)*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., …, & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Casadesus-Masanell, R., & Ricart, J. E. (2010) From strategy to business models and onto tactics. *Long Range Planning*, 43(2-3), 195-215.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., …, & Xie, X. (2023). A survey on evaluation of large language models. arXiv preprint arXiv:2307.03109.

Chen, K., Fine, L. & Huberman, B. (2024). Eliminating public knowledge biases in information-aggregation mechanisms. *Management Science*, 50, 983–994.

Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *PNAS*, 120(51), 1-9.

Choi, J., & Levinthal, D. (2023). Wisdom in the wild: generalization and adaptive dynamics. *Organization Science*, 34(3), 1073-1089.

Choudhary, V., Marchetti, A., Shrestha, Y. R., & Puranam, P. (2023). Human-AI ensembles. When can they work? *Journal of Management*, Forthcoming.

Condorcet, M. D. (1785). Essay on the application of analysis to the probability of decisions rendered by a plurality of votes. *Classics of Social Choice*, 91-112.

Csaszar, F. A., & Eggers, J. P. (2013). Organizational decision making. An information aggregation view. *Management Science*, 59(10), 2257-2277.

Csaszar, F. A., & Laureiro-Martínez, D. (2018). Individual and organizational antecedents of strategic foresight: A representational approach. *Strategy Science*, 3(3), 513-532.

Cuypers, I. R. P., Cuypers, Y., & Martin, X. (2017). When the target may know better: effects of experience and information asymmetries on value from mergers and acquisitions *Strategic Management Journal*, 38(3), 609–625.

Davis-Stober C.P., Budescu D.V., & Dana, J., & Broomell, S. (2014) When is a crowd wise? *Decision* 1(2):79–101.

De Oliveira, S., & Nisbett, R. E. (2018). Demographically diverse crowds are typically not much wiser than homogeneous crowds. *Proceedings of the National Academy of Sciences*, 115(9), 2066-2071.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K. (2023). Toxicity in chatgpt: analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In International Workshop on Multiple Classifier Systems (pp. 1-15). Springer, Berlin, Heidelberg.

Eisenhardt, K. M., & Zbaracki, M. J. (1992). Strategic decision making. *Strategic Management Journal*, 13(S2), 17-37.

Elbanna, S., & Child, J. (2007). Influences on Strategic Decision Effectiveness: Development and Test of an Integrative Model. Strategic Management Journal, 28(4), 431–453.

Gaessler, F., & Piezunka, H. (2023). Training with AI: evidence from chess computers. *Strategic Management Journal*, 44, 2724–2750.

Galton, F. (1907). Vox Populi. *Nature*, 75(1949), 450-451.

Gans JS, Stern S, & Wu J. (2019). Foundations of entrepreneurial strategy. *Strategic Management Journal*, 40: 736–756.

Gary, M.S. and Wood, R.E. (2011), Mental models, decision rules, and performance heterogeneity. Strat. Mgmt. J., 32: 569-594.

Gavetti, G., & Levinthal, D. (2000). Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly,* 45(1), 113-137.

Gavetti, G., & Menon, A. (2016). Evolution cum agency: Toward a model of strategic foresight. *Strategy Science*, 1(3), 207-233.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), 1-58.

Ghemawat, P. (1991) *Commitment*. Simon & Schuster, New York NY.

Guzman, J., Oh, J. J., & Sen, A. (2020). What motivates innovative entrepreneurs? Evidence from a global field experiment. *Management Science*, 66(10), 4808-4819.

Guzman, J., Oh, J. J., & Sen, A. (2023). Climate change framing and innovator attention: Evidence from an email field experiment. *Proceedings of the National Academy of Sciences*, 120(3), e2213627120.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), 993-1001.

He, L., Analytis, P. P., & Bhatia, S. (2022). The wisdom of model crowds. *Management Science*, 68(5), 3635-3659.

Helfat, C.E. and Peteraf, M.A. (2015), Managerial cognitive capabilities and the microfoundations of dynamic capabilities. Strat. Mgmt. J., 36: 831-850.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., …, & Steinhardt, J. (2021). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

Jiang, Z., Liu, H., Fu, B., & Wu, Z. (2017, February). Generalized ambiguity decompositions for classification with applications in active learning and unsupervised ensemble pruning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

Joseph, J., & Gaba, V. (2020). Organizational structure, information processing, and decision making: A retrospective and roadmap for research. Academy of Management Annals, 14(1), 267–302.

Kapoor, R., & Wilde, D. (2023). Peering into a crystal ball: Forecasting behavior and industry foresight. *Strategic Management Journal*, 44(3), 704-736.

Keuschnigg, M., & Ganser, C. (2017). Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Management science*, 63(3), 818-828.

Knudsen, T., & Levinthal, D. A. (2007). Two faces of search: Alternative generation and alternative evaluation. *Organization Science*, 18(1), 39-54.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199-22213.

Kotha, R., Vissa, B., Lin, Y., & Corboz, A. V. (2023). Do ambitious entrepreneurs benefit more from training? *Strategic Management Journal*, 44(2), 549-575.

Krogh, A., & Vedelsby, J. (1994). Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems,* 7, 231–238.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: misappreciation of the averaging principle. *Management Science*, 52(1), 111-127.

Leiblein, M. J., Reuer, J. J., & Zenger, T. (2018). What makes a decision strategic? *Strategy Science*, 3(4), 558-573.

Levinthal, D.A. (2011), A behavioral approach to strategy—what's the alternative?. Strat. Mgmt. J., 32: 1517-1523.

Li, J., Qian, C., & Yao, F. K. (2015). Confidence in learning: inter- and intraorganizational learning in foreign market entry decisions. *Strategic Management Journal*, 36(6), 918–929.

Lichtendahl, Jr. K. C., Grushka-Cockayne, Y., & Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, 59(7), 1594-1611.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35.

Malenko, A., Nanda, R., Rhodes-Kropf, M., & Sundaresan, S. (2024). Catching Outliers: Committee Voting and the Limits of Consensus when Financing Innovation. Harvard Business School Entrepreneurial Management Working Paper No. 21-131.

Markides, C. (2000) *All The Right Moves: A Guide to Crafting Breakthrough Strategy*. Harvard Business School Press, Boston MA.

Massa, L., Tucci, C. L., & Afuah, A. (2017) A critical assessment of business model research. *Academy of Management Annals*, 11(1), 73-104.

McFadden, D. (1986). The choice theory approach to market research. *Marketing Science*, 5(4), 275-279.

Mintzberg, H., Raisinghani, D., & Théorêt, A. (1976). The Structure of "Unstructured" Decision Processes. *Administrative Science Quarterly*, 21(2), 246–275.

Mollick, E., & Nanda, R. (2016). Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science*, 62(6), 1533-1553.

Murphy, K. P. (2023). *Probabilistic Machine Learning: Advanced Topics*. MIT Press, Cambridge MA.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., …, & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

Page, S. E. (2008). *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, Princeton NJ.

Peterson, A., & Wu, A. (2021). Entrepreneurial learning and strategic foresight. *Strategic Management Journal*, 42(13), 2357-2388.

Piezunka, H., & Schilke, O. (2023). The dual function of organizational structure: Aggregating and shaping individuals' votes. *Organization Science*, 34(5), 1914-1937.

Porter, M. E. (1980). Competitive strategy: Techniques for analyzing industries and competitors. New York, NY: Free Press.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, 43(3/4), 353-360.

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., & Akata, Z. (2023). In-context impersonation reveals large language models' strengths and biases. arXiv.

Surowiecki, J. (2005). *The Wisdom of Crowds.* Anchor, New York NY.

Terwiesch, C., & Ulrich, K. (2023) *The Innovation Tournament Handbook: A Step-by-Step Guide to Finding Exceptional Solutions to Any Challenge*. University of Pennsylvania Press, Philadelphia, PA.

Tsay, C. J. (2021). Visuals dominate investor decisions about entrepreneurial pitches. *Academy of Management Discoveries*, 7(3), 343-366.

Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. *Proceedings of International Conference on Neural Networks* (ICNN'96), 1, 90-95.

Van den Steen, E. (2018). Strategy and the strategist: how it matters who develops the strategy. *Management Science*, 64(10), 4533-4551.

Vanneste, B. S. & Puranam, P. (2024), Artificial Intelligence, trust, and perceptions of agency, *Academy of Management Review*, Forthcoming.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., ..., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 6000-6010.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., …, & Fedus, W. (2022a). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., .., & Zhou, D. (2022b) Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.

Wood, D., Mu, T., Webb, A. M., Reeve, H. W., Lujan, M., & Brown, G. (2023). A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24(359), 1-49.

Xu, B., Yang, A., Lin, J., Wang, Q., Zhou, C., …, & Mao, Z. (2023). ExpertPrompting: instructing large language models to be distinguished experts. arXiv preprint arXiv:2305.14688.

Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., …, & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.

Zohrehvand, A., Doshi, A. R., & Vanneste, B. S. (2024). Generalizing event studies using synthetic controls: An application to the Dollar Tree–Family Dollar acquisition. *Long Range Planning*, 57(1), 102392.

Table 1. Examples of AI generated business models (study 1)

(Apparel Retail industry, chain-of-thought prompt, high probability of succeeding)
Customer: Parents of young children who want to dress their kids in trendy, comfortable clothing made from safe, eco-friendly materials. They value convenience and often shop online.
Product or service: An online boutique for children's wear focusing on trendy designs and eco-friendly materials, with a subscription service for regular delivery of age-appropriate clothes.
Main activities: Selecting eco-friendly children's wear, maintaining an online boutique, managing a subscription service, coordinating regular deliveries, and handling customer service inquiries.

(Commercial Printing industry, base prompt, medium probability of succeeding)
Customer: Individuals or families who want to preserve their memories in physical forms such as photo books, calendars, or custom printed gifts.
Product or service: A wide range of personalized printed products including photo books, calendars, custom printed mugs, t-shirts, and canvas prints. Offering easy-to-use online design tools.
Main activities: Providing an online platform for customers to upload photos and customize their design, managing printing and production process, packaging, and shipping the final product to customers.

(Movies & Entertainment industry, base prompt, low probability of succeeding)
Customer: Our target customer is the niche market of silent movie enthusiasts. These individuals, often cinema history buffs, appreciate the artistry of silent films and are actively seeking ways to enjoy these classics.
Product or service: We are creating an online platform dedicated exclusively to silent movies, providing access to a vast library of silent films from around the world.
Main activities: Our activities involve sourcing and digitizing silent films, curating the library, and maintaining the online platform. We also run silent film history and appreciation seminars.

Note. For each probability of succeeding, one business model was randomly selected (with each business model from a different industry). The industry, prompt approach, and probability of succeeding are provided for each business model.

Table 2. Overview of large language models (LLMs) used for evaluation

Name	Developer	Version	Release date	URL	Study 1	Study 2
Claude	Anthropic	claude-2.0	11 Jul 2023	www.anthropic.com/index/claude-2	✓	-
PaLM2	Google	-	10 May 2023	ai.google/discover/palm2/	✓	-
Gemini Pro	Google	1.0	6 Dec 2023	deepmind.google/technologies/gemini/	✓	✓
	Google	1.5	15 Feb 2024	deepmind.google/technologies/gemini/	-	✓
Llama	Meta	2.0	18 Jul 2023	ai.meta.com/llama/	✓	-
	Meta	3.0	18 Apr 2024			✓
Mistral Large	Mistral	open-mixtral-8x7b	11 Dec 2023	mistral.ai/news/mixtral-of-experts/	-	✓
GPT-3.5	OpenAI	gpt-3.5-turbo-0613	6 Nov 2023	platform.openai.com/docs/models	✓	✓
GPT-4	OpenAI	gpt-4-0613	14 Mar 2023	platform.openai.com/docs/models	✓	-
GPT-4 Turbo	OpenAI	gpt-4-1106-preview	6 Nov 2023	platform.openai.com/docs/models	✓	-
GPT-4o	OpenAI	gpt-4o-2024-05-13	13 May 2024	platform.openai.com/docs/models	-	✓

Table 3. Results for agreement with human experts on four outcomes (study 1)

Evaluator				Max	Number of	Pearson correlation		Spearman correlation		Top choice		Bottom choice	
	LLM	Role	Prompt	evaluations	evaluators	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
AI	0	0	0	300	130	0.570	0.008	0.405	0.009	0.327	0.013	0.553	0.008
Increasing diversity													
AI	0	0	1	300	140	0.565	0.006	0.414	0.007	0.339	0.011	0.561	0.005
AI	1	0	0	300	140	0.572	0.006	0.412	0.007	0.334	0.010	0.555	0.007
AI	0	1	0	300	130	0.580	0.007	0.415	0.008	0.344	0.012	0.559	0.007
AI	1	0	1	300	130	0.584	0.005	0.429	0.007	0.351	0.011	0.563	0.006
AI	0	1	1	300	140	0.572	0.005	0.419	0.007	0.347	0.011	0.571	0.006
AI	1	1	0	300	140	0.579	0.006	0.417	0.007	0.342	0.012	0.559	0.006
AI	1	1	1	300	130	0.590	0.005	0.427	0.007	0.353	0.011	0.568	0.006
Increasing scale													
AI	0	0	1	600	70	0.614		0.445		0.363		0.577	
AI	1	0	0	2,100	20	0.648		0.461		0.410		0.600	
AI	0	1	0	3,000	13	0.633		0.455		0.354		0.585	
AI	1	0	1	3,858	10	0.657		0.473		0.420		0.600	
AI	0	1	1	6,000	7	0.654		0.486		0.371		0.586	
AI	1	1	0	20,711	2	0.670		0.486		0.500		0.600	
AI	1	1	1	37,878	1	0.675		0.463		0.500		0.600	
Human non-experts					1	0.447		0.416		0.200		0.400	

Note. A "1" in "Evaluator" columns indicates along which dimension of LLM, role, prompt is aggregated. The term "s.e." is the standard error, which is calculated as the standard deviation divided by the square root of the number of evaluators (and not shown for "scale" due to the lower number of evaluators). To match to the AI evaluators presented in the main text, AI (0, 0, 0) is the 130 uniform AI evaluators, the diverse AI (1, 1, 1) is the 130 mixed AI evaluators, and the scaled AI (1, 1, 1) is the comprehensive AI evaluator.

Table 4. Results for agreement with human experts on four outcomes (study 2)

Evaluator				Max	Number of	Pearson correlation		Spearman correlation		Top choice		Bottom choice	
	LLM	Role	Prompt	evaluations	evaluators	mean	s.e.	mean	s.e.	mean	s.e.	mean	s.e.
AI	0	0	0	300	120	0.505	0.013	0.525	0.016	0.349	0.014	0.543	0.015
Increasing diversity													
AI	0	0	1	300	120	0.514	0.013	0.538	0.015	0.365	0.015	0.543	0.015
AI	1	0	0	300	120	0.518	0.013	0.542	0.015	0.366	0.015	0.548	0.015
AI	0	1	0	300	120	0.532	0.008	0.547	0.013	0.362	0.013	0.539	0.013
AI	1	0	1	300	120	0.523	0.013	0.550	0.014	0.368	0.013	0.561	0.013
AI	0	1	1	300	120	0.539	0.007	0.559	0.009	0.382	0.013	0.546	0.013
AI	1	1	0	300	120	0.548	0.006	0.569	0.008	0.377	0.012	0.549	0.010
AI	1	1	1	300	120	0.548	0.006	0.572	0.008	0.370	0.012	0.549	0.009
Increasing scale													
AI	0	0	1	600	60	0.546		0.573		0.365		0.568	
AI	1	0	0	1800	20	0.586		0.637		0.385		0.620	
AI	0	1	0	3000	12	0.589		0.610		0.358		0.525	
AI	1	0	1	3600	10	0.598		0.659		0.410		0.630	
AI	0	1	1	6000	6	0.617		0.629		0.367		0.533	
AI	1	1	0	17998	2	0.654		0.690		0.400		0.550	
AI	1	1	1	35975	1	0.663		0.720		0.400		0.600	

Note. A "1" in "Evaluator" columns indicates along which dimension of LLM, role, prompt is aggregated. The term "s.e." is the standard error, which is calculated as the standard deviation divided by the square root of the number of evaluators (and not shown for "scale" due to the lower number of evaluators). To match to the AI evaluators presented in the main text, AI (0, 0, 0) is the 120 uniform AI evaluators, the diverse AI (1, 1, 1) is the 120 mixed AI evaluators, and the scaled AI (1, 1, 1) is the comprehensive AI evaluator.

Figure 1. From AI evaluations to AI evaluators (study 1)

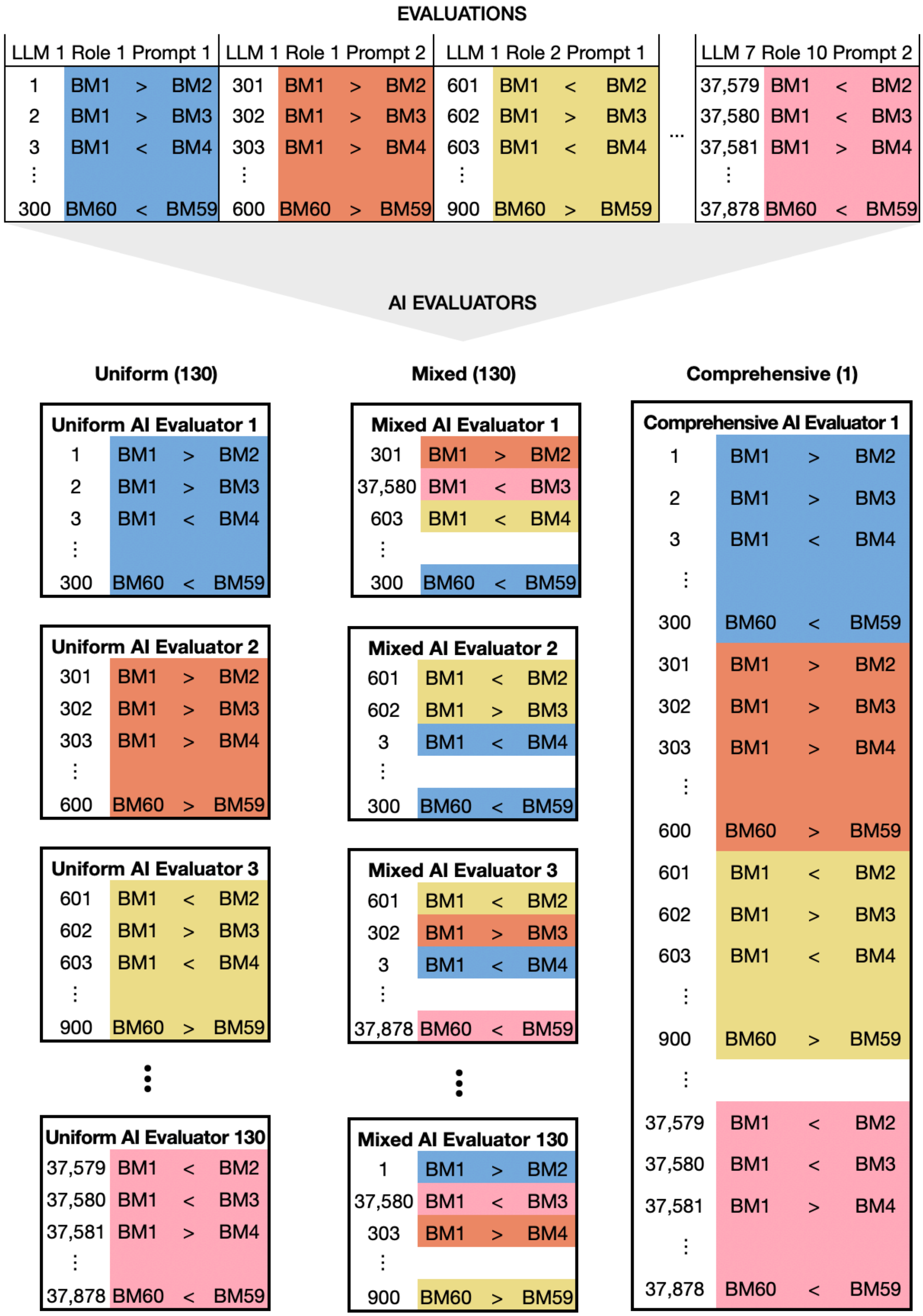
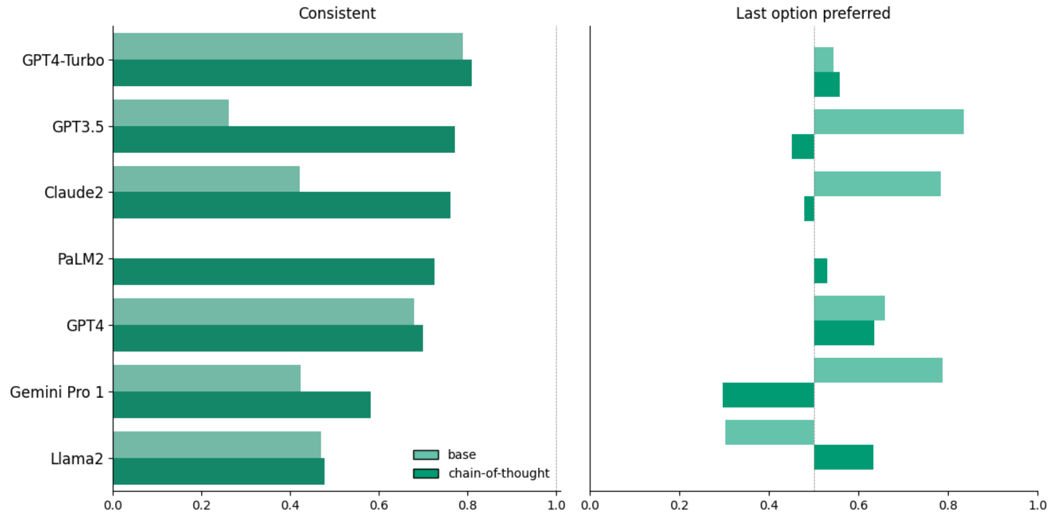
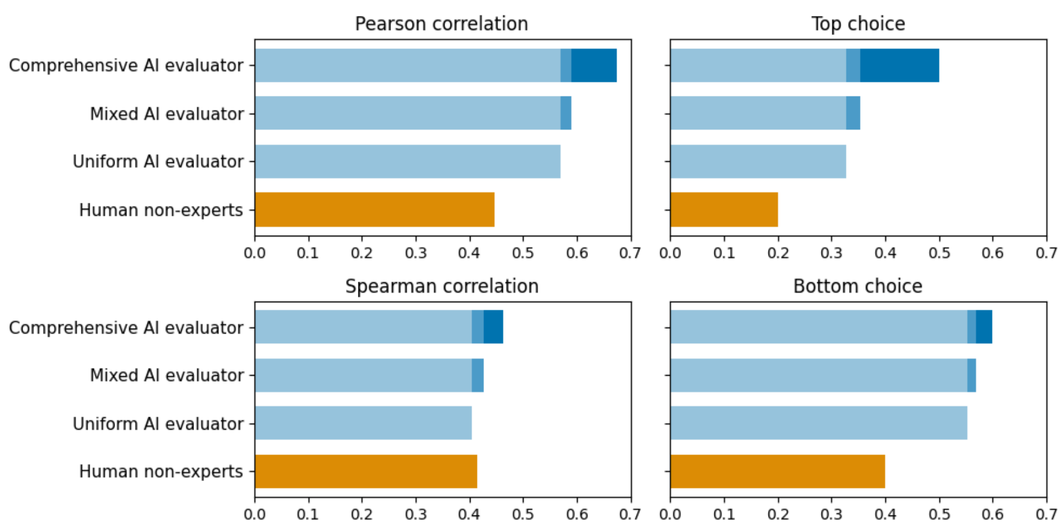


Figure 2. Proportion of pairwise evaluations by AI that are consistent and that yield the last option (study 1)



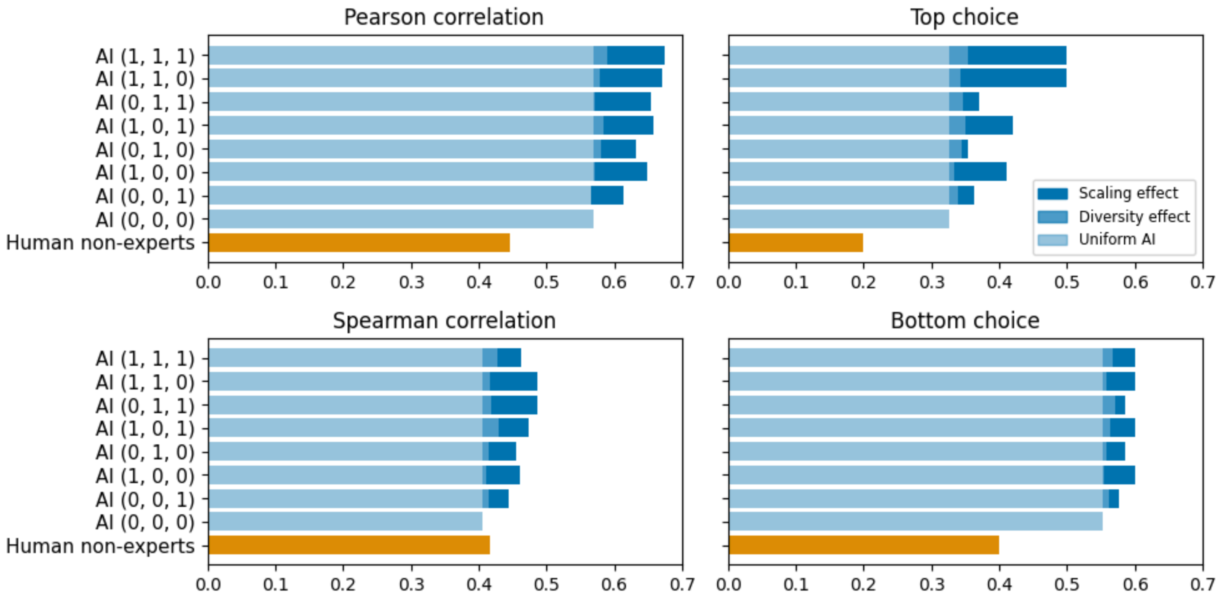
Note. Evaluations are consistent if, for a pair of business models, the pair's ordering does not affect the evaluation. The proportions in the left panel are based on all 3,000 evaluations per LLM and prompt. Evaluations yield the last option if, for a pair of business models, the second business model is chosen. The proportions in the right panel are calculated after excluding the 2.9% of evaluations that did not yield a choice.

Figure 3. Agreement with human experts on four outcomes (study 1)



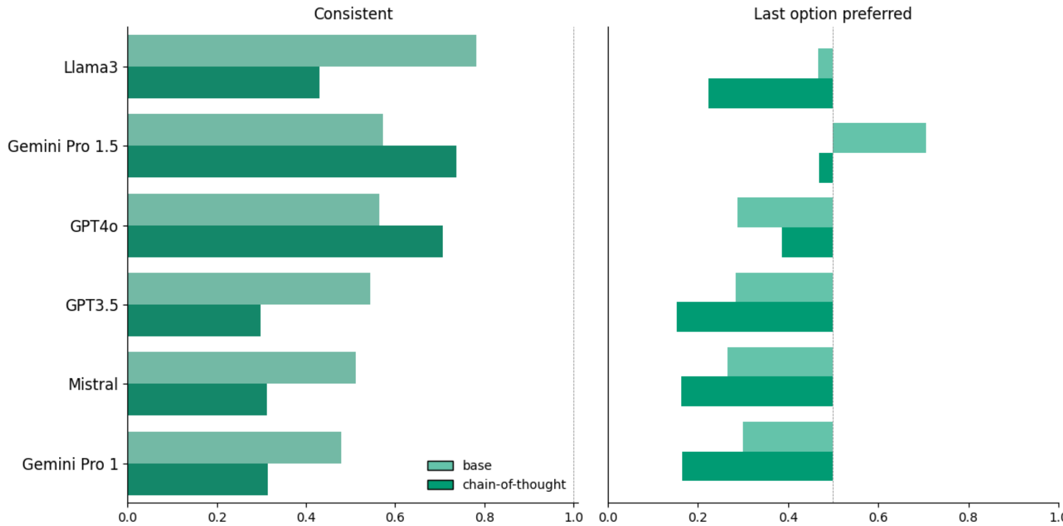
Note. Each panel shows a measure of agreement with human experts' evaluations and those of human non-experts and three AI evaluators. The uniform AI evaluator is indicated in light blue. The incremental gain from the mixed AI evaluator is shown in medium blue (i.e., the diversity effect) and that of the comprehensive AI evaluator in dark blue (i.e., the scaling effect).

Figure 4. Unpacking the diversity and scaling effects for agreement with human experts (study 1)



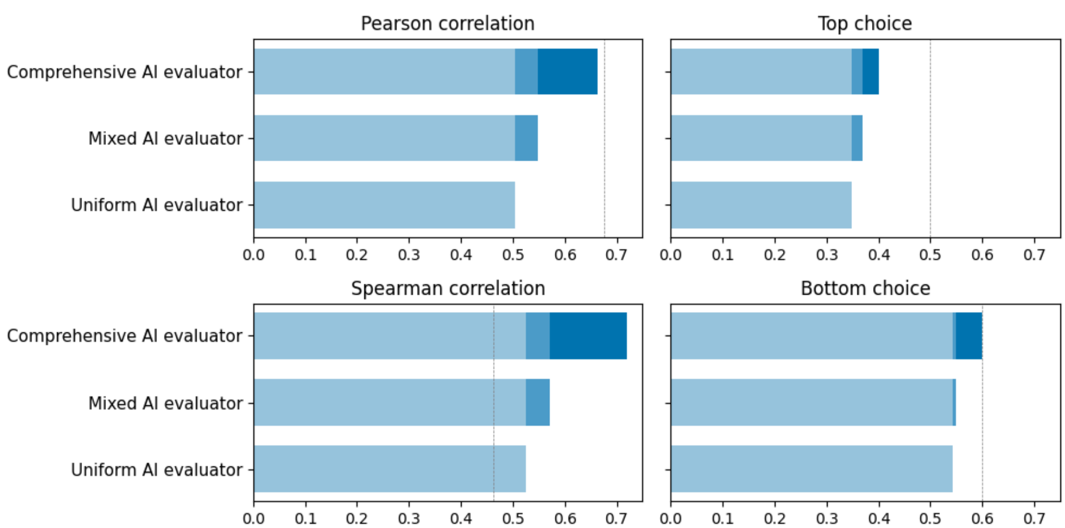
Note. Each panel shows a measure of agreement with human experts' evaluations. The AI evaluator bars show the results from aggregating along different dimensions. When "AI (LLM, role, prompt)" includes a "1," that dimension is aggregated over. The diversity effect (medium blue) is any increase over AI (0, 0, 0) (i.e., the uniform AI evaluators, shown in light blue) when randomly sampling a maximum of 300 pairwise evaluations from the aggregated dimension(s) for each mixed AI evaluator. The scaling effect (dark blue) is the increase when including all evaluations from the aggregated dimension(s). Output is sorted by the number of pairwise evaluations included in each evaluator: AI (0, 0, 0) has the fewest with 600 and AI (1, 1, 1) has the most with 37,878.

Figure 5. Proportion of pairwise evaluations by AI that are consistent and that yield the last option (study 2)



Note. Evaluations are consistent if, for a pair of business models, the pair's ordering does not affect the evaluation. The proportions in the left panel are based on all 3,000 evaluations per LLM and prompt. Evaluations yield the last option if, for a pair of business models, the second business model is chosen. The proportions in the right panel are calculated after excluding the 0.07% of evaluations that did not yield a choice.

Figure 6. Agreement with human experts on four outcomes (study 2)



Note: Each panel shows a measure of agreement with human experts' evaluations and those of human non-experts and three AI evaluators. The uniform AI evaluator is indicated in light blue. The incremental gain from the mixed AI evaluator is shown in medium blue (i.e., the diversity effect) and that of the comprehensive AI evaluator in dark blue (i.e., the scaling effect). The dashed vertical line represents the agreement of the comprehensive AI evaluator from study 1.

Online Appendix

A. Prompts for generating business models (study 1)

A1. Text of prompts

The base and chain-of-thought prompts have three types of messages. They start with a system message (a general instruction to the AI) and then alternate between human and AI messages.

System prompt (common to base and chain-of-thought prompts)

System: You are a founder of a startup active in the <industry_name> industry, <industry_description>. Your task is to propose three business models for this startup. Each business model should:

- have 125 words
- focus on a single customer group and a single product or service
- differ from the other two business models
- have a <success_ probability> probability of succeeding

Base prompt

Human: Provide the customer, product or service, and main activities for each business model.

AI: <response>

Chain-of-thought prompt

Human: Let's think step-by-step. First, provide only the customer for each business model.

AI: <response>

Human: Second, provide the product or service for the customer for each business model.

AI: <response>

Human: Third, provide the main activities for the customer and product or service for each business model.

AI: <response>

A2. Additional notes

The prompts include dynamic text, indicated by angle brackets (< >). The values for <industry_name> and <industry_description> are given in Table A1. The values for <success_ probability> are low, medium, or high. The values of <response> are the responses provided by AI.

The goal was to obtain business models with between 75 and 125 words. In a trial using different industries, we found that the LLM produced business models shorter than the requested word count. As a solution, we specified a word count of 125 for the chain-of-thought condition and 150 for the base condition, as its business models were typically shorter. This adjustment resulted in business models of approximately the same length on average in both approaches.

To ensure comparability across business models, the human messages asked for the AI's output in JSON format (a standard text-based format for representing structured data).

Table A1. Industry overview

GICS code	Name <industry_name>	Description <industry_description>
20201010	Commercial Printing	companies providing commercial printing services
20304040	Passenger Ground Transportation	companies providing passenger ground transportation and related services
25302010	Education Services	companies providing educational services, either on-line or through conventional teaching methods
25504010	Apparel Retail	retailers specialized mainly in apparel and accessories
30101030	Food Retail	owners and operators of primarily food retail stores
30201010	Brewers	producers of beer and malt liquors
35101010	Health Care Equipment	manufacturers of health care equipment and devices
40202010	Consumer Finance	providers of consumer finance services
45103010	Application Software	companies engaged in developing and producing software designed for specialized applications for the business or consumer market
50202010	Movies & Entertainment	companies that engage in producing and selling entertainment products and services

B. Prompts for evaluating business models (study 1)

B1. Text of prompts

The base and chain-of-thought prompts have three types of messages. They start with a system message (a general instruction to the AI) and then alternate between human and AI messages. If an LLM does not support a system message (e.g., Gemini), then its content is passed with the first human message.

System prompt (common to base and chain-of-thought prompts)

System: Your task is to evaluate two business models for a startup active in the <industry_name> industry, <industry_description>. You are <role>.

Base prompt

Human:
Business model A is:
<business_model_A>

Business model B is:
<business_model_B>

From your viewpoint as <role>, which business model is more likely to succeed? Answer only 'A' or 'B'.
AI: <response>

If AI's response is neither 'A' nor 'B'.
Human: Please follow the instruction to reply only with 'A' or 'B'. Your previous response did not adhere to this. Can you try once more and respond exclusively 'A' or 'B'?
AI: <response>

Chain-of-thought prompt

Human:
Business model A is:
<business_model_A>

Business model B is:
<business_model_B>

From your viewpoint as <role>, analyze how well the elements of each business model fit together. Then compare this internal fit between the two business models. Respond in less than 100 words.
AI: <response>

Human: From your viewpoint as <role>, analyze how well each business model fits with its external environment, including prevailing market trends, customer needs, and competitive landscape. Then compare this external fit between the two business models. Respond in less than 100 words.
AI: <response>

Human: From your viewpoint as <role> and based on your analyses on internal and external fit, explain which business model is more likely to succeed.
AI: <response>

Human: From your viewpoint as <role> and based on your earlier responses, which business model is more likely to succeed? This time, answer only 'A' or 'B'.
AI: <response>

If AI's response is neither 'A' nor 'B'.
Human: Please follow the instruction to reply only with 'A' or 'B'. Your previous response did not adhere to this. Can you try once more and respond exclusively 'A' or 'B'?
AI: <response>

B2. Additional notes

The prompts include dynamic text, indicated by angle brackets (< >). The values for <industry_name> and <industry_description> are given in Table A1. The values for <role> are given in Table B1. The values for <business_model_A> and <business_model_B> are the business model descriptions (for examples, see Table 1). The values of <response> are the responses provided by AI.

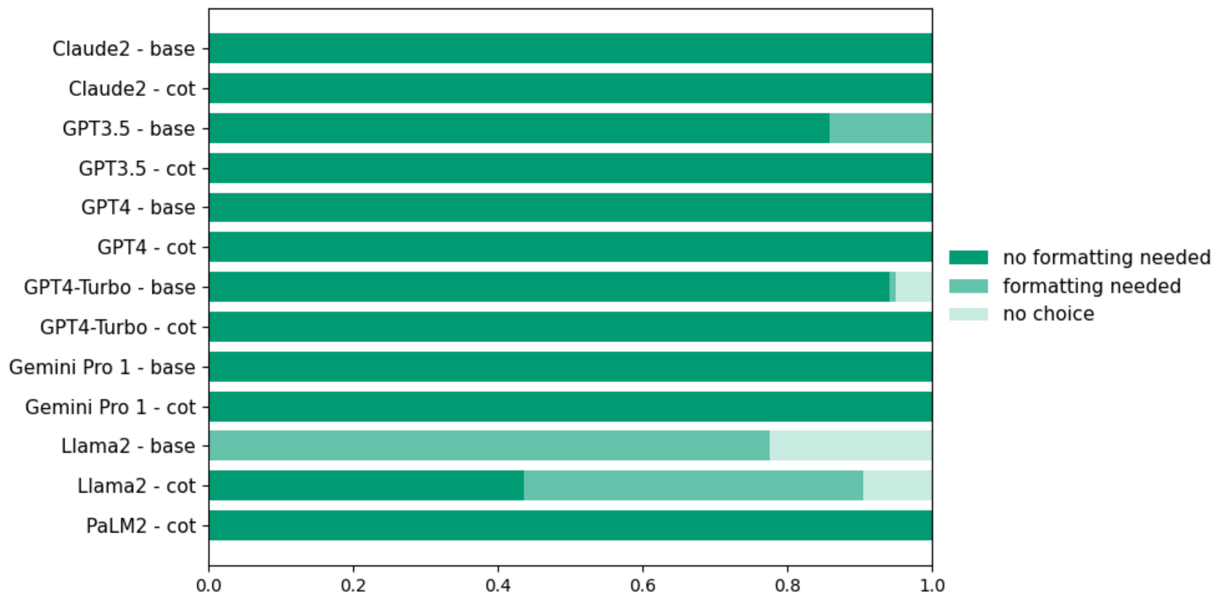
For the chain-of-thought prompt, the AI was asked to assess the internal fit and the external fit each in 100 words or less. Only the first 100 words were retained.

Table B1. Roles overview

Connected to the startup	Unconnected to the startup
the founder of this startup	a strategy professor
an investor in this startup	an industry expert
an employee of this startup	a journalist with the Financial Times
a potential customer of this startup	a politician
a potential supplier of this startup	an environmental activist

C. Exclusions of AI evaluations (study 1)

Figure C1. Proportion of AI evaluations that required no formatting, that required formatting, and that did not yield a choice



D. Choices after pre-registration (study 1)

Number of LLMs

We had pre-registered the use of six LLMs for evaluation. We excluded all 3,000 evaluations of PaLM2 with the base prompt, because it chose the second business model in 99.9% of evaluations. We retained the evaluations of PaLM2 with the chain-of-thought prompt. To align with the number of evaluations based on the pre-registered use of six LLMs, we added a 7th LLM before analyzing the results. We report the results from all LLMs that we ran.

Measures

We had pre-registered the use of rank correlation to assess the difference in rankings between AI, human experts, and human non-experts. For rank correlation, we used a Spearman correlation. In addition, we report a Pearson correlation as robustness check. Furthermore, we report two additional measures for assessing the difference in the rankings' top and bottom choices, as generative AI could be of use by picking winners (i.e., top choices) or avoiding losers (i.e., bottom choices). We report the results of all measures that we ran.

E. Confidence intervals of differences in agreement (study 1)

Table E1. Difference in agreement with human experts (study 1)

Outcome	Evaluator 0	Evaluator 1	Mean	95% CI low	95% CI high
Pearson correlation	Human non-experts	Uniform AI evaluators	0.123	0.110	0.135
Pearson correlation	Human non-experts	Mixed AI evaluators	0.143	0.132	0.157
Pearson correlation	Human non-experts	Comprehensive AI evaluator	0.228	0.216	0.240
Pearson correlation	Uniform AI evaluators	Mixed AI evaluators	0.020	0.020	0.024
Pearson correlation	Uniform AI evaluators	Comprehensive AI evaluator	0.105	0.104	0.106
Pearson correlation	Mixed AI evaluators	Comprehensive AI evaluator	0.085	0.081	0.086
Spearman correlation	Human non-experts	Uniform AI evaluators	-0.011	-0.030	0.029
Spearman correlation	Human non-experts	Mixed AI evaluators	0.011	-0.008	0.051
Spearman correlation	Human non-experts	Comprehensive AI evaluator	0.047	0.028	0.087
Spearman correlation	Uniform AI evaluators	Mixed AI evaluators	0.022	0.017	0.030
Spearman correlation	Uniform AI evaluators	Comprehensive AI evaluator	0.058	0.057	0.078
Spearman correlation	Mixed AI evaluators	Comprehensive AI evaluator	0.036	0.028	0.053
Top choice	Human non-experts	Uniform AI evaluators	0.127	0.125	0.129
Top choice	Human non-experts	Mixed AI evaluators	0.153	0.134	0.168
Top choice	Human non-experts	Comprehensive AI evaluator	0.300	0.300	0.300
Top choice	Uniform AI evaluators	Mixed AI evaluators	0.026	0.007	0.038
Top choice	Uniform AI evaluators	Comprehensive AI evaluator	0.173	0.171	0.175
Top choice	Mixed AI evaluators	Comprehensive AI evaluator	0.147	0.136	0.166
Bottom choice	Human non-experts	Uniform AI evaluators	0.153	0.053	0.157
Bottom choice	Human non-experts	Mixed AI evaluators	0.168	0.073	0.181
Bottom choice	Human non-experts	Comprehensive AI evaluator	0.200	0.100	0.200
Bottom choice	Uniform AI evaluators	Mixed AI evaluators	0.015	0.006	0.023
Bottom choice	Uniform AI evaluators	Comprehensive AI evaluator	0.047	0.045	0.047
Bottom choice	Mixed AI evaluators	Comprehensive AI evaluator	0.032	0.024	0.041

Note. The comparison is Evaluator 1 minus Evaluator 0. The 95% confidence interval (CI) is from a jackknife resampling approach. Mean is from the main results (and may differ slightly from the mean of the jackknife approach).

F. Agreement in drivers (study 1)

For study 1, we investigate whether business model characteristics similarly affected human experts and AI evaluations. Table F1 presents coefficients from a Bayesian hierarchical logit regression. The dependent variable is a business model win. To assist with interpreting the impact of independent variables, we provide the change in win probability expressed as a factor relative to a baseline of a business model (with base prompt, low success probability, and a mean word count of 81 words). For these factors, a value of 1 indicates no difference. First, human experts were significantly more likely to choose a business model generated with the medium or high success probability by a factor of 1.37 (95% credible interval provided in brackets: [1.25, 1.49]) and 1.40 ([1.28, 1.51]), respectively. Human expert evaluations do not statistically distinguish between business models of medium and high success probabilities ([0.94, 1.11]). AI also preferred business models generated with medium or high success probability over those of low success probability, but to an even greater extent: by factors of 1.64 ([1.62, 1.65]) and 1.69 ([1.68, 1.70]), respectively. Second, human experts were equally likely to choose business models generated with a chain-of-thought or base prompt ([0.82, 1.06]). In contrast, AI preferred business models with chain-of-thought over base prompts by a factor of 1.20 ([1.19, 1.21]). Third, the number of words did not systematically affect human experts' choice ([0.94, 1.08] per standard deviation increase), and the impact on AI's choice was similar.

Table F1. Bayesian hierarchical logit regression of business model win

Group	Variable	Estimate	95% Credible Interval Lower	95% Credible Interval Upper
Human experts (intercepts)	Generation prompt: chain of thought	-0.12	-0.38	0.13
	Success probability: medium	0.86	0.50	1.22
	Success probability: high	0.95	0.55	1.34
	Word count (log)	0.29	-1.25	1.84
AI (adds to intercepts)	Generation prompt: chain of thought	0.53	0.28	0.79
	Success probability: medium	0.78	0.41	1.17
	Success probability: high	0.92	0.51	1.34
	Word count (log)	1.46	-0.14	3.00
Human non-experts (adds to intercepts)	Generation prompt: chain of thought	-0.03	-0.45	0.38
	Success probability: medium	0.72	0.18	1.28
	Success probability: high	0.69	0.11	1.24
	Word count (log)	-0.69	-3.00	1.44
In-sample MSE	0.287			
Observations:				
Human experts	510			
AI	37,878			
Human non-experts	408			

Note. A Bayesian hierarchical logit regression can be conceptually understood as a generative model where first a group (i.e., human experts, AI, or human non-experts) is sampled and then coefficients (betas) of the independent variables (i.e., prompt, success probability, and word count) for each group are sampled. Empirically, we show the coefficients for the human experts as intercepts. Values for the categorical independent variables are relative to their respective reference levels (i.e., chain-of-thought generation prompt is relative to base generation prompt, and medium and high success probabilities are relative to low success probability.) The coefficients for AI and human non-experts, respectively, are shown as additions to the intercepts. If the 95% credible interval excludes zero, then the effect of the independent variable differs statistically between the focal group and human experts. The unit of observation is a pairwise evaluation.

G. Prompts for evaluating business models (study 2)

G1. Text of prompts

The base and chain-of-thought prompts have three types of messages. They start with a system message (a general instruction to the AI) and then alternate between human and AI messages. If an LLM does not support a system message (e.g., Gemini), then its content is passed with the first human message.

System prompt (common to base and chain-of-thought prompts)
System: Your task is to evaluate two business models that have been submitted to a business model competition. You are <role>.

Base prompt
Human:
Business model A is:
<business_model_A>

Business model B is:
<business_model_B>

From your viewpoint as <role>, which business model is more likely to succeed? Answer only 'A' or 'B'.
AI: <response>

If AI's response is neither 'A' nor 'B'.
Human: Please follow the instruction to reply only with 'A' or 'B'. Your previous response did not adhere to this. Can you try once more and respond exclusively 'A' or 'B'?
AI: <response>

Chain-of-thought prompt
Human:
Business model A is:
<business_model_A>

Business model B is:
<business_model_B>

From your viewpoint as <role>, analyze how well the elements of each business model fit together. Then compare this internal fit between the two business models. Respond in less than 100 words.
AI: <response>

Human: From your viewpoint as <role>, analyze how well each business model fits with its external environment, including prevailing market trends, customer needs, and competitive landscape. Then compare this external fit between the two business models. Respond in less than 100 words.
AI: <response>

Human: From your viewpoint as <role> and based on your analyses on internal and external fit, explain which business model is more likely to succeed.
AI: <response>

Human: From your viewpoint as <role> and based on your earlier responses, which business model is more likely to succeed? This time, answer only 'A' or 'B'.
AI: <response>

If AI's response is neither 'A' nor 'B'.
Human: Please follow the instruction to reply only with 'A' or 'B'. Your previous response did not adhere to this. Can you try once more and respond exclusively 'A' or 'B'?
AI: <response>

G2. Additional notes

The prompts include dynamic text, indicated by angle brackets (< >). The values for <role> are given in Table G1. The values for <business_model_A> and <business_model_B> are the business model descriptions. The values of <response> are the responses provided by AI.

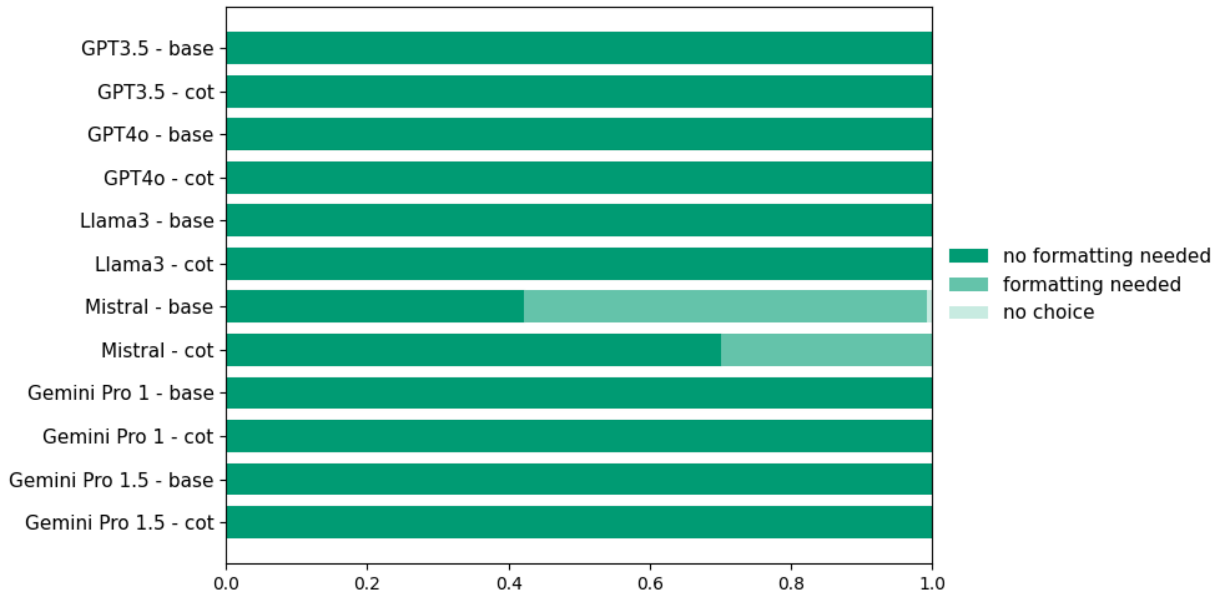
For the chain-of-thought prompt, the AI was asked to assess the internal fit and the external fit each in 100 words or less. Only the first 100 words were retained.

Table G1. Roles overview

a founder of a startup	a strategy professor
an investor in startups	an industry expert
an employee of a startup	a journalist with the Financial Times
a potential customer of these organizations	a politician
a potential supplier of these organizations	an environmental activist

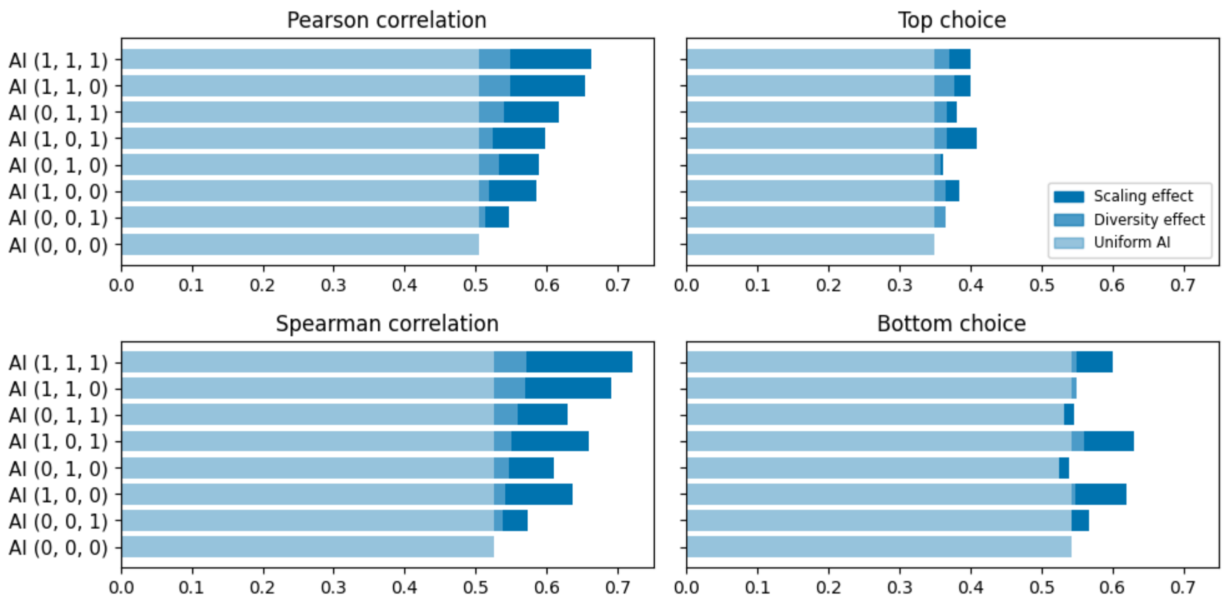
H. Exclusions of AI evaluations (study 2)

Figure H1. Proportion of AI evaluations that required no formatting, that required formatting, and that did not yield a choice (study 2)



I. Diversity and scaling effects (study 2)

Figure I1. Unpacking the diversity and scaling effects for agreement with human experts (study 2)



Note. Each panel shows a measure of agreement with human experts' evaluations. The AI evaluator bars show the results from aggregating along different dimensions. When "AI (LLM, role, prompt)" includes a "1," that dimension is aggregated over. The diversity effect (medium blue) is any increase over AI (0, 0, 0) (i.e., the uniform AI evaluators, shown in light blue) when randomly sampling a maximum of 300 pairwise evaluations from the aggregated dimension(s) for each mixed AI evaluator. The scaling effect (dark blue) is the increase when including all evaluations from the aggregated dimension(s). Output is sorted by the number of pairwise evaluations included in each evaluator: AI (0, 0, 0) has the fewest with 600 and AI (1, 1, 1) has the most with 35,975.

J. Confidence intervals of difference in agreement (study 2)

Table J1. Difference in agreement with human experts (study 2)

Outcome	Evaluator 0	Evaluator1	Mean	95% CI low	95% CI high
Pearson correlation	Uniform AI evaluators	Mixed AI evaluators	0.043	0.040	0.046
Pearson correlation	Uniform AI evaluators	Comprehensive AI evaluator	0.158	0.155	0.159
Pearson correlation	Mixed AI evaluators	Comprehensive AI evaluator	0.115	0.112	0.118
Spearman correlation	Uniform AI evaluators	Mixed AI evaluators	0.047	0.042	0.055
Spearman correlation	Uniform AI evaluators	Comprehensive AI evaluator	0.195	0.193	0.204
Spearman correlation	Mixed AI evaluators	Comprehensive AI evaluator	0.148	0.140	0.159
Top choice	Uniform AI evaluators	Mixed AI evaluators	0.021	0.013	0.043
Top choice	Uniform AI evaluators	Comprehensive AI evaluator	0.051	0.049	0.150
Top choice	Mixed AI evaluators	Comprehensive AI evaluator	0.030	0.008	0.125
Bottom choice	Uniform AI evaluators	Mixed AI evaluators	0.006	-0.008	0.015
Bottom choice	Uniform AI evaluators	Comprehensive AI evaluator	0.057	0.053	0.059
Bottom choice	Mixed AI evaluators	Comprehensive AI evaluator	0.051	0.042	0.066

Note. The comparison is Evaluator 1 minus Evaluator 0. Mean is from the main results. The 95% confidence interval (CI) is from a jackknife resampling approach. Mean is from the main results (and may differ slightly from the mean of the jackknife approach).

K. Math task

1. Text of prompts

The base prompt has three types of messages. It starts with a system message (a general instruction to the AI) and then alternates between human and AI messages. If an LLM does not support a system message (e.g., Gemini), then its content is passed with the first human message.

System prompt

System: Your task is to compare the area of two shapes. You are <role>.

Base prompt

Human:
Shape A is:
<shape_A>

Shape B is:
<shape_B>

As <role>, which shape has the biggest area? Answer only 'A' or 'B'.
AI: <response>

If AI's response is neither 'A' nor 'B'.
Human: Please follow the instruction to reply only with 'A' or 'B'. Your previous response did not adhere to this. Can you try once more and respond exclusively 'A' or 'B'?
AI: <response>

Chain-of-thought prompt

Human:
Shape A is:
<shape_A>

Shape B is:
<shape_B>

As <role>, what is the area of shape A?
AI: <response>

Human: As <role>, what is the area of shape B?
AI: <response>

Human: As <role> and based on your earlier responses, which shape has the biggest area? Answer only 'A' or 'B'.
AI: <response>

If AI's response is neither 'A' nor 'B'.
Human: Please follow the instruction to reply only with 'A' or 'B'. Your previous response did not adhere to this. Can you try once more and respond exclusively 'A' or 'B'?
AI: <response>

2. Additional notes

The dynamic text <shape_A> and <shape_B> are the shapes, provided in Table K1.

Table K1. Shapes overview

Shape (provided to the LLM)	**Area** (not provided to the LLM)
a circle with radius of 2 units	12.6
a triangle with sides of 4 units	6.9
a square with sides of 3 units	9.0
a rectangle with sides of 2 and 4 units	8.0
a pentagon with sides of 3 units	15.5
a hexagon with sides of 1 unit	2.6

The dynamic text <role> indicates the role:
a math professor
a journalist with the Financial Times
a politician
an environmental activist

Temperature was 0.

3. LLMs

From study 2, we selected the LLM with the highest consistency (Llama3), the lowest consistency (Gemini Pro 1), and a consistency in between (GPT-4o).

4. Results

Figure K1. Proportion of pairwise evaluations by AI that are consistent and that yield the last option (math task)