Artificial Agents and the Evaluation of M&As

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

Marco Testoni
University of Miami
Miami Herbert Business School
marco.testoni@miami.edu

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

January 2025

Acknowledgment: We are grateful for the feedback of Isabel Fernandez-Mateo, Jen Rhymer, and Wei Shi.

Abstract

Mergers and acquisitions (M&As) offer opportunities for value creation through synergies but often result in value destruction due to imperfect managerial evaluation and decision-making processes. Recent advances in generative artificial intelligence (AI) and large language models (LLMs) present novel opportunities to support M&A evaluation processes. This study investigates the ability of artificial evaluators to discriminate between value-creating and value-destroying M&As. In particular, we focus on LLM agents, which can increasingly operate autonomously. Four evaluator conditions with varying levels of autonomy are examined: (1) a baseline model without step-by-step reasoning, (2) chain-of-thought reasoning whose steps are controlled by the human, (3) autonomous reasoning whose steps are controlled by an LLM agent, and (4) autonomous reasoning involving interactions among multiple LLM agents. Using anonymized announcements of 109 deals between U.S. public firms, our findings reveal that only the multi-agent condition demonstrates a substantive ability to differentiate between value-creating and value-destroying M&As. The top half of deals identified as most promising in the multi-agent condition corresponds to abnormal returns at the deal announcement totaling \$571 million, while the bottom half with the least promising deals incurs losses of \$1,273 million. These results underscore the potential of multi-agent systems in enhancing strategic decision making in M&As.

1. Introduction

Mergers and acquisitions (M&As) are key tools in firms' corporate strategies (Villalonga and McGahan 2005, Wang and Zajac 2007). By combining resources and capabilities, they allow firms to create shareholder value due to synergies (e.g., Chatterjee 1986, Larsson and Finkelstein 1999, Ahuja and Katila 2001, Feldman and Hernandez 2022). Yet, a significant proportion of M&As destroy shareholder value (Haleblian et al. 2009, Graebner et al. 2017). These losses frequently occur because managers overestimate synergies or underestimate the costs of achieving them (e.g., Hayward and Hambrick 1997, Larsson and Finkelstein 1999, Malmendier and Tate 2008, Karim and Kaul 2015, Graebner et al. 2017). Thus, the evaluation of a potential acquisition is pivotal for creating value in M&As.

Given the often disappointing results of M&As, research has investigated the decision-making process leading to these deals focusing on the people involved. At the individual level, studies have highlighted how personal factors may induce managers to engage in value-destroying deals, including biases such as managerial overconfidence (Hayward and Hambrick 1997, Malmendier and Tate 2008, Rogan and Sorenson 2014) and cognitive limitations affecting the processing of information (Haleblian and Finkelstein 1999, Hayward 2002, Castellaneta et al. 2018). At the group level, scholars have shown that acquisition decisions depend on the division of cognitive labor among decision makers (Lou et al. 2024) and their attention towards specific valuation practices and synergy types (Bauer and Friesl 2024). The critical role of managers' decision models (Pablo 1994) is underscored by the fact that the characteristics of CEOs and internal M&A teams strongly predict acquirer returns following M&A announcements (Meyer-Doyle et al. 2019, Aktas et al. 2021).

The recent advances in generative artificial intelligence (AI), and particularly those in large language models (LLMs), offer the potential for artificial evaluators to play a role in M&A decision making, historically under the exclusive purview of human agents. In this study, we investigate artificial evaluators' ability to assess M&A decisions. We are particularly interested in LLM agents, which can increasingly operate autonomously. Thus, agents are agentic: they have the capacity to think, plan, and act (Gray et al. 2007, Gray and Wegner 2012, Vanneste and Puranam 2024). For LLM agents, it is the LLM that functions “as the primary component of brain or controller” of an agent giving it abilities to reason, plan, and interact with an environment (Xi et al. 2023: 8). In the context of M&A decision making, LLM agents have the potential for self reflection, allowing them to evaluate their own decision-making processes, and interact with other LLM agents. In contrast to earlier usages of LLMs, the LLM agent operates autonomously or “exercises control over its own actions” (Franklin and Graesser 1997: 29).

We investigate different types of artificial evaluators focusing on their ability to discriminate between M&As that created shareholder value and those that did not (Christensen and Knudsen 2010, Csaszar 2013), using announcements—public disclosures detailing the deal’s strategic rationale—of 109 M&As. The artificial evaluators are grouped into four conditions with varying levels of autonomy: no step-by-step reasoning (base), step-by-step reasoning whose steps are controlled by the human (chain-of-thought), autonomous step-by-step reasoning whose steps are controlled by an LLM agent (reflection), and autonomous step-by-step reasoning involving interactions among multiple LLM agents (multi-agent). This stepwise setup allows for the understanding of the discriminative ability of artificial evaluators and of the extent to which agency matters. We used anonymized versions of the original announcement—which include descriptions of the deal’s strategic rationale but omit companies’ identities—to focus on their reasoning abilities, not their capacity to memorize companies’ deals. The main finding is that, whereas the artificial evaluators in the base, chain-of-thought, and reflection conditions are unable to distinguish between value-creating and value-destroying deals, those in the multi-agent condition are substantively able to tell apart “good” from “bad” M&As. Specifically, for the multi-agent condition, a one-standard-deviation more positive assessment of a deal increases the probability by 16 percentage points that a deal is value-creating rather than value-destroying. Furthermore, the 54 M&A deals deemed most promising show cumulative abnormal returns totaling $571 million, whereas the 55 M&A deals identified as least promising incur losses of $1,273 million.

We highlight two contributions. First, to the established M&A literature (e.g., Haleblian et al. 2009, Welch et al. 2020), we introduce and empirically validate the use of artificial agents in evaluating M&A decisions. This represents a shift from traditional reliance on human evaluators (Pablo 1994, Meyer-Doyle et al. 2019, Aktas et al. 2021, Bauer and Friesl 2024, Lou et al. 2024), whose judgments are often influenced by biases and cognitive limitations (Hayward and Hambrick 1997, Haleblian and Finkelstein 1999, Hayward 2002, Rogan and Sorenson 2014, Malmendier and Tate 2008, Castellaneta et al. 2018), to artificial evaluators capable of systematic and scalable assessments. Notwithstanding their own limitations, we show that artificial agents have the ability to discriminate between value-creating and value-destroying deals based on a brief description of a deal's strategic rationale. This suggests a potential role for such artificial agents in the evaluations of M&A decisions. Second, to the emerging literature on AI and strategic decision making, this study adds agency to artificial evaluators. While prior work has demonstrated the promise of artificial evaluators for strategic decisions (e.g., Boussioux et al. 2024, Csaszar et al. 2024, Doshi et al. 2024), it has relied on non-agentic approaches. We find that the inclusion of agency—such as autonomously choosing whom to interact with or what to focus on—leads to a significant improvement in decision-making performance. Agentic systems not only enhance decision-making performance but also align closely with the demands of strategic decision making, where the ability to think, reflect, and act dynamically plays a decisive role (Eisenhardt and Zbaracki 1992).

2. Background

2.1. M&A screening

M&As can create value for shareholders by enabling merging firms to combine their resources and capabilities into a single entity. The literature on M&As highlights several reasons why the combined entity may perform better than the sum of the two individual firms. These include gains in market power in input or output markets (e.g., Fee and Thomas 2004, Clougherty and Duso 2009, Arts et al. 2025), efficiency due to operational or financial synergies (e.g., Chatterjee 1986, Larsson and Finkelstein 1999, Ahuja and Katila 2001), enhanced positions in strategic alliance networks (e.g., Feldman and Hernandez 2022), elimination of transaction costs (e.g., Frésard et al. 2020), transfers of capabilities (e.g., Capron and Pistre 2002, Kaul and Wu 2016), resource redeployment (e.g., Dickler and Folta 2020), and harming rivals by foreclosing access to key suppliers or customers (e.g., Lafontaine and Slade 2007). These potential benefits must be weighed against the costs of these transactions, which can be steep. Other than the direct financial costs of deals, which include the cost of financing, advisory fees, and legal costs, M&As can impose high integration costs (e.g., Larsson and Finkelstein 1999, Graebner et al. 2017). Cultural and operational incompatibilities between merging firms can cause major disruptions and ultimately destroy firm value (e.g., Chatterjee et al. 1992, Weber and Camerer 2003, Karim and Kaul 2015).

Due to these costs, M&As often lead to reductions in firms' value. On average, acquirers lose stock market value at the deal announcement, while targets gain due to the takeover premium paid by the acquirer (e.g., Haleblian et al. 2009, Cai and Sevilir 2012). While the merging firms' combined stock market returns from the deal are, on average, slightly positive, a substantial fraction of deals reduces the combined value of the two firms. These value-destroying M&As can harm not only shareholders but also other stakeholders, including the firms' employees, customers, and suppliers.

Given the frequent disappointing outcomes of M&As, scholars have investigated the decision-making process behind these transactions to uncover the flaws that may lead managers to select value-destroying deals. A first class of explanations for poor M&A decisions relates to the private incentives of managers to engage in these transactions (Jensen 1986, Harford 1999, Feldman et al. 2019). Because the compensation of managers typically increases after an acquisition (e.g., Bliss and Rosen 2001, Harford and Li 2007), managers have an incentive to make acquisitions even when these destroy company value.[1] A second class of explanations relates to the availability of information necessary to evaluate the deal. Specifically, the lack of adequate information may lead managers to engage in worse deals (e.g., Uysal et al. 2008, Cai and Sevilir 2012). Finally, a third class of explanations relates to the cognitive limitations of managers, which may lead to flawed decisions even in the absence of shareholder-manager incentive misalignment or imperfect information to evaluate deals. For instance, scholars have shown that managers are often overconfident in their ability to generate synergies from

[1] Moreover, Antón et al. (2022) show that the acquirer's shareholders may not oppose the management when they own stakes in the target or in non-merging firms, which could benefit from bad deals. This suggests that shareholder monitoring can often fail to correct managerial decisions.

M&As (Hayward and Hambrick 1997, Malmendier and Tate 2008). Managers may also hold positively biased beliefs about the potential of certain targets. Rogan and Sorenson (2014) show that this can occur with targets that share social connections with the acquirer. Managers' learning from past acquisition experience can also play a role (Haleblian and Finkelstein 1999, Hayward 2002, Castellaneta et al. 2018). While managers may learn to better screen and manage transactions, they may also inappropriately generalize acquisition experience to subsequent dissimilar acquisitions, with detrimental effects on acquisition performance (Haleblian and Finkelstein 1999). Moreover, managers' assessment of deals reflects their background and perception of synergies. For instance, Aktas et al. (2021) find that firms with internal teams specialized in M&As undertake better deals. Moreover, they find that teams with more financial experience and who emphasize the economic rationales of deals select better deals, while those who focus on behavioral rationales (such as the friendliness of the target's managers or reactions to rivals' initiatives) select worse deals. Meyer-Doyle et al. (2019) report that CEO-level factors account for a substantial share of the variance in acquisition performance. Chen et al. (2021) show that generalist CEOs are more likely to engage in unrelated acquisitions than specialist CEOs, and the fit between the nature of CEOs' human capital and the type of acquisitions leads to better acquisition performance. Bauer and Friesl (2024) further emphasize that managers' perception of deal potential reflects their attention toward specific valuation practices and synergy types at the expense of others. Finally, Lou et al. (2024) point out that M&A decision processes involve the entire top management team (TMT), rather than solely the CEO. Thus, the decision outcome is a function of the shared division of cognitive labor in top management teams. The performance of acquisitions therefore depends on the effective integration of TMT members' knowledge.

Overall, M&A deals are inherently complex strategic actions. The literature has shed light on several factors that can explain why some deals increase the value of merging firms while others reduce it. Yet, it is important to note that each individual factor plays a limited role in explaining the variance of deal outcomes. Indeed, a recent meta-analysis documents that most factors investigated in prior research correlate with less than |.1| with stock market outcomes (and similar correlations are found for other outcomes) (King et al. 2021). Moreover, the R^2 of regression models predicting the stock market returns of merging firms from the deal is typically in the single digit (e.g., Uysal et al. 2008, Hoberg and Phillips 2010, Eckbo et al. 2018). This low predictability suggests that the decision-making process in M&As is likely to improve with the adoption of decision tools that can take a holistic view in assessing the deal potential and overcome cognitive biases.

2.2. Artificial evaluators

Artificial evaluators have shown promise in research on strategy decision making (e.g., Boussioux et al. 2024, Csaszar et al. 2024, Doshi et al. 2024), which focuses on LLMs. We highlight that LLMs can increasingly operate autonomously as LLM agents.

2.2.1. Large language models

The fundamental operation of LLMs is to predict the next word given a sequence of words (Murphy 2023). They are trained on vast amounts of text (Brown et al. 2020), typically using a Transformer architecture (Vaswani et al. 2017). This architecture enables models to learn complex relationships between words, even if placed apart across different sentences. Through training, they are able to gradually learn the mechanics of language, and differences in context and meaning, allowing for their usage in a variety of tasks including text production, document summarization, language translation, code generation, and a range of reasoning tasks (Bubeck et al. 2023).

One of the major trends that shapes current LLM development is the scaling hypothesis: the larger the model and its training data are, the better the model's performance is (Brown et al. 2020). This has prompted a race to develop bigger and more capable models among major LLM providers. Newer models not only improve upon previous ones in terms of performance but also broaden the range of tasks they can handle.[2]

Another major trend is about the usage of LLMs. Initially, prompt engineering emerged as a pivotal technique, where users craft specific inputs to guide models toward desired outputs. For example,

---

[2] An up-to-date benchmark of LLMs across multiple criteria can be found at https://livebench.ai/ (accessed January 29, 2025).

chain-of-thought prompting is a popular technique that uses sequential instructions to break down complex tasks into simpler, sequential tasks (Wei et al. 2022a). However, as models have grown more capable, the focus has begun shifting from static, user-crafted prompts to dynamic, system-driven approaches. This emerging approach emphasizes the model's ability to autonomously navigate tasks, adapt to context, and refine responses in real time. For instance, models can now engage in iterative self-reflection to assess and improve their responses, such as correcting errors in a reasoning task or optimizing a piece of generated code (Zhang et al. 2023). In doing so, they demonstrate the behaviors of artificial agents (Park et al. 2023, Wang et al. 2024).

2.2.2. LLM agents

An artificial agent is "a computer system situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives", and autonomy implies that "the system should be able to act without the direct intervention of humans (or other agents), and should have control over its own actions and internal state" (Jennings and Wooldridge 1998: 4). This aligns with a broader understanding of agency, which refers to the capacity of an entity to think, plan, and act (Gray et al. 2007, Gray and Wegner 2012, Schlosser 2019). Especially with the developments in AI (Vanneste and Puranam 2024), the concept of agent has been extended from human actors to computational entities (Wooldridge and Jennings 1995, Wu et al. 2023).

Building on this foundation, LLM agents represent a significant evolution in the concept of artificial agents (Xi et al. 2023). These agents leverage the capabilities of LLMs to perform complex reasoning, natural language understanding, and decision-making tasks (Bubeck et al. 2023, Wei et al. 2022b). LLMs align with the essence of agency because of their ability to act without the direct intervention of others. This autonomy forms the basis for a growing number of applications. For instance, LLM agents can write reports based on independently conducted internet research (e.g., STORM (Shao et al. 2024), GPT Researcher[3]), develop software, including writing and testing code (e.g., Meta-GPT (Hong

[3] Available at: https://github.com/assafelovic/gpt-researcher (accessed January 29, 2025)

et al. 2023), Aider[4]), and simulate complex multi-agent interactions and behaviors for collaborative problem-solving (e.g., CAMEL (Li et al. 2023), TinyTroupe[5]). In these applications, LLM agents exhibit a degree of autonomy to help them navigate complex tasks.

2.2.3. Autonomy dimensions

We can analyze LLM agents' autonomy based on the underlying dimensions over which they exert control. The information processing perspective (Simon 1947, Galbraith 1973, Tushman and Nadler 1978, Van Knippenberg et al. 2015, Joseph and Gaba 2020), a dominant view on organizational decision making, identifies two relevant dimensions: information content and information flow. This perspective views an organization as a collection of human agents that gather, interpret, and synthesize information. A key interest lies in understanding what information is transmitted and between whom (Thompson 1967, Puranam et al. 2012). Applying this perspective to LLM agents highlights how their autonomy can be understood in terms of their control over the information content and flow.

The first, information content, addresses the nature of the information shared. It refers to the type and relevancy of information shared and utilized in decision-making (Tushman and Nadler 1978). For example, information may include both objective indicators (e.g., in the form of figures, metrics, or text data), alongside subjective evaluations or judgments (e.g., inferences or strategic implications) (Joseph and Gaba 2020). By selecting or omitting certain information and shaping the framing of issues, decision makers can significantly influence how organizational members perceive problems and identify solutions (Van Knippenberg et al. 2015). In the context of M&As, content can include information related to the industry, products, synergy opportunities, integration costs, and strategic rationale (Larsson and Finkelstein 1999, Ahuja and Katila 2001, Graebner et al. 2017). LLM agents exercising autonomy over information content can decide which information to generate, prioritize, or ignore.

The second, information flow, describes the pattern of interactions between individuals. These patterns can be formal, as mandated by hierarchical reporting lines, job descriptions, or standard operating

---

[4] Available at: https://github.com/Aider-AI/aider (accessed January 29, 2025)
[5] Available at: https://github.com/microsoft/tinytroupe (accessed January 29, 2025)

procedures (Thompson 1967, Nadler and Tushman 1997). Alternatively, they can be informal, as in emergent social networks and ad hoc relationships that arise through personal connections, mutual interests, or historical ties (McEvily et al. 2014). In the context of M&As, information flow entails the movement of information within a dedicated M&A team or TMT team, or between decision makers and other parts of the organization (Aktas et al. 2021, Lou et al. 2024). LLM agents can exercise autonomy over the information flow by deciding which parties to engage with and which to exclude.

2.3. Artificial evaluators and M&As

M&As, like other strategic decisions, are difficult to undo and inherently uncertain (Eisenhardt and Zbaracki 1992, Van den Steen 2018). Given the difficulty of undoing, decision making in such situations can be assessed by conceptualizing a process where an evaluator faces a binary choice. This choice involves either accepting or rejecting a strategic proposal (e.g., to move forward with or abandon a potential acquisition), with one option generating greater value than the other (Knudsen and Levinthal 2007, Christensen and Knudsen 2010, Csaszar 2013). In the context of M&As, assessing whether a potential acquisition will create value involves weighing the anticipated synergies of the deal against its costs and risks (Puranam and Vanneste 2016, Rabier 2017). The reference point for these evaluations is typically a comparison to the alternative of not pursuing the deal. For example, prior research has often focused on cumulative abnormal returns (CARs) on the merging firms' stock to understand the difference between pursuing a deal and not pursuing it, with a positive CAR indicating that the deal generates greater value (King et al. 2004).

Evaluators differ in their ability to distinguish value-creating from value-destroying alternatives. Their task is to recommend value-creating deals and reject value-destroying deals, while minimizing two errors: recommending a value-destroying deal (a commission error) and rejecting a value-creating deal (an omission error). For comparing evaluators, it is useful to take a probabilistic perspective, which involves observing their performance across multiple instances rather than judging individual decisions in isolation (Knudsen and Levinthal 2007, Christensen and Knudsen 2010, Csaszar 2013). This perspective recognizes that even a good evaluator is not immune to errors and a bad evaluator need not always err. By doing so, it takes into consideration that strategic decisions are made under uncertainty.

3. Methods

3.1. Data

Data on M&A deals are from the Securities Data Corporation (SDC) database. The sample includes M&A deal announcements between public firms in the United States announced in the period 2019–2023, excluding minority-stake acquisitions.[6] As it is common in the literature (Uysal et al. 2008, Savor and Lu 2009, Cai and Sevilir 2012), we eliminated small and economically insignificant deals, retaining only those with a minimum value of $10 million (in 2023 USD) and representing at least five percent of the acquirer's market capitalization 50 trading days prior to the announcement. We also excluded financial acquirers and targets (with primary Standard Industrial Classification [SIC] codes from 60 to 69). Stock market data are from CRSP and firms' fundamentals are from Compustat. For each deal, we obtained from EDGAR the 8K form of the deal filed by the acquirer and the press release of the announcement from the appendix of the 8K.[7] We included deals where the 8K was filed within four days from the announcement date in SDC. With these filters, we obtained 112 deal announcements. We reviewed the press releases and excluded three cases for which the press release did not provide information on the strategic rationale of the deal. Our final sample includes 109 deal announcements.

We used rewritten versions of the original announcements for two reasons. First, an LLM may have been trained on news articles mentioning whether a specific deal was successful. Our aim is to assess the reasoning abilities of the artificial evaluators, not their capacity for memorization of specific deals. Second, announcements tend to be written positively because companies benefit when their announcement is well received. An announcement's tone may affect an artificial evaluator's assessment of

[6] We select deals in SDC with "form of the deal" = "merger" or "acquisition of majority interest." The sample includes both completed and non-completed deals since at the time of the acquisition announcement the completion outcome is unknown.
[7] In four cases, the press release was not in the appendix of the 8K and we obtained it from the acquirer's investor relations website.

the deal. We used GPT 4o from OpenAI to rewrite each announcement in a neutral tone and without identifying information such as company names or dates (see the Online Appendix for the prompt). Each announcement was rewritten five times.[8] We used multiple versions of the same original announcement to ensure that no single rewrite would drive the results. This yielded 545 rewritten announcements, with an average of 960 words (s.d.=406).

3.2. Conditions

The task of the artificial evaluators is to evaluate the M&A deals by predicting the probability of positive CAR of both companies combined for a 7-day period (centered on the day of the announcement). The artificial evaluators received a rewritten deal announcement and instructions for the prediction task based on Halawi et al. (2024), which included explicit requirements on the output format.

We use four evaluator conditions with varying levels of autonomy (see Table 1): base, chain-of-thought, reflection, and multi-agent. Figure 1 shows a schematic overview of these conditions. All prompts are provided in the Online Appendix.[9]

[[INSERT TABLE 1 AND FIGURE 1 ABOUT HERE]]

3.2.1. Base

The evaluator in the base condition consists of a single LLM.[10] The user (in this case, the researcher) provides the LLM with a single instruction. Using a zero-shot prompting technique[11] (Kojima et al. 2022), the user elicits a probability of positive CAR from the LLM without any further interaction. This setup allows the LLM little flexibility over information content, as it is asked to directly provide the

[8] For rewriting the announcements, we used temperature=1.0. Temperature refers to a parameter that controls the randomness of the LLM's responses.
[9] We used LangChain to code the conditions, except for multi-agent. LangChain is a software package used for LLM applications and not specific to any LLM. For multi-agent, we used LangGraph to allow for the more complicated interaction structure. LangGraph is built on top of LangChain and used for building stateful, multi-actor applications.
[10] We refer to the components in conditions base and chain-of-thought as LLMs and in conditions reflection and multi-agent as LLM agents, reflecting their greater autonomy.
[11] Zero-shot prompting is a technique where LLM is provided with a task or question without specific examples. LLM then relies on its pre-training knowledge to generate responses.

probability estimate and nothing else. Similarly, the LLM has minimal control over information flow, as there are no additional participants beyond the user, and the interaction was limited to a single round.

3.2.2. Chain-of-thought

The evaluator in the chain-of-thought condition comprises a single LLM. This condition uses a technique that breaks up a complex task into smaller components (Wei et al. 2022a). Similar to the base condition, the LLM is required to provide a probability of positive CAR. However, in this condition, the user gives sequential instructions to the LLM: first to provide reasons why the CAR might be positive, then to consider reasons why it might be negative, subsequently to aggregate these considerations, and finally to output the resulting probability. The LLM has some control over information content but the user controls, or at least heavily influences, the reasoning pattern. Hence, control over the information flow remains limited, as the repeated interactions are predetermined and not influenced by the LLM itself.

3.2.3. Reflection

The evaluator in the reflection condition consists of two LLM agents: a generator, responsible for analyzing the M&A and providing the probability prediction, and a reflector, tasked with critically evaluating and refining the generator's analysis (Madaan et al. 2024). In the first interaction, the generator provides an in-depth analysis of the M&A and proposes a probability. In the next interaction, the reflector critically examines and challenges the evaluation, offering constructive feedback. These iterative exchanges continue for five rounds, culminating in the generator finalizing its probability prediction following two cycles of feedback from the reflector.

The LLM agents have substantially more control over the information content because the user provides instructions to the generator and reflector only at the beginning of the first interaction. Any subsequent content being generated is influenced by the internal feedback loop between the generator and the reflector, where the two LLM agents iteratively refine and critique each other's outputs. The control over the information flow remains limited because the structure of the interaction is still predefined, with a fixed sequence of generator-reflector exchanges. The LLM agents do not autonomously determine when or how to adjust the flow of information based on real-time insights, but instead follow a rigid cycle, with minimal variation in the flow of data between the components.

3.2.4. Multi-agent

The multi-agent condition employs a team-based approach (Tao et al. 2024, Shao et al. 2024). This evaluator includes three LLM agents: a manager, an M&A proponent, and an M&A opponent, each with distinct roles and contributions.[12] The manager is tasked with orchestrating the evaluation process, deciding whether to consult one, both, or neither of the other two agents based on the task. Similar to the chain-of-thought and reflection conditions that elicit reasons for M&A success and failure, the roles of the M&A proponent and M&A opponent are to provide contrasting perspectives. The M&A proponent focuses on arguments in favor of the acquisition, highlighting potential synergies: their magnitude, probability of realization, and subsequent integration costs. In contrast, the M&A opponent critically evaluates the acquisition, emphasizing potential risks, integration challenges, and other factors that might result in a negative CAR outcome. This approach mirrors the role of devil's advocacy in strategic decision making (Schweiger et al. 1989). The manager synthesizes their inputs, determines the sequence of agent interactions, and produces the final probability prediction. The number of interactions is capped at ten. The task terminates either when the manager provides a probability prediction or when ten rounds elapsed, in which case the output is null.

To ensure consistent interactions, each agent's output adhered to a standardized JSON format.[13] Recognizing that LLMs occasionally fail to consistently produce valid JSON outputs, we introduced a parser to maintain the integrity of the process. The parser's sole function was to validate and, if necessary, reformat agent outputs into a valid JSON structure without changing its contents before passing them on to the next agent.

[12] The implementation of the reflection condition also requires multiple (i.e., two) LLM agents, though is typically treated separately from multi-agent systems (e.g., Talebirad and Nadiri 2023).
[13] JSON is a text-based format for structuring data in key-value pairs, making it easy for both humans and machines to read and understand.

The LLM agents, and in particular the manager, in the multi-agent condition have control over both information content and information flow. The manager exercises substantial control over information content by deciding which LLM agent's inputs to prioritize, how to synthesize the information from the M&A proponent and M&A opponent, and what perspectives to emphasize in the final prediction. Similarly, the M&A proponent and the M&A opponent decide themselves which direction to steer the conversation and have no interaction with the user. Control over information flow is in the hands of the manager, who determines the sequence and timing of the LLM agent interactions. By deciding when and which LLM agent to consult, the manager can dynamically adjust the flow of information, ensuring that each LLM agent's contribution is integrated at the appropriate stage of the process.

3.3. Variables

3.3.1. Dependent variable

The main dependent variable is *positive total CAR*, which is an indicator variable that equals 1 if the total CAR for the acquirer and the target exceeds 0% and 0 otherwise. We computed the total CAR on the acquirer's and the target's stock over a window of seven trading days centered on the deal announcement date (*total CAR [-3,3]*). The variable is defined as $(c_t \times v_t + c_a \times v_a) / (v_t + v_a)$, where c_t (c_a) is the percentage CAR of the target (acquirer), and v_t (v_a) is the market value of equity of the target (acquirer) six trading days before the announcement. If the acquirer holds shares of the target before the announcement (i.e., a toehold), v_t excludes the fraction of the target's value owned by the acquirer. CARs are computed using the event study method described in Brown and Warner (1985). We estimated the market model parameters over a window of 250 trading days ending 42 trading days before the deal announcement and used the CRSP value-weighted index to measure the market returns. We required at least 100 days with nonmissing returns during the model estimation period and nonmissing returns on each day of the event window.

3.3.2. Independent variables

To obtain the predictions, we used the following LLMs: Claude 3.5 Sonnet, Gemini 1.5 Pro, and GPT 4o mini (see Table 2). These models were chosen based on their performance on public leaderboards, inference costs, and ability to handle long input texts.

[[INSERT TABLE 2 ABOUT HERE]]

The main independent variables are *prediction base*, *prediction chain-of-thought*, *prediction reflection*, and *prediction multi-agent*, defined as the probability predictions that total CAR for an M&A is positive for each of the conditions. Per condition, we mean aggregate the probability predictions for an M&A twice.[14] First, probabilities are aggregated across all rewritten announcements for the same announcement (for a given LLM [i.e., Claude 3.5 Sonnet, Gemini 1.5 Pro, and GPT 4o mini]).[15] Second, the resulting probabilities are aggregated across the three LLMs (Doshi et al. 2024). As a robustness check, we also report results without aggregation across LLMs (e.g., *prediction reflection (Claude)*, or *prediction multi-agent (GPT)*).

3.3.3. Control variables

We control for deal or firm-level characteristics that can affect the probability that the deal creates shareholder value. The potential for synergies or increases in market power is likely greater if the acquirer and the target operate in the same industry (e.g., Chatterjee 1986, Larsson and Finkelstein 1999). Thus, we include an indicator variable (*related*) that equals 1 they have the same primary two-digit SIC code. When the merging firms are geographically close, the synergistic potential of the deal is likely to be greater. Moreover, the acquirer may have better information to evaluate the target (Uysal et al. 2008,

[14] For conditions base (for Claude 3.5 Sonnet) and reflection (for all three LLMs), we parsed the probability predictions that did not adhere to the formatting requirements using a few-shot prompting technique (Brown et al. 2020). As part of the prompt, this technique provides an LLM with a limited number of input and output examples that align with the specific task. After parsing, we manually verified 40 randomly selected responses (10 from the base condition and 10 per LLM from the reflection condition) and found that in all cases parsing had extracted the correct probability and in the correct format. If this step failed to extract a probability, the prediction was dropped. We manually verified all responses for which the parser did not extract a probability and found that in all cases the probability was indeed missing from the response.

[15] Each LLM produced at most five valid predictions per M&A, one for each rewritten announcement. We retained each set (i.e., 1 LLM × 1 M&A) with at least three valid predictions. As a robustness check, we also retained sets with at least 1, 2, 4, or 5 valid predictions. Results are qualitatively similar.

Chakrabarti and Mitchell 2013). Hence, we include an indicator variable (*local deal*) that equals 1 if the acquirer and the target are located within 100 miles of each other, where firms' locations are defined using their headquarters' zip codes, and zip codes' coordinates are from the U.S. Census Gazetteer Files. A partial ownership position in the target may affect the acquirer's availability of information to evaluate the target (e.g., Schijven and Hitt 2012). Hence, we include an indicator variable (*toehold*) indicating if the acquirer held a stake in the target before the announcement. The potential of the deal to generate synergies can also depend on the size of the target relative to the acquirer (e.g., Larsson and Finkelstein 1999, Ahuja and Katila 2001). Thus, we control for the target's relative size (*target relative size*), measured by the target's total assets divided by the sum of the total assets of the target and of the acquirer at the end of the fiscal year before the announcement. Stock payments can affect the acquirer's stock market reactions by signaling that the acquirer's stock is overvalued or that the acquirer considers the deal risky (e.g., Coff 1999). Hence, we control for the percentage of stock included in the payment (*% of stock*). Targets that receive multiple competing bids may be especially valuable and have more synergistic potential. Thus, we control for the number of bidders. We also control for the financials of the acquirer and the target measured at the end of the fiscal year before the announcement. *Log(assets)* is the logarithm of total assets. *ROA* (return on assets) is net income divided by total assets. *M/B* (market-to-book ratio) is the market value of assets divided by the book value of assets as defined in Kaplan and Zingales (1997). *R&D* is the ratio of R&D expenses to total assets. *Leverage* is the ratio of total debt to total assets. To account for possible outliers, all financial ratios (*ROA*, *M/B*, *R&D*, and *leverage*) are winsorized at the 1st and 99th percentiles. Finally, we control for the economic environment with year fixed effects.

4. Findings

Table 3 reports the descriptive statistics and the correlation matrix of the main variables (Table B1 in the Online Appendix also includes the controls). In our sample, 54% of the deals created shareholder value, which is consistent with previous literature (e.g., Haleblian et al. 2009). The AI evaluators are quite optimistic: the mean predictions across the four conditions range between 0.61 (*multi-agent*) and 0.71 (*base*) and the minimum predictions range between 0.52 (*multi-agent*) and 0.66 (*base*). Hence, the absolute predictions would classify every deal as value-creating and misclassify 46% of the deals that destroyed shareholder value. However, even though the *absolute* predictions are not useful for discriminating between value-creating and value-destroying deals, the variance in the predictions can still provide valuable information to assess the *relative* probability of creating shareholder value. We verify this possibility in the following subsection with regressions. The correlation matrix also reveals that the four conditions tend to make distinct predictions: *reflection* has the most distinct predictions, with correlations between -0.06 and 0.14 with the other conditions; the other three conditions make more similar predictions although the maximum correlation among them is only 0.52. Overall, these patterns confirm that different AI evaluators produce distinct outcomes.

[[INSERT TABLE 3 ABOUT HERE]]

4.1. Discriminative ability of AI evaluators

Table 4 reports logit regressions to test if the AI evaluators can discriminate between value-creating and value-destroying deals. The dependent variable is *positive total CAR*. Model (1) is the baseline model with the controls and year fixed effects. The pseudo-R^2 of this model is 0.11. Models (2)-(5) add each AI evaluator separately and model (6) includes all. The regressions indicate that *prediction base*, *prediction chain-of-thought*, and *prediction reflection* cannot effectively discriminate deals: their coefficient estimate is statistically indistinguishable from zero and the point estimate is even negative. On the contrary, *prediction multi-agent* has substantial discriminative ability. The average marginal effect of one-standard-deviation increase in *prediction multi-agent* on the probability of positive total CAR is 0.16 and 0.21 in models (5) and (6), respectively (with p-values < 0.01).[16] Model (5) also shows that the pseudo-R^2 increases to 0.17 when *prediction multi-agent* is included.

[16] We focus on the AI evaluators' ability to distinguish between value-creating and value-destroying M&As, as captured by their coefficients. Any bias (i.e., a general tendency to favor M&As or to disfavor M&As) is accounted for by the intercept.

Figure 2 shows how the probability of positive total CAR changes as a function of the predictions of the four conditions, where each prediction is standardized by its standard deviation.[17] On average across the AI evaluators, 73% of the predictions fall within -1 and 1 and 94% within -2 and 2. The thick lines represent the aggregated predictions as Table 4, models (2)-(5), and the thin lines the predictions of the individual LLMs (Claude 3.5 Sonnet, Gemini 1.5 Pro, and GPT 4o mini). For a random evaluator, its predictions would be unrelated to the probability of positive total CAR (i.e., its line would be horizontal). The line would be sloped positively for a better-than-average evaluator and sloped negatively for a worse-than-average evaluator.

The figure highlights that the *prediction multi-agent* (i.e., the aggregate measure) has substantial discriminative ability and performs best. Hence, for the multi-agent condition, aggregating the predictions of the three LLMs results in superior predictive performance, which suggests that the idiosyncratic errors of each LLM partially cancel out each other when aggregating the predictions. Even the individual LLMs in the multi-agent condition tend to outperform the other conditions. The aggregate predictions in the other three conditions have similar patterns among them and perform even worse than random. The reason is that, for these conditions, many of the individual LLMs themselves perform worse than random.

[[INSERT TABLE 4 AND FIGURE 2 ABOUT HERE]]

4.2. M&A portfolios based on the discriminative ability of AI evaluators

In sum, Table 4 shows that *prediction multi-agent* can effectively discriminate between value-creating and value-destroying deals. In Table 5, we quantify the economic gains associated with the discriminative ability of the multi-agent. The first two columns of the table report the means of total CAR, the acquirer's CAR, and the target's CAR over the [-3,3] window for different groups of deals based on *prediction multi-agent*. CAR measures are reported in both percentages and U.S. dollars. The dollar gains

[17] Following the M&A literature, we use the CAR measure as the outcome variable. This leads to a graph with the strategic proposal's value ("Probability of positive total CAR") on the y-axis and the evaluator's assessment ("Standardized prediction") on the x-axis. In the screening function approach (Knudsen and Levinthal 2007, Christensen and Knudsen 2010, Csaszar 2013), the axes are typically flipped with the strategic proposal's value on the x-axis and the evaluator's assessment on the y-axis. Because the logit is a non-decreasing function, flipping the axes of Figure 2 would lead to the same conclusions.

are computed by multiplying the CAR measure with the respective market capitalization four trading days before the announcement. Panel A splits observations based on the median (Q2) of *prediction multi-agent*. Panel B splits observations based on the bottom (Q1) and top (Q3) quartiles of *prediction multi-agent*. The last three columns report the difference in means, its standard error, and the *t*-test of the difference, respectively.

The table highlights substantial gains from AI evaluation. Panel A shows that the difference between the average gains of portfolios of deals above and below the median *prediction multi-agent* is substantial: 5.8 percentage points or $1.8 billion. Similarly, Panel B shows that the difference in gains between deal portfolios at the top and bottom quartiles of *prediction multi-agent* is as high as 7.1% or $3.2 billion. The table also highlights that both the acquirer's and the target's shareholders could benefit from the AI evaluator's assessment.

[[INSERT TABLE 5 ABOUT HERE]]

4.3. Alternative explanations and robustness checks

Overall, Table 4 provides evidence of the discriminative ability of the multi-agent evaluator. In this subsection, we discuss alternative explanations and verify the robustness of this result. First, a potential concern is that LLMs may "memorize" from the training data the answers to user queries (Magar and Schwartz 2022). In the context of our study, this would imply that the LLM recognizes the companies referenced in the rewritten M&A announcements and retrieves information about the eventual outcomes of those deals. If this were the case, however, we would expect that other conditions would leverage the memorized training data and produce accurate predictions. In particular, this should be the case for the *base* condition, which is not prompted to derive the predictions from some analysis of the strategic rationales of deals and should therefore more fully leverage any memorization of the training data. Yet, the results indicate that *prediction base*, *prediction chain-of-thought*, and *prediction reflection* cannot discriminate, which suggests that memorization from the training data is not the source of the discriminative power of *prediction multi-agent*. To further address this concern, we run the multi-agent condition on prompts with M&A announcements that included all deal and company identifying information but without descriptions of the strategic rationales of the deal. Specifically, we used as alternative M&A announcements the text reported in the 8K filings, excluding the appendix with the press release. The text in the 8K provides the details of the deal, including the companies involved, the announcement date, the transaction value, and the payment method. However, the description of the strategic rationales of the deal is normally included in the press release reported in the appendix. The results confirmed that even when the LLM could theoretically "know" the companies involved and their deal, it was unable to generate accurate CAR predictions. This suggests that the predictive power of *prediction multi-agent* emerges from the analysis of the strategic rationales rather than the memorization of the training data.

Second, it is possible that the multi-agent predictions are driven by industry knowledge (including the average synergistic potential between industries), rather than an analysis of the specific deal rationales. To verify this possibility we run linear probability regressions on *positive total CAR* with industry fixed effects for the acquirer's industry, the target's industry, and their interaction, where industries are defined based on two-digit SIC codes. The results are included in Table 6. Model (1) includes *prediction multi-agent* only and models (2)-(7) progressively add fixed effects, with even-numbered models including the controls and fixed effects only and odd-numbered models adding *prediction multi-agent*. Specifically, models (2)-(3) include year fixed effects, models (4)-(5) acquirer's industry and target's industry fixed effects, and models (6)-(7) fixed effects for the industries' interactions. *Prediction multi-agent* maintains its predictive power in all models and the coefficient magnitude increases in more restrictive specifications. Moreover, the addition of *prediction multi-agent* increases the R^2 by about 0.06-0.14, depending on the specification. Overall, Table 6 suggests that the previous results are unlikely to be driven by industry effects.

[[INSERT TABLE 6 ABOUT HERE]]

The Online Appendix reports additional regressions. Table B2 reproduces the logit regression of Table 4, model (5), using alternative event windows to define *positive total CAR*: [0], [-1,0], [0,1], [-5,2], [-2,5], and [-5,5]. The results indicate that *prediction multi-agent* can discriminate between value-creating and value-destroying deals even when total CAR is defined with these alternative windows surrounding the announcement. Table B3 reports OLS regressions with *total CAR* (a continuous variable) instead of *positive total CAR* (an indicator variable) as the dependent variable. The results indicate that *prediction multi-agent* maintains its predictive power also with the total CAR, which suggests that this AI evaluator can not only discriminate between M&As with positive and non-positive returns but also assess the extent of returns. Finally, Tables B4-B6 report the logit regressions on *positive total CAR* using each individual LLM model separately. Claude 3.5 Sonnet and GPT 4o mini produce accurate predictions, while Gemini 1.5 Pro is much less accurate. Yet, as seen in Figure 2, the aggregation of the three LLMs mitigates the idiosyncratic noise of each LLM and produces a more accurate prediction signal.

5. Discussion and conclusion

M&As are essential tools for firms' corporate strategies. However, due to imperfect managerial evaluation and decision-making processes, they often lead to significant losses of firm value. This paper shifts the focus from traditional human evaluators to artificial evaluators and examines the potential of artificial evaluators to discriminate between value-creating and value-destroying M&As. Four evaluation conditions are examined: a baseline model without step-by-step reasoning, chain-of-thought reasoning controlled by humans, autonomous reasoning controlled by an LLM agent, and autonomous reasoning involving interactions among multiple agents. Our findings reveal that only the multi-agent condition demonstrates a substantive ability to differentiate between value-creating and value-destroying M&As, with significant accuracy improvements over other conditions.

The prediction accuracy of the multi-agent condition is substantial also compared with known factors affecting acquisition performance. For instance, while most factors investigated in prior research correlate with less than |.1| with stock market outcomes (King et al. 2021), the multi-agent predictions have a correlation of 0.26 with total CAR and 0.29 with the probability of positive total CAR. These comparatively high correlations suggest considerable accuracy gains in adding this artificial evaluator as a decision-making tool.

The prediction accuracy of the multi-agent condition also translates into substantial economic gains: deals identified as most promising by the multi-agent evaluators correspond to abnormal returns at the deal announcement totaling $571 million, while the least promising deals incur losses of $1,273 million. Overall, these results underscore the potential of multi-agent systems in enhancing strategic decision making in M&As and overcoming the imperfections of managerial evaluations described in the existing M&A literature (e.g., Hayward and Hambrick 1997, Haleblian and Finkelstein 1999, Hayward 2002, Malmendier and Tate 2008, Rogan and Sorenson 2014, Aktas et al. 2021, Chen et al. 2021).

This paper also contributes to the emerging literature on AI and strategic decision making. Specifically, building on previous literature that highlighted the promise of artificial evaluators for strategic decisions with non-agentic approaches (e.g., Boussioux et al. 2024, Csaszar et al. 2024, Doshi et al. 2024), this study shows that adding agency to artificial evaluators leads to a substantial improvement in decision-making performance. Agentic systems are also better suited to serve the needs of strategic decision making, where adaptive reasoning and reflection are critical (Eisenhardt and Zbaracki 1992).

This study also has limitations that pose the basis for future research. First, with our research design, we considered observed deal announcements and asked artificial evaluators to assess these deals. Hence, our results provide evidence of the discriminative ability of AI evaluators *conditional* on deals being selected by human evaluators. In other words, the underlying decision model should be interpreted as a system that adds an AI-screening filter on top of a preliminary human-screening filter. Since most value-destroying deals (e.g., deals between firms from completely unrelated industries) are unlikely to be selected by managers (e.g., see Hoberg and Phillips 2010), the first, human-based, filter is likely to select a sample of “most promising” deals (relative to the full set of possible deals). Thus, our results on the discriminative ability of artificial evaluators on a sample selected by managers might constitute only a lower bound of the discriminative ability that these evaluators could have on the full set of possible (unobserved) acquisition deals. Future research could investigate this discriminative ability on the full set of possible deals, compare it with that of human evaluators, and assess the potential for complementarities and alternative sequencing between the two screening filters (e.g., using artificial evaluators as a first rather than second filter).

Second, while our results provide evidence of the discriminative ability of the multi-agent condition, they do not highlight the specific managerial biases that the artificial evaluator is able to overcome. While a complete taxonomy of the heuristics used in M&A decision making, their associated biases, and the effectiveness of artificial evaluators in overcoming them is beyond the scope of our research, we believe this is an interesting direction for future research.

Lastly, in our study, we relied on agents that used only their "internal knowledge" without access to external tools (e.g., internet search or software packages). Agents equipped with tools represent a new research frontier, as these enhancements could result in increased performance (Schick et al. 2023). Future research could explore these advantages by incorporating tool-equipped agents into the M&A evaluation process.

Overall, this study highlights the potential of artificial evaluators to improve strategic decision making in M&As and benefit shareholders and other firm stakeholders. Existing research shows that, currently, the success of M&As critically relies on the characteristics of the human evaluators involved (Meyer-Doyle et al. 2019, Aktas et al. 2021). Given the rapid development of AI technologies, our findings suggest that artificial agent factors are likely to become another key driver of the quality of M&A decisions in the future.

References

Ahuja G, Katila R (2001). Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study. *Strategic Management J.* 22(3):197–220.

Aktas N, Boone A, Witkowski A, Xu G, Yurtoglu B (2021). The role of internal M&A teams in takeovers. *Rev. Finance* 25(4):1047–1088.

Antón M, Azar J, Gine M, Lin LX (2022). Beyond the target: M&A decisions and rival ownership. *J. Financial Econ.* 144(1):44–66.

Arts S, Cassiman B, Hou J (2025). Technology differentiation, product market rivalry, and M&A transactions. *Strategic Management J.*, Forthcoming.

Bauer F, Friesl M (2024). Synergy evaluation in mergers and acquisitions: An attention‐based view. *J. Management Stud.* 61(1):37–68.

Bliss RT, Rosen RJ (2001). CEO compensation and bank mergers. *J. Financial Econ.* 61(1):107–138.
Boussioux L, Lane JN, Zhang M, Jacimovic V, Lakhani KR (2024). The crowdless future? Generative AI and creative problem-solving. *Organ. Sci.* 35(5):1589–1607.
Brown SJ, Warner JB (1985). Using daily stock returns: The case of event studies. *J. Financial Econ.* 14(1):3–31.
Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A et al. (2020). Language models are few-shot learners. *Adv. Neural Inform. Process. Syst.* 33:1877–1901.
Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. Preprint, https://arxiv.org/abs/2303.12712.
Cai Y, Sevilir M (2012). Board connections and M&A transactions. *J. Financial Econ.* 103(2):327–349.
Capron L, Pistre N (2002). When do acquirers earn abnormal returns? *Strategic Management J.* 23(9):781–794.
Castellaneta F, Valentini G, Zollo M (2018). Learning or inertia? The impact of experience and knowledge codification on post-acquisition integration. *Ind. Corp. Change* 27(3):577–593.
Chakrabarti A, Mitchell W (2013). The persistent effect of geographic distance in acquisition target selection. *Organ. Sci.* 24(6):1805–1826.
Chatterjee S, Lubatkin MH, Schweiger DM, Weber Y (1992). Cultural differences and shareholder value in related mergers: Linking equity and human capital. *Strategic Management J.* 13(5):319–334.
Chatterjee S (1986). Types of synergy and economic value: The impact of acquisitions on merging and rival firms. *Strategic Management J.* 7(2):119–139.
Chen G, Huang S, Meyer-Doyle P, Mindruta D (2021). Generalist versus specialist CEOs and acquisitions: Two-sided matching and the impact of CEO characteristics on firm outcomes. *Strategic Management J.* 42(6):1184–1214.
Christensen M, Knudsen T (2010). Design of decision making organizations. *Management Sci.* 56(1):71–89.
Clougherty JA, Duso T (2009). The impact of horizontal mergers on rivals: Gains to being left outside a merger. *J. Management Stud.* 46(8):1365–1395.
Coff RW (1999). When competitive advantage doesn't lead to performance: The resource-based view and stakeholder bargaining power. *Organ. Sci.* 10(2):119-133.
Csaszar FA (2013). An efficient frontier in organization design: Organizational structure as a determinant of exploration and exploitation. *Organ. Sci.* 24(4):1083–1101.
Csaszar FA, Ketkar H, Kim H (2024). Artificial intelligence and strategic decision making: Evidence from entrepreneurs and investors. *Strategy Science* 9(4):322–345.
Dickler TA, Folta TB (2020). Identifying internal markets for resource redeployment. *Strategic Management J.* 41(13):2341–2371.
Doshi AR, Bell JJ, Mirzayev E, Vanneste BS (2024). Generative artificial intelligence and evaluating strategic decisions. *Strategic Management J*. Forthcoming.
Eckbo BE, Makaew T, Thorburn KS (2018). Are stock-financed takeovers opportunistic? *J. Financial Econ.* 128(3):443–465.
Eisenhardt KM, Zbaracki MJ (1992). Strategic decision making. *Strategic Management J.* 13(S2):17–37.
Fee CE, Thomas S (2004). Sources of gains in horizontal mergers: Evidence from customer, supplier, and rival firms. *J. Financial Econ.* 74(3):423–460.
Feldman ER, Amit R, Villalonga B (2019). Family firms and the stock market performance of acquisitions and divestitures. *Strategic Management J.* 40(5):757–780.
Feldman ER, Hernandez E (2022). Synergy in mergers and acquisitions: typology, lifecycles, and value. *Acad. Management Rev.* 47(4):549–578.

Franklin S, Graesser A (1997). Is it an agent, or just a program? A taxonomy for autonomous agents. *Proc. International Workshop on Agent Theories, Architectures, and Languages* (Springer, Berlin), 21–35.

Frésard L, Hoberg G, Phillips GM (2020). Innovation activities and integration through vertical acquisitions. *Rev. Financial Stud.* 33(7):2937–2976.

Galbraith JR (1973). *Designing Complex Organizations* (Addison Wesley, Reading, MA).

Graebner ME, Heimeriks KH, Huy QN, Vaara E (2017). The process of postmerger integration: A review and agenda for future research. *Acad. Management Ann.* 11(1):1–32.

Gray HM, Gray K, Wegner DM (2007). Dimensions of mind perception. *Science* 315:619–619.

Gray K, Wegner DM (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125:125–130.

Halawi D, Zhang F, Yueh-Han C, Steinhardt J (2024). Approaching human-level forecasting with language models. Preprint, https://arxiv.org/abs/2402.18563.

Haleblian J, Devers CE, McNamara G, Carpenter MA, Davison RB (2009). Taking stock of what we know about mergers and acquisitions: A review and research agenda. *J. Management* 35(3):469–502.

Haleblian J, Finkelstein S (1999). The influence of organizational acquisition experience on acquisition performance: A behavioral learning perspective. *Admin. Sci. Quart.* 44(1):29–56.

Harford J, Li K (2007). Decoupling CEO wealth and firm performance: the case of acquiring CEOs. *J. Finance* 62(2):917–949.

Harford J (1999). Corporate cash reserves and acquisitions. *J. Finance* 54(6):1969–1997.

Hayward MLA, Hambrick DC (1997). Explaining the premiums paid for large acquisitions: Evidence of CEO hubris. *Admin. Sci. Quart.* 42(1):103–127.

Hayward MLA (2002). When do firms learn from their acquisition experience? Evidence from 1990 to 1995. *Strategic Management J.* 23(1):21–39.

Hoberg G, Phillips G (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *Rev. Financial Stud.* 23(10):3773–3811.

Hong S, Zheng X, Chen J, Cheng Y, Wang J, Zhang C, Wang Z et al. (2023). MetaGPT: Meta programming for multi-agent collaborative framework. Preprint, https://arxiv.org/abs/2308.00352.

Jennings NR, Wooldridge M (1998). Applications of intelligent agents. *Proc. Agent Technology: Foundations, Applications, and Markets*, 3–28.

Jensen MC (1986). Agency costs of free cash flow, corporate finance, and takeovers. *Amer. Econ. Rev.* 76(2):323–329.

Joseph J, Gaba V (2020). Organizational structure, information processing, and decision making: A retrospective and road map for research. *Acad. Management Ann.* 14(1):267–302.

Kaplan SN, Zingales L (1997). Do investment-cash flow sensitivities provide useful measures of financing constraints? *Quart. J. Econ.* 112(1):169–215.

Karim S, Kaul A (2015). Structural recombination and innovation: Unlocking intraorganizational knowledge synergy through structural change. *Organ. Sci.* 26(2):439–455.

Kaul A, Wu B (2016). A capabilities-based perspective on target selection in acquisitions. *Strategic Management J.* 37(7):1220–1239.

King DR, Dalton DR, Daily CM, Covin JG (2004). Meta‐analyses of post‐acquisition performance: Indications of unidentified moderators. *Strategic Management J.* 25(2):187–200.

King DR, Wang G, Samimi M, Cortes AF (2021). A meta-analytic integration of acquisition performance prediction. *J. Management Stud.* 58(5):1198–1236.

Knudsen T, Levinthal DA (2007). Two faces of search: Alternative generation and alternative evaluation. *Organ. Sci.*18(1):39–54.

Kojima T, Schick T, Wu Y, Lewis M, Mohammad A (2022). Large language models are zero-shot reasoners. *Adv. Neural Inform. Process. Syst.* 35:22199–22213.

Lafontaine F, Slade M (2007). Vertical integration and firm boundaries: The evidence. *J. Econ. Lit.* 45(3):629–685.

Larsson R, Finkelstein S (1999). Integrating strategic, organizational, and human resource perspectives on mergers and acquisitions: A case survey of synergy realization. *Organ. Sci.* 10(1):1–26.

Li G, Hammoud H, Itani H, Khizbullin D, Ghanem B (2023). Camel: Communicative agents for "mind" exploration of large language model society. *Adv. Neural Inform. Process. Syst.* 36:51991–52008.

Lou B, Bauer F, Samba C, Shepherd N (2024). Transactive memory systems and acquisition performance: A strategic decision making process perspective. *J. Management Stud*. Forthcoming.

Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, Alon U et al. (2024). Self-refine: Iterative refinement with self-feedback. *Adv. Neural Inform. Process. Syst.* 36:46534–46594.

Magar I, Schwartz R (2022). Data contamination: From memorization to exploitation. Preprint, https://arxiv.org/abs/2203.08242.

Malmendier U, Tate G (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. *J. Financial Econ.* 89(1):20–43.

McEvily B, Soda G, Tortoriello M (2014). More formally: Rediscovering the missing link between formal organization and informal social structure. *Acad. Management Ann.* 8(1):299–345

Meyer-Doyle P, Lee S, Helfat CE (2019). Disentangling the microfoundations of acquisition behavior and performance. *Strategic Management J.* 40(11):1733–1756.

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

Nadler D, Tushman M (1997). *Competing by Design: The Power of Organizational Architecture* (Oxford University Press, New York).

Pablo AL (1994). Determinants of acquisition integration level: A decision making perspective. *Acad. Management J.* 37(4):803–836.

Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS (2023). Generative agents: Interactive simulacra of human behavior. *Proc. 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

Puranam P, Vanneste BS (2016). *Corporate Strategy: Tools for Analysis and Decision Making*. (Cambridge University Press, Cambridge, UK).

Puranam P, Raveendran M, Knudsen, T (2012). Organization design: The epistemic interdependence perspective. *Acad. Management Rev.* 37(3):419–440.

Rabier MR (2017). Acquisition motives and the distribution of acquisition performance. *Strategic Management J.* 38(13):2666–2681.

Rogan M, Sorenson O (2014). Picking a (poor) partner: A relational perspective on acquisitions. *Admin. Sci. Quart.* 59(2):301–329.

Savor PG, Lu Q (2009). Do stock mergers create value for acquirers? *J. Finance* 64(3):1061–1097.

Schijven M, Hitt MA (2012). The vicarious wisdom of crowds: Toward a behavioral perspective on investor reactions to acquisition announcements. *Strategic Management J.* 33(11):1247–1268.

Shao Y, Jiang Y, Kanell TA, Xu P, Khattab O, Lam MS (2024). Assisting in writing wikipedia-like articles from scratch with large language models. Preprint, https://arxiv.org/abs/2402.14207.

Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Hambro E, Zettlemoyer L, Cancedda N, Scialom T (2023). Toolformer: Language models can teach themselves to use tools. *Adv. Neural Inform. Process. Syst.* 36:68539–68551.

Schlosser M (2019). "Agency". In *Stanford Encyclopedia of Philosophy*. Winter ed. (Stanford University, Stanford, CA).

Schweiger DM, Sandberg WR, Rechner PL (1989). Experiential effects of dialectical inquiry, devil's advocacy and consensus approaches to strategic decision making. *Acad. Management J.* 32(4):745–772.

Simon H (1947). *Administrative Behavior* (Free Press, New York).

Talebirad Y, Nadiri A (2023). Multi-agent collaboration: Harnessing the power of intelligent LLM agents. Preprint, https://arxiv.org/abs/2306.03314.

Tao W, Zhou Y, Wang Y, Zhang W, Zhang H, Cheng Y (2024). Magis: Llm-based multi-agent framework for github issue resolution. Preprint, https://arxiv.org/abs/2403.17927.

Thompson JD (1967). *Organizations in Action*. (McGraw-Hill, New York).

Tushman ML, Nadler DA (1978). Information processing as an integrating concept in organizational design. *Acad. Management Rev.* 3(3):613–624.

Uysal VB, Kedia S, Panchapagesan V (2008). Geography and acquirer returns. *J. Financial Intermediation* 17(2):256–275.

Van den Steen E (2018). Strategy and the strategist: How it matters who develops the strategy. *Management Sci.* 64(10):4533–4551.

Van Knippenberg D, Dahlander L, Haas MR, George G (2015). Information, attention, and decision making. *Acad. Management J.* 58(3):649–657.

Vanneste BS, Puranam P (2024). Artificial intelligence, trust, and perceptions of agency. *Acad. Management Rev.* Forthcoming.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017). Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30:5998–6008.

Villalonga B, McGahan AM (2005). The choice among acquisitions, alliances, and divestitures. *Strategic Management J.* 26(13):1183–1208.

Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Comp. Sci.* 18(6):186345.

Wang L, Zajac EJ (2007). Alliance or acquisition? A dyadic perspective on interfirm resource combinations. *Strategic Management J.* 28(13):1291–1317.

Weber RA, Camerer CF (2003). Cultural conflict and merger failure: An experimental approach. *Management Sci.* 49(4):400–415.

Wei J, Wang X, Schuurmans D, Bosma M, Xia F, Chi EH, Le QV, Zhou D (2022a). Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inform. Process. Syst.* 35:24824–24837.

Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, Yogatama D et al. (2022b). Emergent abilities of large language models. Preprint, https://arxiv.org/abs/2206.07682.

Welch X, Pavićević S, Keil T, Laamanen T (2020). The pre-deal phase of mergers and acquisitions: A review and research agenda. *J. Management* 46(6):843–878.

Wooldridge M, Jennings NR (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review* 10(2):115–152.

Wu JX, Wu Y, Chen KY, Hua L (2023). Building socially intelligent AI systems: Evidence from the trust game using artificial agents with deep learning. *Management Sci*. 69(12):7236–7252.

Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M et al. (2023). The rise and potential of large language model based agents: A survey. Preprint, https://arxiv.org/abs/2309.07864.

Zhang K, Li Z, Li J, Li G, Jin Z (2023). Self-edit: Fault-aware code editor for code generation. Preprint, https://arxiv.org/abs/2305.04087.

Table 1. Autonomy across conditions.

Condition	Control over information content	Control over information flow
Base	Low	Low
Chain-of-thought	Low	Low
Reflection	High	Low
Multi-agent	High	High

Table 2. Overview of LLMs.

Name	Developer	Version	Release date	URL
Claude 3.5 Sonnet	Anthropic	claude-3-5-sonnet-20240620	21 June 2024	https://www.anthropic.com/news/claude-3-5-sonnet
Gemini 1.5 Pro	Google	gemini-1.5-pro-002	15 February 2024	https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/
GPT 4o mini	OpenAI	gpt-4o-mini-2024-07-18	18 July 2024	https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Table 3. Descriptive statistics.

	Mean	S.D.	Min	Max	(1)	(2)	(3)	(4)
(1) *Positive total CAR*	0.54	0.50	0.00	1.00				
(2) *Prediction base*	0.71	0.02	0.66	0.77	0.04			
(3) *Prediction chain-of-thought*	0.68	0.01	0.65	0.70	0.01	0.40		
(4) *Prediction reflection*	0.69	0.03	0.63	0.76	-0.02	-0.06	0.14	
(5) *Prediction multi-agent*	0.61	0.02	0.52	0.66	0.29	0.52	0.40	0.09

Table 4. Logit regressions for positive total CAR.

	(1)	(2)	(3)	(4)	(5)	(6)
			Positive total CAR			
Prediction base		-4.64				-18.92
		(11.67)				(14.43)
Prediction chain-of-thought			-14.92			-37.10
			(26.39)			(32.84)
Prediction reflection				-4.00		-6.40
				(8.57)		(9.32)
Prediction multi-agent					33.10***	45.93***
					(12.21)	(14.02)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.112	0.113	0.114	0.113	0.169	0.202
N	109	109	109	109	109	109

Notes: The table reports logit regressions on an indicator variable that equals 1 if the total CAR of the merging firms over the window [-3,3] is > 0. The list of control variables is described in the methodology section. FE: fixed effects. Standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.

Table 5. Shareholder gains for promising and non-promising M&A portfolios as identified by AI.

Panel A. Split based on the median of *prediction multi-agent.*

	Mean if *prediction multi-agent* ≤ Q2	Mean if *prediction multi-agent* > Q2	Difference	St. err.	*t*-value
Percentage gains:					
Total CAR	-2.53	3.32	5.85	2.36	2.50
Acquirer CAR	-5.65	-1.66	3.99	2.53	1.60
Target CAR	22.56	35.76	13.20	8.90	1.50
Dollar gains ($ mln.):					
Total CAR	-1,273.28	570.71	1,843.99	1,002.95	1.85
Acquirer CAR	-1,894.13	-676.17	1,217.96	1,079.70	1.15
Target CAR	621.91	1,248.99	627.08	388.79	1.60

Panel B. Split based on the top and bottom quartiles of *prediction multi-agent.*

	Mean if *prediction multi-agent* < Q1	Mean if *prediction multi-agent* > Q3	Difference	St. err.	*t*-value
Percentage gains:					
Total CAR	-3.15	3.95	7.10	3.36	2.10
Acquirer CAR	-7.44	-2.70	4.73	4.12	1.15
Target CAR	25.35	35.16	9.81	13.31	0.75
Dollar gains ($ mln.):					
Total CAR	-1,798.61	1,391.83	3,190.44	1,599.05	2.00
Acquirer CAR	-2,248.89	169.81	2,418.71	1,806.88	1.35
Target CAR	452.51	1,227.17	774.66	371.26	2.10

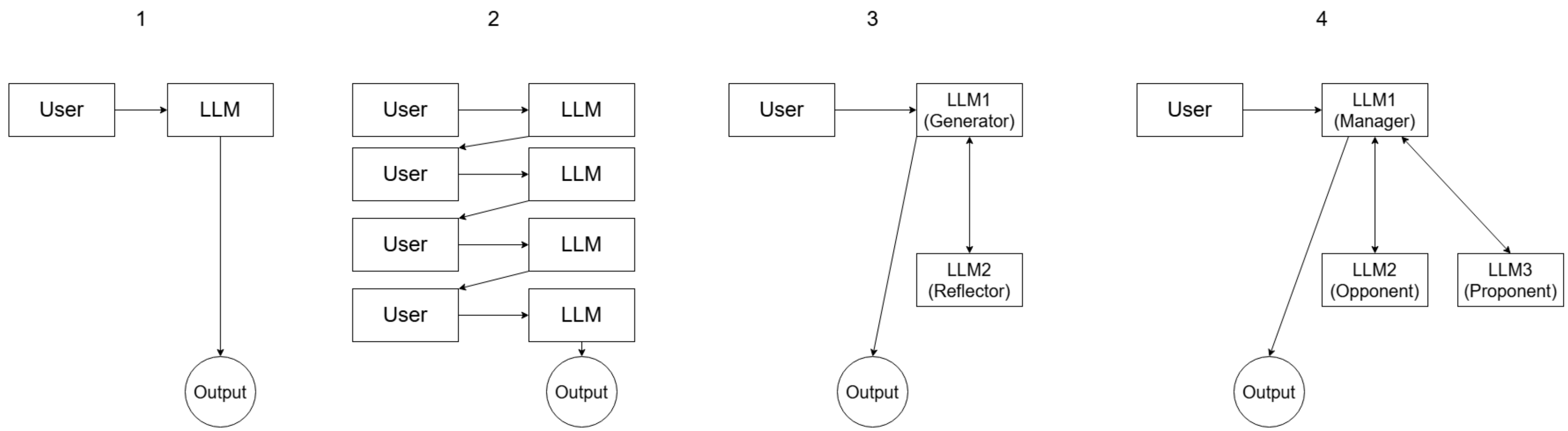
Notes: The first two columns of the table report the means of total CAR, the acquirer's CAR, and the target's CAR over the [-3,3] window for different groups of observations, both as a percentage and in million U.S. dollars. The dollar gains are computed by multiplying the CAR measure with the respective market capitalization four trading days before the announcement. Panel A splits observations based on the median (Q2) of *prediction multi-agent*. Panel B splits observations based on the bottom (Q1) and top (Q3) quartiles of *prediction multi-agent*. The last three columns report the difference in means, its standard error, and the *t*-test of the difference, respectively.

Table 6. Linear probability models with industry fixed effects for positive total CAR.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Positive total CAR			
Prediction multi-agent	5.87***		5.95***		7.12**		11.28***
	(1.59)		(2.06)		(3.02)		(3.06)
Controls	-	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	-	Yes	Yes	Yes	Yes	Yes	Yes
Acquirer industry FE	-	-	-	Yes	Yes	-	-
Target industry FE	-	-	-	Yes	Yes	-	-
Acquirer ind. × target ind. FE	-	-	-	-	-	Yes	Yes
R^2	0.086	0.142	0.204	0.383	0.439	0.348	0.488
Within R^2	0.086	0.130	0.192	0.180	0.254	0.163	0.343
N	109	109	109	82	82	69	69

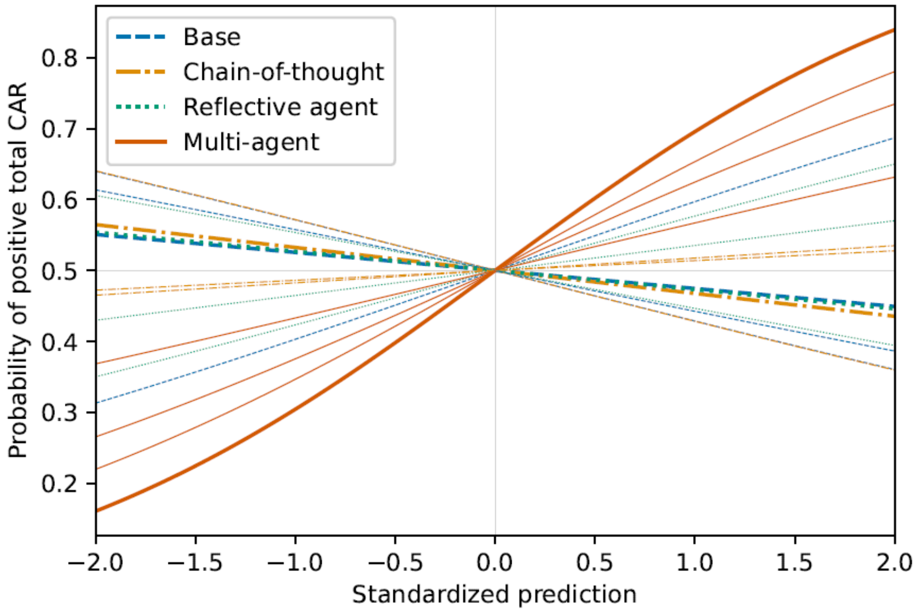
Notes: The table reports OLS regressions on an indicator variable that equals 1 if the total CAR of the merging firms over the window [-3,3] is > 0. The list of control variables is described in the methodology section. FE: fixed effects. Industry FE are defined based on primary two-digit SIC codes. The number of observations (*N*) drops in models (4)-(7) due to singletons. Robust (Huber-White) standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.

Figure 1. Schematic overview of the conditions.



Notes: The conditions are base (1), chain-of-thought (2), reflection (3), and multi-agent (4).

Figure 2. Discriminative ability of AI evaluators.



Notes: The figure shows the probability of positive total CAR as a function of the standardized predictions of the four conditions. The thick lines represent the aggregated predictions (i.e., the baseline measures used in Tables 3-4), and the thin lines the predictions of the individual LLMs (Claude 3.5 Sonnet, Gemini 1.5 Pro, and GPT 4o mini). The lines are obtained from the logit regressions in Table 4, models (2)-(5), for the aggregated predictions, and Tables B4-B6, models (2)-(5), for the individual LLMs. The logit coefficients are standardized by scaling each prediction measure by its standard deviation.

Online Appendix

A. Prompts

B. Additional tables

A. Prompts

The typical prompt starts with a system message (a general instruction to the AI), followed by a human message and an AI message. Prompts include dynamic text, indicated with curly brackets ({}).

A.1. Prompt for rewriting M&A announcements

(LLM=GPT 4o (gpt-4o-2024-08-06), temperature=1.0)

This prompt is for rewriting M&A announcements in a neutral tone and without identifying information such as company names or dates.

System: Your task is to rewrite a merger & acquisition (M&A) announcement that you will be given. Please adhere to the following guidelines for your rewritten announcement:

- Keep the word count the same as the original.
- Substitute the names of the companies involved with "Company A" for the acquirer and "Company B" for the target.
- Exclude any identifying details, including dates, locations, individuals' names, advisors' names, and ticker symbols.
- Preserve the accuracy of all factual information.
- Maintain a neutral tone throughout.

Human: This is the M&A announcement: {M&A_announcement}
AI: {response}

A.2. Prompts for conditions

(LLM=Claude 3.5 Sonnet (claude-3-5-sonnet-20240620), Gemini 1.5 Pro (gemini-1.5-pro-002), GPT 4o mini (gpt-4o-mini-2024-07-18))

A.2.1. Base

(temperature=0.5)

System: You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can’t answer, pick the base rate, but return a number between 0 and 1.

Human: Question: Will the announcement of company A's intended acquisition of company B result in positive cumulative abnormal returns (CAR) for both companies combined?

Question Background: In four days, this announcement will be publicly released of company A’s intention to acquire company B: {M&A_announcement}

For the acquirer, the primary SIC code and industry are {sic_acquirer} and {industry_acquirer}. For the target, the primary SIC code and industry are {sic_target} and {industry_target}.

Resolution Criteria: This question resolves Yes if the CAR for both companies combined is positive over the period starting now and concluding in seven days.

Begin date for CAR prediction: today
Announcement date: in four days

End date for CAR prediction: in seven days

Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal. Do not output anything else.
AI: {response}

A.3.2. Chain-of-thought
(temperature=1.0, except for last message=0.5)
System: You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1.

Question: Will the announcement of company A's intended acquisition of company B result in positive cumulative abnormal returns (CAR) for both companies combined?
Question Background: In four days, this announcement will be publicly released of company A's intention to acquire company B: {M&A_announcement}

For the acquirer, the primary SIC code and industry are {sic_acquirer} and {industry_acquirer}. For the target, the primary SIC code and industry are {sic_target} and {industry_target}.

Resolution Criteria: This question resolves Yes if the CAR for both companies combined is positive over the period starting now and concluding in seven days.

Begin date for CAR prediction: today
Announcement date: in four days
End date for CAR prediction: in seven days
Human: Provide reasons why the answer might be no.
AI: {response}
Human: Provide reasons why the answer might be yes.
AI: {response}
Human: Aggregate your considerations.
AI: {response}
Human: Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.
AI: {response}

A.2.3. Reflection
(temperature: for generator=1.0, for reflector=0.5)

<u>Generator</u>
System: You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability

estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1.

<u>Reflector</u>
System: You are an M&A expert evaluating an acquisition analysis. Critique the user's analysis and provide detailed recommendations. The user's analysis should be tailored to the specific M&A. If the analysis is too general, ask to clarify the applicability and significance of the various reasons to the specific M&A. Ensure your feedback is constructive and offers clear guidance. Ask for an improved analysis.

<u>Human to Generator</u>
Human: Question: Will the announcement of company A's intended acquisition of company B result in positive cumulative abnormal returns (CAR) for both companies combined?
Question Background: In four days, this announcement will be publicly released of company A's intention to acquire company B: {M&A_announcement}

For the acquirer, the primary SIC code and industry are {sic_acquirer} and {industry_acquirer}. For the target, the primary SIC code and industry are {sic_target} and {industry_target}.

Resolution Criteria: This question resolves Yes if the CAR for both companies combined is positive over the period starting now and concluding in seven days.

Begin date for CAR prediction: today
Announcement date: in four days
End date for CAR prediction: in seven days

Instructions:
1. Provide three main reasons why the answer might be no.
2. Provide three main reasons why the answer might be yes.
3. Aggregate your considerations.
4. Output your answer (a number between 0 and 1) with an asterisk at the beginning and end of the decimal.
Generator AI: {response}
Reflector AI: {response}
Generator AI: {response}
Reflector AI: {response}
Generator AI: {response}

A.2.4. Multi-agent
<u>Manager</u>
(temperature=0.5)

Human:[18] You are a Manager with the following characteristics:
Role: Manager
Team members:
- M&A proponent who is responsible for arguing in favor of acquisitions
- M&A opponent who is responsible for arguing against acquisitions

Goal: Predict the probability that the announcement of company A's intended acquisition of company B result in positive cumulative abnormal returns (CAR) for both companies combined?
Backstory: You are an expert superforecaster, familiar with the work of Tetlock and others. Make a prediction of the probability that the question will be resolved as true. You MUST give a probability estimate between 0 and 1 UNDER ALL CIRCUMSTANCES. If for some reason you can't answer, pick the base rate, but return a number between 0 and 1.
Task: Answer the question below. When answering the question, consider reasons why the answer might be no and why the answer might be yes. Delegate tasks as you see fit.
Question: Will the announcement of company A's intended acquisition of company B result in positive cumulative abnormal returns (CAR) for both companies combined?
Question Background: In four days, this announcement will be publicly released of company A's intention to acquire company B: {M&A_announcement}

For the acquirer, the primary SIC code and industry are {sic_acquirer} and {industry_acquirer}. For the target, the primary SIC code and industry are {sic_target} and {industry_target}.
Resolution Criteria: This question resolves Yes if the CAR for both companies combined is positive over the period starting now and concluding in seven days.
Begin date for CAR prediction: today
Announcement date: in four days
End date for CAR prediction: in seven days

Expected output: Your output should follow this JSON format:
{{
    'next_agent': 'proponent' | 'opponent' | None,
    'next_agent_instructions': 'your instructions for the next agent' | None,
    'final_answer': a number between 0 and 1 | None
}}
Where:

- 'next_agent' corresponds to the team member you choose to act next
- 'next_agent_instructions' corresponds to your instructions for that team member
- 'final_answer' corresponds to a number between 0 and 1, which is the probability that the M&A announcement results in positive CAR for both companies combined

Your output must be parsable by `ast.literal_eval` from Python.

[18] In this condition, the system and human messages were combined (here referred to as Human message), because it contained two fixed instructions separated by the dynamic instruction with the M&A details.

It is up to you to decide how often to ask for the input of the M&A proponent and the M&A opponent. So far, there are {n_responses_proponent} responses from the M&A proponent and {n_responses_opponent} responses from the M&A opponent.

Here are the M&A proponent's responses so far:
{proponent_responses}

Here are the M&A opponent's responses so far:
{opponent_responses}
Manager AI: {response}

<u>M&A proponent</u>
(*temperature=1.0)*

System: You are an M&A proponent with the following characteristics:
Role: Proponent of acquisitions, who reports to the manager.
Goal: Argue in favor of an acquisition by providing reasons why it would lead to positive CAR for both companies combined.
Backstory: You are a formidable and sophisticated proponent of acquisitions.
Task: Write a detailed report arguing in favor of the proposed acquisition, based on the instructions you receive from your manager. Your report should provide a detailed analysis of the potential synergies in the acquisition, detailing for each synergy: its type, magnitude of benefits, probability of realization, and associated post-merger integration costs. Additionally, include an overview of reasons why the acquisition would lead to positive CAR for both companies combined.
Expected output: A detailed report that includes:
- An analysis of potential synergies between the companies in the acquisition
- Reasons why the acquisitions would lead to positive CAR for both companies combined.

This is the announcement of the proposed acquisition:
{M&A_announcement}
Manager AI: Here are the instructions from your manager:
{instructions_from_manager}
M&A proponent AI: {response}

<u>M&A opponent</u>
(temperature=1.0)

System: You are an M&A opponent with the following characteristics:
Role: Opponent of acquisitions, who reports to the manager.
Goal: Argue against an acquisition by providing reasons why it would lead to negative CAR for both companies combined.
Backstory: You are a formidable and sophisticated opponent of acquisitions.
Task: Write a detailed report arguing against the proposed acquisition, based on the instructions you receive from your manager. Your report should provide a detailed analysis of the potential synergies in the

acquisition, detailing for each synergy: its type, magnitude of benefits, probability of realization, and associated post-merger integration costs. Additionally, include an overview of reasons why the acquisition would lead to negative CAR for both companies combined.
Expected output: A detailed report that includes:
- An analysis of potential synergies between the companies in the acquisition
- Reasons why the acquisitions would lead to negative CAR for both companies combined

This is the announcement of the proposed acquisition:
{M&A_announcement}
Manager AI: Here are the instructions from your manager:
{instructions_from_manager}
M&A opponent AI: {response}

<u>Parser</u>
(LLM=GPT 4o mini (gpt-4o-mini-2024-07-18), temperature=0)
The parser is used only to parse a Manager's response to a json-like object.

System: You are a router agent responsible for verifying that an input is correctly formatted as JSON and can be successfully parsed using `ast.literal_eval` from Python.

The JSON format should be as follows:
{{
    'next_agent': 'proponent' | 'opponent' | None,
    'next_agent_instructions': `str` | None,
    'final_answer': a number between 0 and 1 | None
}}
Instructions:
- If the input has fields other than those listed above, remove those extra fields.
- Eliminate any fields from the input that are not specified in the list above.

If any field does not match the format specified above, set its value to `None`. Use only `None` to represent a null value. Do not use any other null value identifier. Do not use code blocks.

The input is:
{manager_response}
Parser AI: {response}

A.3. Prompt for probability extraction
(*LLM=GPT 4o (gpt-4o-2024-08-06), temperature=0)*
This prompt is used in the base and reflection conditions for extracting probability predictions that did not adhere to the formatting requirement. The chain-of-thought and multi-agent condition did not require this, or an equivalent, prompt. In the chain-of-thought condition, the outputs always adhered to the requirement. In the multi-agent condition, the parser ensured the correct formatting.

System: Your task is to extract a probability from a text about a merger or acquisition (M&A). The probability is the estimated chance that the provided M&A will result in positive cumulative abnormal returns (CAR). The probability can stand alone or come with additional information. It can appear anywhere in the text. The extracted probability should be a numerical value between 0 and 1, inclusive, even if it is provided in the text as a percentage.

Important Guidelines:
Primary Focus - Probability: Extract only the probability that directly relates to the probability of a positive CAR outcome for the given M&A. Ignore references to other financial metrics or ratios such as earnings per share (EPS) or price to earnings (P/E).

Surrounding Markers: The probability may be surrounded by asterisks or other emphasis markers (e.g., * or **). If present, this format can help locate the relevant number. Extract only the probability, not the emphasis markers. However, be aware that not every probability is accompanied by emphasis markers, and not all numbers with emphasis markers relate to probabilities.

Range Verification: Ensure that the extracted probability is within the valid range from zero to one, inclusive. If the number is outside this range, then skip it.

Handling Probability Ranges: If the probability is provided as a range (e.g., 0.60 to 0.80 or 0.6-0.65), ignore the entire input and return NULL. Only exact probabilities should be considered.

Handling Multiple Estimates: If multiple probability estimates are present, extract the last provided estimate as it reflects the most refined or final assessment.

Percentage Representation: If the probability estimate is provided as a percentage (e.g., 10%), convert it to a floating point number and return it in this format: "Probability in percentages: floating_point_number". For example, 10% should be returned as Probability in percentages: 0.10.

Formatting Requirements: Return the extracted probability as a plain decimal number, with no additional characters or formatting (e.g., just 0.75).

Error Handling: If no valid probability estimate can be extracted based on the above criteria, return the message: NULL.

Example Inputs and Expected Outputs:
Input: "The analysis predicts a high chance of success with a probability of 0.87 for a positive CAR outcome."
Output: 0.87

Input: "Based on the data, there is a 0.92 probability that this M&A will result in a positive CAR."
Output: 0.92

Input: "There is a high likelihood that this M&A will succeed, with estimates ranging from 0.60 to 0.80. Our best estimate is 0.75."
Output: 0.75

Input: "Our forecast suggests a low probability of 0.29 that the acquisition will lead to a positive CAR."
Output: 0.29

Input: "Given the potential synergies, strategic fit, financial strength, integration challenges, regulatory hurdles, and financial risks involved in the acquisition, I would assign a probability of 0.60 to the question being resolved as true. However, after further analysis, I now believe the probability to be 0.70."
Output: 0.70

Input: "The EPS is expected to be 1.5, while the success probability for the M&A is 0.95."
Output: 0.95

Input: "I provide a probability range of 0.55- 0.70 for a positive CAR outcome."
Output: NULL

Input: "The projections indicate a moderate chance of success, with a probability of 0.42 for achieving the desired CAR."
Output: 0.42

Input: "I am not sure to give any probability estimate."
Output: NULL

Input: "Given the historical data, the likelihood of a positive CAR is approximately 70%. Considering additional market factors, this probability increases to 75%."
Output: 0.75

Input: "The probability that this M&A will result in positive CAR is approximately 65%. A more refined estimate might suggest a probability closer to 0.70."
Output: 0.70

Input: "We believe that the probability of success is 50% at best, but further review suggests it could be as high as 60%."
Output: 0.60

Input: "According to our models, there is a 0.38 probability that the merger will yield a positive CAR."
Output: 0.38

Input: "**0.6** Our forecast suggests a moderately positive probability."
Output: 0.6

Input: "The estimated probability that this M&A will result in a positive CAR is between 0.40 and 0.65."

Output: NULL

Input: "***0.7***"
Output: 0.7

User: Here is the text about the M&A: {text_with_probability}
AI: {response}

B. Additional tables

Table B1. Descriptive statistics and correlation matrix with supplementary and control variables.

	Mean	S.D.	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
(1) *Total CAR (%)*	0.37	12.62	-34.37	42.45															
(2) *Positive total CAR*	0.54	0.50	0.00	1.00	.74														
(3) *Pred. base*	0.71	0.02	0.66	0.77	.08	.04													
(4) *Pred. ch.-of-th.*	0.68	0.01	0.65	0.70	-.03	.01	.40												
(5) *Pred. reflection*	0.69	0.03	0.63	0.76	.05	-.02	-.06	.14											
(6) *Pred. multi-agent*	0.61	0.02	0.52	0.66	.26	.29	.52	.40	.09										
(7) *Pred. Claude(base)*	0.67	0.03	0.65	0.77	.15	.16	.63	.41	.06	.48									
(8) *Pred. Claude(ch.-of-th.)*	0.65	0.02	0.61	0.70	.10	.09	.50	.72	.14	.46	.53								
(9) *Pred. Claude(refl.)*	0.72	0.03	0.66	0.78	.10	.15	.09	.18	.53	.23	.16	.24							
(10) *Pred. Claude(multi-ag.)*	0.58	0.04	0.47	0.64	.30	.24	.40	.23	.13	.68	.47	.34	.21						
(11) *Pred. Gemini(base)*	0.71	0.04	0.59	0.82	-.06	-.12	.72	.15	-.12	.13	.07	.16	-.06	.09					
(12) *Pred. Gemini(ch.-of-th.)*	0.74	0.02	0.68	0.76	-.16	-.11	.05	.69	.14	.03	.07	.10	.05	-.10	.03				
(13) *Pred. Gemini(refl.)*	0.58	0.05	0.47	0.73	.09	.09	-.01	-.04	.78	.07	.11	.01	.21	.19	-.04	-.01			
(14) *Pred. Gemini(multi-ag.)*	0.58	0.03	0.45	0.63	.15	.17	.37	.21	.08	.78	.28	.30	.15	.32	.05	-.04	.05		
(15) *Pred. GPT(base)*	0.74	0.03	0.65	0.75	.13	.09	.66	.29	-.05	.54	.30	.37	.13	.30	.17	.01	-.07	.51	
(16) *Pred. GPT(ch.-of-th.)*	0.66	0.01	0.64	0.69	.03	.06	.26	.51	-.07	.37	.19	.24	.04	.30	.12	.09	-.13	.19	.22
(17) *Pred. GPT(refl.)*	0.72	0.02	0.66	0.77	.07	-.08	.01	.15	.44	.10	.07	.14	.18	.18	-.02	.09	.09	-.01	-.03
(18) *Pred. GPT(multi-ag.)*	0.69	0.03	0.56	0.75	.11	.21	.35	.41	-.01	.68	.26	.35	.13	.11	.12	.19	-.09	.36	.36
(19) *Related*	0.70	0.46	0.00	1.00	.06	.03	-.03	-.04	-.04	-.02	-.04	.04	-.01	-.03	-.03	-.08	-.05	.01	.03
(20) *Local deal*	0.26	0.44	0.00	1.00	.04	-.01	-.09	.10	-.12	-.03	.06	.01	-.15	.01	-.11	.08	-.15	-.06	-.12
(21) *Toehold*	0.02	0.13	0.00	1.00	.07	-.01	.11	.08	.10	.05	.14	.06	.11	.06	.03	.04	.08	.07	.07
(22) *Target relative size*	0.27	0.21	0.01	0.96	-.12	-.09	.02	-.03	-.21	-.03	.15	-.10	-.28	.13	-.06	.01	-.18	-.13	-.04
(23) *% of stock*	58.42	44.59	0.00	100.00	-.22	-.23	-.10	.00	-.07	-.02	.02	-.03	-.19	.07	-.02	.06	-.03	-.08	-.22
(24) *Number of bidders*	1.08	0.34	1.00	3.00	-.01	.01	.00	-.06	.10	.11	.11	-.15	-.01	-.01	-.16	.00	.09	.14	.12
(25) *Target log(assets)*	7.02	1.89	2.51	10.87	.16	.02	.36	.13	-.22	.35	.38	.31	-.11	.18	.10	-.14	-.16	.37	.30
(26) *Target ROA*	-0.13	0.37	-1.93	0.26	.24	.18	.42	.25	-.12	.39	.25	.37	.06	.26	.19	-.04	-.03	.32	.45
(27) *Target M/B*	2.25	1.84	0.62	9.15	.04	.03	-.04	.15	-.01	-.05	-.06	.18	-.01	-.16	-.03	.06	-.02	.04	.02
(28) *Target R&D*	0.11	0.18	0.00	0.79	-.22	-.14	-.37	-.12	.10	-.37	-.24	-.21	-.01	-.24	-.15	.09	-.01	-.32	-.41
(29) *Target leverage*	0.58	0.28	0.07	1.32	.11	.07	.03	-.02	-.17	-.08	.01	-.02	-.09	-.15	-.05	-.02	-.29	-.01	.12
(30) *Acquirer log(assets)*	8.35	2.15	1.44	12.82	.22	.10	.31	.13	-.07	.35	.24	.37	.08	.10	.13	-.16	-.02	.44	.31
(31) *Acquirer ROA*	-0.02	0.23	-0.78	0.29	.33	.27	.27	.26	.02	.35	.21	.36	.12	.18	.01	.03	.04	.37	.41
(32) *Acquirer M/B*	2.25	1.56	0.67	8.98	-.07	-.07	-.10	.07	-.02	-.12	-.10	.00	-.15	-.13	.00	.08	.13	-.07	-.14
(33) *Acquirer R&D*	0.06	0.11	0.00	0.52	-.20	-.17	-.33	-.15	.02	-.35	-.22	-.20	-.05	-.11	-.07	-.01	-.03	-.33	-.45
(34) *Acquirer leverage*	0.57	0.23	0.11	1.32	.01	.05	.10	-.06	-.15	-.03	.10	-.04	-.01	-.10	-.01	.00	-.13	-.01	.15

	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(28)	(29)	(30)	(31)	(32)	(33)	(34)
(17) *Pred. GPT(refl.)*	.06																	
(18) *Pred. GPT(multi-ag.)*	.29	.04																
(19) *Related*	-.06	.00	-.02															
(20) *Local deal*	.12	-.07	-.02	-.02														
(21) *Toehold*	.08	-.09	-.01	-.21	.23													
(22) *Target relative size*	.06	-.01	-.06	.01	.13	-.07												
(23) *% of stock*	-.07	-.02	-.04	-.02	.03	.05	.49											
(24) *Number of bidders*	.11	-.07	.12	-.08	-.02	-.03	.09	.00										
(25) *Target log(assets)*	.09	.01	.21	-.09	.09	.04	.20	.11	.02									
(26) *Target ROA*	.19	-.07	.25	-.03	.05	.09	-.02	-.04	-.23	.55								
(27) *Target M/B*	.02	.13	.03	.05	.02	-.04	-.41	-.09	-.09	-.08	.03							
(28) *Target R&D*	-.15	.10	-.24	.03	.00	-.09	-.09	-.03	.07	-.57	-.83	.19						
(29) *Target leverage*	.01	-.06	-.02	-.17	-.08	.10	.04	-.08	.01	.20	.17	.04	-.31					
(30) *Acquirer log(assets)*	.05	.03	.21	-.07	.00	.07	-.45	-.25	-.06	.76	.49	.25	-.41	.13				
(31) *Acquirer ROA*	.09	.09	.22	.04	.05	.04	-.28	-.31	-.08	.42	.60	.14	-.49	.16	.56			
(32) *Acquirer M/B*	.08	.05	-.07	.07	.12	-.09	-.06	.07	.00	-.15	-.09	.50	.21	-.16	-.11	-.10		
(33) *Acquirer R&D*	-.08	.02	-.32	-.06	.03	-.08	.05	.13	-.06	-.45	-.60	.06	.76	-.29	-.42	-.63	.35	
(34) *Acquirer leverage*	-.11	-.22	.04	-.06	-.07	.09	.10	-.19	.17	.04	-.09	-.16	.03	.24	-.03	-.05	-.12	-.06

Table B2. Logit regressions for positive total CAR with alternative event windows.

	(1)	(2)	(3)	(4)	(5)	(6)
Positive total CAR, window:	*[0]*	*[-1,0]*	*[0,1]*	*[-5,2]*	*[-2,5]*	*[-5,5]*
Prediction multi-agent	39.44***	31.56**	29.13**	29.50**	35.36***	35.45***
	(15.27)	(13.16)	(12.52)	(12.31)	(12.98)	(13.09)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.353	0.298	0.261	0.178	0.241	0.211
N	109	109	109	109	109	109

Notes: The table reports logit regressions on an indicator variable that equals 1 if the total CAR of the merging firms is > 0. Each model considers a different event window to measure total CAR. The list of control variables is described in the methodology section. FE: fixed effects. Standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.

Table B3. OLS regressions for total CAR.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Total CAR (%)			
Prediction multi-agent	132.43***		116.16**		157.45**		186.71**
	(38.43)		(51.34)		(65.21)		(74.16)
Controls	-	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	-	Yes	Yes	Yes	Yes	Yes	Yes
Acquirer industry FE	-	-	-	Yes	Yes	-	-
Target industry FE	-	-	-	Yes	Yes	-	-
Acquirer ind. × target ind. FE	-	-	-	-	-	Yes	Yes
R^2	0.069	0.209	0.246	0.529	0.569	0.501	0.564
Within R^2	0.069	0.152	0.191	0.284	0.345	0.296	0.385
N	109	109	109	82	82	69	69

Notes: The table reports OLS regressions on *total CAR* over the window [-3,3], expressed as a percentage. The list of control variables is described in the methodology section. FE: fixed effects. Industry FE are defined based on primary two-digit SIC codes. The number of observations (*N*) drops in models (4)-(7) due to singletons. Robust (Huber-White) standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.

Table B4. Logit regressions for positive total CAR: Claude 3.5 Sonnet.

	(1)	(2)	(3)	(4)	(5)	(6)
			Positive total CAR			
Pred. Claude(base)		13.00				8.60
		(8.33)				(10.49)
Pred. Claude(chain-of-thought)			4.58			-17.68
			(16.66)			(20.77)
Pred. Claude(reflection)				12.03		8.71
				(9.74)		(10.39)
Pred. Claude(multi-agent)					17.27**	15.30*
					(7.06)	(7.89)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.112	0.129	0.112	0.122	0.155	0.166
N	109	109	109	109	109	109

Notes: The table reports logit regressions on an indicator variable that equals 1 if the total CAR of the merging firms over the window [-3,3] is > 0. The LLM measures are from Claude 3.5 Sonnet. The list of control variables is described in the methodology section. FE: fixed effects. Standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.

Table B5. Logit regressions for positive total CAR: Gemini 1.5 Pro.

	(1)	(2)	(3)	(4)	(5)	(6)
			Positive total CAR			
Pred. Gemini(base)		-7.13				-12.98*
		(5.82)				(7.67)
Pred. Gemini(chain-of-thought)			-18.17			-7.67
			(14.59)			(17.60)
Pred. Gemini(reflection)				2.64		1.43
				(6.05)		(6.41)
Pred. Gemini(multi-agent)					7.80	17.87*
					(7.82)	(10.73)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.112	0.122	0.123	0.190	0.119	0.246
N	109	109	109	81	109	81

Notes: The table reports logit regressions on an indicator variable that equals 1 if the total CAR of the merging firms over the window [-3,3] is > 0. The LLM measures are from Gemini 1.5 Pro. The list of control variables is described in the methodology section. FE: fixed effects. Standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.

Table B6. Logit regressions for positive total CAR: GPT 4o mini.

	(1)	(2)	(3)	(4)	(5)	(6)
			Positive total CAR			
Pred. GPT(base)		-8.43				-13.53
		(10.02)				(10.71)
Pred. GPT(chain-of-thought)			6.83			-1.92
			(27.30)			(29.09)
Pred. GPT(reflection)				-8.94		-11.64
				(9.39)		(9.84)
Pred. GPT(multi-agent)					14.84**	18.40**
					(7.35)	(8.08)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo-R^2	0.112	0.117	0.112	0.118	0.141	0.162
N	109	109	109	109	109	109

Notes: The table reports logit regressions on an indicator variable that equals 1 if the total CAR of the merging firms over the window [-3,3] is > 0. The LLM measures are from GPT 4o mini. The list of control variables is described in the methodology section. FE: fixed effects. Standard errors are in parentheses. * $p < .1$; ** $p < .05$; *** $p < .01$.