

# 000 PARTNERMAS: AN LLM HIERARCHICAL MULTI- 001 AGENT FRAMEWORK FOR BUSINESS PARTNER SELEC- 002 TION ON HIGH-DIMENSIONAL FEATURES 003

004  
005  
006 **Anonymous authors**  
007 Paper under double-blind review  
008  
009  
010  
011  
012

## ABSTRACT

013 High-dimensional decision-making tasks, such as business partner selection, in-  
014 involve evaluating large candidate pools with heterogeneous numerical, categorical,  
015 and textual features. While large language models (LLMs) offer strong in-context  
016 reasoning capabilities, single-agent or debate-style systems often struggle with  
017 scalability and consistency in such settings. We propose PARTNERMAS, a hi-  
018 erarchical multi-agent framework that decomposes evaluation into three layers:  
019 a Planner Agent that designs strategies, Specialized Agents that perform role-  
020 specific assessments, and a Supervisor Agent that integrates their outputs. To sup-  
021 port systematic evaluation, we also introduce a curated benchmark dataset of ven-  
022 ture capital co-investments, featuring diverse firm attributes and ground-truth syn-  
023 dicates. Across 140 cases, PARTNERMAS consistently outperforms single-agent  
024 and debate-based multi-agent baselines, achieving up to 10–15% higher match  
025 rates. Analysis of agent reasoning shows that planners are most responsive to  
026 domain-informed prompts, specialists produce complementary feature coverage,  
027 and supervisors play an important role in aggregation. Our findings demonstrate  
028 that structured collaboration among LLM agents can generate more robust out-  
029 comes than scaling individual models, highlighting PARTNERMAS as a promis-  
030 ing framework for high-dimensional decision-making in data-rich domains. Our  
031 implementation is available at this anonymous link.  
032

## 1 INTRODUCTION

033 In real-world decision-making, practitioners often navigate high-dimensional data including exten-  
034 sive option sets and numerous evaluative features (Sandanayake et al., 2018; Sigle et al., 2023).  
035 Business partner selection which includes partner shortlisting and strategic alliance formation ex-  
036emplifies this challenge (Mindruta et al., 2016): firms often face a vast pool of potential candidates,  
037 each described by diverse attributes ranging from quantitative indicators (e.g., financial metrics,  
038 geographic presence) to text-rich information (e.g., strategic fit, investment preferences) (Shah &  
039 Swaminathan, 2008). The scale and complexity of such data can easily overwhelm human decision-  
040 makers, incurring significant costs (Li et al., 2008). This underscores the need for intelligent systems  
041 capable of analyzing large candidate sets and diverse features.  
042

043 Large language models (LLMs) have emerged as promising tools for addressing reasoning tasks  
044 in data-rich domains (Lee et al., 2025; Mischler et al., 2024). Unlike traditional machine learn-  
045 ing algorithms that often demand extensive training data, LLMs draw on pretrained knowledge and  
046 in-context reasoning to interpret heterogeneous information (Li et al., 2025a). With appropriate  
047 prompting (e.g., few-shot learning) or information retrieval techniques (e.g., RAG), these models  
048 can identify salient features using only feature and task descriptions, achieving performance compa-  
049 rable to established methods (Li et al., 2025a; Jeong et al., 2024). As task complexity has increased  
050 across domains, researchers have increasingly moved beyond single-agent approaches toward multi-  
051 agent systems (MAS), wherein complex problems are decomposed into specialized sub-tasks man-  
052 aged by agents operating within structured collaborative workflows (Li et al., 2024). This enables  
053 more sophisticated problem-solving by distributing cognitive load across multiple specialized com-  
ponents, each optimized for specific aspects of the overall task. Recent studies have demonstrated

054 the effectiveness of MAS across diverse domains, including software development (Tao et al., 2024),  
 055 mathematical reasoning (Li et al., 2025b), and healthcare decision-making (Chen et al., 2025).  
 056

057 Despite these advances, a significant gap remains in the application of MAS to high-dimensional  
 058 decision-making tasks within high-stakes domains such as finance, where effective automation could  
 059 substantially reduce cognitive burden on human experts while improving decision quality. Current  
 060 research on financial MAS has concentrated primarily on a narrow range of applications, such as  
 061 individual stock trading (Yu et al., 2024) and portfolio management (Luo et al., 2025), leaving  
 062 numerous critical financial decision-making areas substantially underexplored. This study addresses  
 063 business partner selection as a representative example of such underexplored domains, where MASs  
 064 need to deal with high-dimensional, heterogeneous features. This setting requires both scalability  
 065 and nuanced feature reasoning, a new aspect that previous MAS works have not explored much.  
 066

067 To investigate these challenges and advance MAS design, we develop a hierarchical MAS frame-  
 068 work, **PARTNERMAS** for business partner selection on high-dimensional features. PARTNERMAS  
 069 follows a three-tier design: a *Planner Agent* first analyzes the investment context and creates special-  
 070 ized evaluators; multiple *Specialized Agents* then assess candidate firms from different perspectives;  
 071 finally, a *Supervisor Agent* integrates their outputs to make the final selection. This hierarchical de-  
 072 sign brings several advantages: it enables decomposition of complex decision-making tasks, allows  
 073 weaker agents to contribute effectively through specialization, and provides robustness by synthe-  
 074 zizing diverse perspectives rather than relying on a single model’s judgment. To summarize, this  
 075 research makes the following two key contributions: (i) We introduce a tabular benchmark for co-  
 076 investor selection that captures real-world decision-making scenarios, featuring diverse candidate  
 077 firms and multifaceted evaluation criteria. (ii) We design and implement PARTNERMAS that mir-  
 078 rrors expert roles in business partner selection. Through extensive empirical validation, we demon-  
 079 strate that PARTNERMAS achieves significant performance improvements of approximately 15%  
 080 compared to single agent and traditional multi-agent debate methods.  
 081

## 082 2 RELATED WORK

083 High-dimensional data refers to datasets characterized by large numbers of heterogeneous fea-  
 084 tures (Tang et al., 2016). These settings pose well-documented challenges (Johnstone & Titterington,  
 085 2009) for both traditional machine learning and LLM-based methods, including overfitting (Kim &  
 086 Simon, 2014), feature redundancy (Ferreira & Figueiredo, 2012), and difficulties in interpreting  
 087 model outputs (Potts & Schmischke, 2021). Such data is prevalent in high-stakes domains like  
 088 healthcare (Patra et al., 2021) and finance (Fallahgoul, 2025).  
 089

090 **Business Partner Selection.** Selecting partners is a crucial first step in establishing business rela-  
 091 tionships (Shah & Swaminathan, 2008; Cummings & Demirkan). Prior research highlights central  
 092 drivers of partner selection, including value creation and trustworthiness. Firms often evaluate po-  
 093 tential partners based on complementary resources and capabilities like knowledge, technology, or  
 094 capital, to assess whether collaboration generates greater value than independent efforts (Furlotti &  
 095 Soda, 2018; Mindruta et al., 2016). Meanwhile, trustworthiness is essential to mitigate risks from  
 096 opportunistic behavior (Das & Rahman, 2010; Li et al., 2008). Trust is often inferred from past  
 097 collaborations or, in their absence, signals like reputation and transaction records (Lumineau et al.,  
 098 2021). Beyond value and trust, firms also weigh coordination costs, which can arise from commu-  
 099 nication demands, geographical distance, or cultural differences, causing misunderstandings even  
 100 with trustworthy partners (Gulati et al., 2012). To reduce risks, firms often prefer partners with ge-  
 101 ographic proximity, similar industry backgrounds, or cultural homophily (Mayer & Argyres, 2004).  
 102 These considerations make partner selection a complex, multidimensional decision-making process.  
 103

104 In the venture capital (VC) sector, which is the focus of our study, co-investor selection often begins  
 105 with a broad search for potential partners. After compiling an initial list, experienced managers  
 106 carefully evaluate multiple factors to identify firms that align best with their objectives. The lead  
 107 VC then extends invitations to the shortlisted candidates. However, this process is often lengthy and  
 108 labor-intensive, and the costs of partner selection are recognized as a major ex-ante transaction cost  
 109 in business exchanges (Lumineau et al., 2021), particularly for VC firms that frequently engage in  
 110 collaborations. This challenge presents the need for more intelligent and efficient selection methods.  
 111

112 **LLM-driven Feature Selection.** By leveraging pretrained knowledge, LLMs can rank, filter, or ex-  
 113 plain the importance of features using their names and task context (Jeong et al., 2024). Jeong et al.  
 114 (2024) has developed pipelines where LLMs like GPT-4 can generate feature importance scores or  
 115

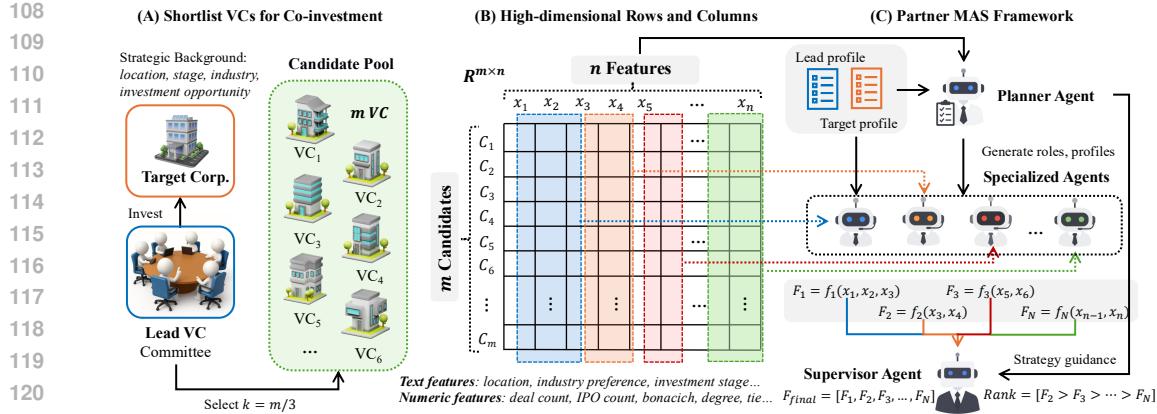


Figure 1: An illustration of the research design. (A) Co-investors shortlisting. (B) High-dimensional feature selection. (C) Hierarchical MAS framework.

explanations. Li et al. (2025a) has shown that LLMs can reliably identify key predictors in domains like healthcare and finance. In biomedical settings, researchers have improved performance by retrieving definitions of gene or protein identifiers to ground LLM reasoning in domain knowledge (Lee et al., 2025). Such LLM-driven selectors can rival standard statistical feature selection techniques (e.g. LASSO) even in zero-shot settings (Zhang et al., 2025).

Beyond single LLM, researchers have increasingly turned to MAS to handle complex decision-making. In MAS, agents assume specialized roles, such as planner, critic, or domain expert, and interact through structured workflows. For instance, MetaGPT assigns LLM agents to emulate software development teams (Hong et al., 2024), while debate-style frameworks allocate agents to critique and refine reasoning in math and coding tasks (Chan et al., 2024; Liang et al., 2024). Similar approaches in healthcare show that a “generalist” can triage cases and delegate to specialists for targeted diagnosis (Zuo et al., 2025). Such architectures outperform single LLMs by combining division of labor with consensus and conflict resolution. Some recent work applies MAS to feature engineering, where selector, generator, and coordinator agents refine feature sets (Gong et al., 2025). However, applying MAS to high-dimensional business data still remains underexplored, and little evidence shows how role-specialized agents can reliably coordinate to produce feature-informed decisions. This gap motivates our exploration of LLM-based MAS for business partner selection.

### 3 METHODOLOGY

#### 3.1 PROBLEM FORMULATION

Figure 1 illustrates our MAS architecture design. We specifically examine how the lead VC evaluates the initial candidate pool to create a more targeted shortlist (Figure 1 (A)). We do not address the final stage of selection, which moves from the shortlist to the actual partnership, as it involves more nuanced negotiation processes between the parties. By focusing on the shortlisting phase, we investigate the potential of LLM-based MAS to support the partner selection process.

We consider the candidate pool as high-dimensional, heterogeneous tabular data with  $m$  candidates (rows) and  $n$  features (columns) (Figure 1 (B)). Let it be  $C = \{c_1, \dots, c_m\}$ , where each candidate  $c_j$  is represented by a feature vector  $x_j \in \mathbb{R}^n$ . The feature matrix is  $X \in \mathbb{R}^{m \times n}$ . Columns are mixed-type consisting of both text fields (e.g., investment preference, firm type, industry preference) and numeric attributes (e.g., deal count, IPO count, network degree, tie strength). Given a task context  $\mathcal{Q}$  (e.g., lead VC and target company profiles), the goal is to produce a shortlist of candidates:

$$S = (s_1, \dots, s_k), \quad S \in C^k, \quad k = \lfloor m/3 \rfloor, \quad (1)$$

i.e., the top  $\lfloor m/3 \rfloor$  candidates from the pool. The proportion is set to one-third to emulate a realistic VC screening process, where a large pool of candidates is progressively narrowed down. For instance, an initial pool of 36 candidates is filtered to a shortlist of 12, from which a final group of four co-investors is selected (feature description detailed in Appendix C).

#### 3.2 MAS ARCHITECTURE DESIGN

To tackle the challenge of partner selection, we introduce a hierarchical MAS framework called **PARTNERMAS** (Figure 1 (C)). This design enables the system to decompose the high-dimensional

task into manageable sub-tasks, each assessing tabular data from different dimensions: (i) **Planner Agent** ( $PA$ ) that designs the evaluation strategy. (ii) **Specialized Agents** ( $SA$ ),  $\{SA_1, \dots, SA_N\}$ , which are dynamically configured by the  $PA$  to execute this strategy based on their role definition and expertise. (iii) **Supervisor Agent** ( $SPA$ ) that aggregates analyses to make a final decision.

**Planner Agent** ( $PA$ ). The primary role of  $PA$  is to interpret the high-level task context  $Q$  (Lead VC and target company profiles) and formulate an evaluation plan. It does so by analyzing  $Q$  with feature names to identify the most critical evaluation dimensions given the case description. The  $PA$ 's output is a set of  $N$  agent configurations,  $\{A_1, \dots, A_N\}$ , where each configuration  $A_i$  contains a specific “profile” that guides the corresponding  $SA_i$ . It also generates “strategic guidance” to support the decision for the  $SPA$ . Formally, the output can be expressed as:  $\{[A_1, \dots, A_k], PA(Q, C_{sample})\}$ , a set of explicit instructions for the team of  $SAs$ .

**Specialized Agent** ( $SA$ ). Each  $SA$  acts as a domain expert, tasked with evaluating the entire candidate pool  $C$  from its specialized perspective. To manage the high dimensionality of the candidate feature vectors ( $x_j \in \mathbb{R}^n$ ), the  $SA$  first performs a feature selection step. Guided by its assigned  $profile_i$ , it identifies and focuses on a relevant feature subset. The agent's core evaluation function  $f_i$ , driven by a backbone LLM, then directly produces a ranked shortlist of candidates. This output,  $S'_i$ , is a list containing the top  $k' = \lceil m/3 \rceil$  firms, where each entry includes the firm's ID, its rank, and an alignment score,  $score_{ij} \in [1, 10]$ .

**Supervisor Agent** ( $SPA$ ). The  $SPA$  is responsible for synthesizing the shortlists  $\{F'_1, \dots, F'_N\}$  from Specialized Agents into the final ranked list  $F$ . It mimics a human-led committee's decision-making process by first establishing consensus and then resolving disagreements based on the strategic priorities of the deal. This is achieved in a two-step process: (i) **Consensus Selection**: The  $SPA$  first identifies candidates with broad support across  $SA$ , by counting how many agents include them in their shortlists. This step can determine robust candidates that perform well across different dimensions:  $F_1(c_j) = \sum_{i=1}^N \mathbb{I}(c_j \in F'_i)$ . (ii) **Conflict Resolution**: To fill the remaining slots in the shortlist, the  $SPA$  resolves disagreements among the  $SAs$ . It first determines an importance weight  $w_i$  for each agent based on the agent's relevance to the specific deal. It then resolves the conflict for the remaining candidates by giving more weight to the opinions of more important  $SAs$ . This allows the  $SPA$  to select candidates who excel in critical areas, even if they lack broad consensus:  $F_2(c_j) = \sum_{i=1}^N w_i \cdot \frac{1}{R_i(c_j)}$ . The final shortlist  $F = [F_1(c_j), F_2(c_j)]$ .

### 3.3 EVALUATION

The system's performance is evaluated against a ground truth of successful co-investor partnerships. For each task, the generated shortlist  $F$  is compared to the set of ground-truth co-investors,  $G$ . The primary metric is the **Match Rate**, which measures the fraction of actual partners that are successfully identified by the MAS. It is formally defined as the recall of the system:

$$\text{Match Rate} = |F \cap G|/|G| \times 100\%. \quad (2)$$

For instance, if the initial candidate pool has 36 VC firms ( $m = 36$ ), our PARTNERMAS generates a shortlist of 12 ( $k = 12$ ). If the ground truth contains 4 actual co-investors ( $|G| = 4$ ) and 3 are found within the system's shortlist of 12 ( $|F \cap G| = 4$ ), the Match Rate is 75%. This metric measures the system effectiveness by ensuring that the true positive candidates are included in the final shortlist.

## 4 EXPERIMENT DESIGN

### 4.1 DATA PREPARATION

Our primary data source is the London Stock Exchange Group (LSEG) Workspace (London Stock Exchange Group, 2024), from which we collect VC investment records from 1980 to 2024. Following prior VC research (Makarevich, 2018; Wang et al., 2022), we restrict the sample to U.S.-based companies to avoid confounding effects from cross-country differences in regulatory frameworks and market institutions. We further exclude solo investments to focus on co-investments, where collaboration among investors is observable. After this initial filtering, the dataset comprise 52,662 companies backed by 16,030 VC firms, and we identify all active VCs in each year, industry, and state to construct the relevant candidate pool for each year-state-industry context.

To examine multiparty syndicate formation, specifically the process by which lead VCs select co-investment partners, we further restrict the sample to companies with complete first-round information and syndicates of at least three investors. We then merge the company list from LSEG with

PitchBook (PitchBook Data, Inc., 2024) to identify the lead VC for first-round investment based on the company name and the headquarter state. Only matches labeled as “high” or “very high” confidence are retained to ensure accurate lead investor identification. For analytical tractability, we limit the sample to cases with a single lead VC, since multiple leads often involve more complex governance and negotiation dynamics (Lerner, 2022; Kaplan & Strömberg, 2003). After these filters, the sample include 2,218 companies. We then merge PitchBook data back into the LSEG sample to retrieve lead VC information, supplemented by manual matching (company name, VC firm name, industry, and state) and excluding samples without necessary information. Applying all restrictions yields a final dataset of 140 cases for subsequent<sup>1</sup>.

## 4.2 EXPERIMENT SETTINGS

We evaluate multiple experiments for partner selection. All experiments use the same dataset and evaluation protocol, and all LLMs are run with temperature set to 0 to minimize output variance.

**Baseline Configurations: Single Agent.** In this baseline, a single LLM agent reviews all candidate firms and produces a ranked shortlist without external feedback. The parameter  $k$  denotes the number of independent runs:  $k=1$  corresponds to a one-shot evaluation, while  $k > 1$  allows the agent to generate multiple candidate shortlists. To obtain a final decision, the agent engages in a self-reflection step, comparing its own  $k$  outputs and selecting the one it deems most reliable. This design tests both the limitations of single-pass reasoning and the potential benefits of repeated deliberation within a single-agent framework. Unless otherwise specified, we set  $k=1$  by default. The details for the Single Agent prompts are shown in Appendix F.2.

**Baseline Configurations: Debate MAS.** The second baseline implements a debate-based multi-agent system (MAS) inspired by prior work (Chan et al., 2024; Liang et al., 2024). Three specialized agents simulate a VC committee: each independently evaluates candidates, critiques peers’ reasoning while scores remain hidden, and then revises its judgments in light of feedback. A supervisor agent synthesizes their inputs into a final shortlist. The details for the Debate MAS design and prompts are shown in Appendix E and Appendix F.3.

**Agent Configuration: PARTNERMAS.** Our MAS adopts the Planner–Specialist–Supervisor design described in Section 3. Unlike Debate MAS, which emphasizes adversarial critique, PARTNERMAS is built on structured collaboration and coordinated division of labor. Its design varies along two main dimensions: prompt guidance and backbone assignment. For prompts that guide Planner Agent and Supervisor Agent, we compare two conditions: (i) generic prompts without business knowledge, and (ii) business-domain guided prompts, which encourage agents to explicitly consider dimensions, including collaboration networks, industry fit, financial capacity, and geography (details in Appendix F). The details for the PARTNERMAS prompts are shown in Appendix F.4.

## 5 EXPERIMENTAL RESULTS

### 5.1 PERFORMANCE BENCHMARKING

We first compare the overall performance of PARTNERMAS against Single Agent and debate-based baselines with regard to the overall match rate across all business cases in our dataset. Results are summarized in Figure 2. Our key observations and analysis are as follows.

**PARTNERMAS achieves the strongest results.** Our hierarchical multi-agent system consistently outperforms all baselines. For example, PARTNERMAS with gpt-4.1-mini as the backbone achieves a match rate of 70.89% with business-domain guided prompt, which exceeds the best-performing single LLMs, such as gpt-5 (medium effort) (61.50%) and gemini-2.5-pro (61.42%). Notably, gpt-4.1-mini is a smaller and more cost-efficient model, its token cost is roughly an order of magnitude lower than gpt-5 or gemini-2.5-pro, yet when embedded in PARTNERMAS it delivers markedly stronger outcomes. This pattern holds even when using smaller backbones like gpt-5-nano, where PARTNERMAS still delivers 8–10% higher match rates than the same model in a Single Agent configuration. These results underscore that coordination among specialized agents can compensate for and often surpass pure scaling of model size. Importantly, the gains remain robust across different backbone LLMs, confirming that our framework is not tied to single LLMs.

<sup>1</sup>Due to intellectual property constraints of VC firms, the dataset cannot be released publicly but is available from the authors upon request for pure research purposes.

270 **Debate alone does not guarantee improvements.** We next assess whether adversarial  
 271 interaction between agents helps. The debate-based MAS baseline sometimes produces  
 272 modest gains relative to Single Agent baselines (e.g., 57.12% for gpt-5-nano in Debate  
 273 MAS vs. 55.14% for its Single Agent counterpart). However, the overall effect is  
 274 inconsistent: for gpt-4.1-mini, Debate MAS reaches only 60.19%, which lags substantially  
 275 behind PARTNERMAS with the same backbone, and in some cases performance even  
 276 drops below the Single Agent level. A likely explanation is that debate can distract agents  
 277 from their original reasoning or amplify minor errors, rather than directing to stronger  
 278 solutions. This finding shows that while debate can reveal reasoning errors, it lacks the  
 279 structured role division and aggregation mechanisms needed for complex business tasks.  
 280

281 **Business-domain guidance helps.** To verify  
 282 the effect of business domain knowledge, we  
 283 conduct experiments comparing model per-  
 284 formance with and without business-domain  
 285 guidance. The text for the guidance is shown in Appendix F.1. Across nearly all model  
 286 families, introducing business domain guidance leads to consistent accuracy gains, although marginal.  
 287 For Single Agent baselines, business-domain guidance improves match rates by 2–5% absolute,  
 288 such as for gpt-4.1-mini (55.95%). The improvements are more substantial for PARTNERMAS,  
 289 with some configurations showing gains of over 7%, rising from 62.55% to 69.03%). These re-  
 290 sults demonstrate that grounding agents in domain-relevant dimensions—such as financial capacity,  
 291 collaboration networks, or geographic compatibility—substantially enhances shortlist quality. In-  
 292 terestingly, the magnitude of improvement varies across models, indicating that stronger backbones  
 293 are better able to exploit domain cues, while smaller models benefit but plateau earlier.  
 294

301 **Backbone LLM effects.** Table 5.1  
 302 shows that among Single Agent  
 303 baselines, gpt-4.1-mini achieves  
 304 the highest match rates when busi-  
 305 ness domain guidance is provided,  
 306 followed by gpt-4o-mini, while  
 307 gpt-5-nano (medium effort) gen-  
 308 erally lags behind. This suggests  
 309 medium-sized models balance rea-  
 310 soning ability and efficiency, whereas  
 311 smaller models underperform and  
 312 larger ones may not justify their cost.  
 313 When incorporated into PARTNER-  
 314 MAS, however, even lightweight models like gpt-5-nano (medium effort) gain from struc-  
 315 tured role division, achieving 8–10% higher match rates relative to their Single Agent counter-  
 316 parts. These findings highlight that the hierarchical design not only amplifies the capabilities of stronger  
 317 backbones but also compensates for the weaknesses of smaller ones, yielding robust improvements.  
 318

319 **Performance–efficiency tradeoff.** Figure 3 illustrates the trade-off between match rate and the  
 320 overall token consumption. To ensure fairness, we evaluate Single Agent baselines with  $k=4$  runs,  
 321 aligning their computational cost with the multiple agents in Debate MAS (3+1 agents) and PART-  
 322 NERMAS (average of 4.27 agents). The results show that large single models such as gpt-5 and  
 323 gemini-2.5-pro fall into a high-cost yet only moderate-accuracy regime. In contrast, PARTNERMAS  
 324 configurations consistently deliver both higher match rates and lower token budgets. For example,  
 325 PARTNERMAS with gpt-4.1-mini achieves over 70% accuracy while consuming fewer tokens than

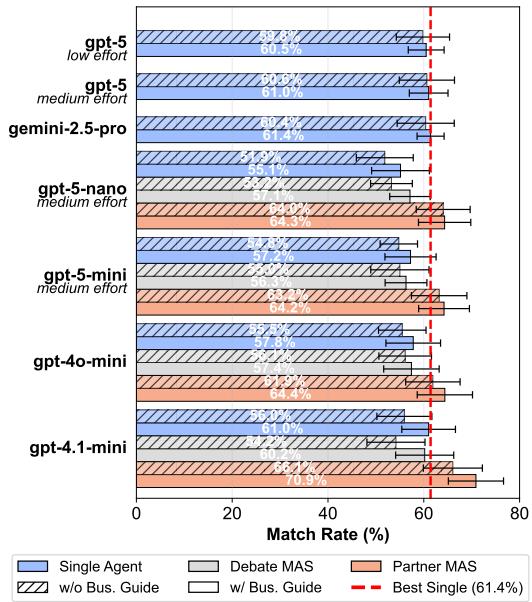


Figure 2: Performance benchmark for Single Agent, Debate MAS, and PARTNERMAS .

Table 1: Match rates (mean  $\pm$  95% CI) across different settings. All gpt-5-nano adopt the medium thinking effort.

PA	SA	SPA	Match Rate
gpt-4o-mini	gpt-4o-mini	gpt-4o-mini	64.40% $\pm$ 5.77
gpt-4.1-mini	gpt-4o-mini	gpt-4o-mini	63.21% $\pm$ 5.81
gpt-5-nano	gpt-4o-mini	gpt-4o-mini	65.19% $\pm$ 5.89
gpt-4o-mini	gpt-4.1-mini	gpt-4o-mini	62.14% $\pm$ 6.03
gpt-4o-mini	gpt-5-nano	gpt-4o-mini	64.70% $\pm$ 5.48
gpt-4o-mini	gpt-4o-mini	gpt-4.1-mini	69.03% $\pm$ 5.94
gpt-4o-mini	gpt-4o-mini	gpt-5-nano	64.70% $\pm$ 5.48

324 gpt-5 (medium effort). This efficiency arises from decomposition: specialized agents narrow their  
 325 focus to smaller feature subsets, reducing redundancy and improving coordination. Consequently,  
 326 PARTNERMAS proves not only more accurate but also more cost-effective, making it well suited  
 327 for practical deployment where API costs and inference latency are critical constraints.

## 328 329 5.2 AGENT PERFORMANCE AND REASONING ANALYSIS

330 To understand the internal dynamics,  
 331 we first analyze agent performance  
 332 (Figure 4) and then use re-  
 333 gression models to examine reason-  
 334 ing at each layer of the PARTNER-  
 335 MAS hierarchy. We evaluate the  
 336 Planner Agent’s deployment strat-  
 337 egy with logistic and linear regres-  
 338 sion (Table 2), visualize Specialized  
 339 Agents’ performance and feature fo-  
 340 cus with heatmaps (Figure 5), and as-  
 341 sess the Supervisor Agent’s decision-  
 342 making through regression analysis  
 343 (Table 3). This analysis reveals how  
 344 different backbones and business-  
 345 domain guidance, shape agent rea-  
 346 soning and overall performance.

### 346 System and agents performance.

347 Figure 4 shows the performance of  
 348 Specialized Agent clusters (see Appendix D for details on how we handle variations in Specialized  
 349 Agent names) on top right corner (achieving 65% to 75% accuracy with less than 45k token  
 350 usage), indicating our multi-agent-system build not only achieve higher accuracy but is also more  
 351 effective. To better understand how different agents perform and contribute to the final success of  
 352 the multi-agent-system, we present Figure 4: sub-figure (A) relates cluster performance to their us-  
 353 age and importance. Partnership History and Network Connectivity agents tend to be among the  
 354 strongest performers overall while Industry and Geographic agents are often ranked highly by the  
 355 planner/supervisor despite delivering comparatively lower accuracy. This misalignment implies that  
 356 the planning and supervision logic may be overvaluing broad topical coverage relative to histori-  
 357 cally effective signals for co-investment discovery. Sub-figure (B) and (C) show that MAS accuracy  
 358 is highest when the number of active agents is modest (approximately 4–5) and opinion diversity  
 359 is more concentrated (lower normalized HHI). As agent count and heterogeneity grow, returns di-  
 360 minish and aggregation becomes harder, consistent with the hypothesis that excessive diversity can  
 361 dilute the supervisor’s ability to extract a coherent signal.

361 **Model choice and business-domain-guided prompt drive planning.** Our analysis in Table 2  
 362 reveals that the Planner’s decisions are most significantly influenced by its core instructions and  
 363 backbone LLM rather than specific case context. The backbone LLM and the inclusion of a business  
 364 domain guidance are overwhelmingly the strongest predictors for the type and number of agents  
 365 deployed. For example, the odds of deploying an “Industry & Sector” agent increase by a factor  
 366 of 57.61 when a business domain guidance is provided. In contrast, for most cases, the contextual  
 367 factors, such as the target company’s industry focus or geolocation, show no statistically significant  
 368 effect. This indicates that the Planner Agent operates at a strategic level, relying on its guiding  
 369 prompts and model architecture to structure the problem-solving approach, rather than reacting to  
 370 the fine-grained details of each case.

371 **Specialized Agents value different features and their performance varies.** As illustrated in Fig-  
 372 ure 5, the effectiveness of Specialized Agents highly depend on their assigned role, leading to sig-  
 373 nificant performance variations across different areas of expertise. For instance, when using the  
 374 gpt-4.1-mini backbone, the “Risk & Compliance” agent excels with an impressive 83.3% accuracy,  
 375 while the Investment Stage” agent struggles, achieving only 37.7%. Similar gaps appear with other  
 376 models, though those driven by gpt-5-nano exhibit noticeably lower variance.

377 This performance difference is probably explained by how agents, guided by their roles, focus on  
 378 different features. Agents can successfully identify and prioritize relevant information; for exam-

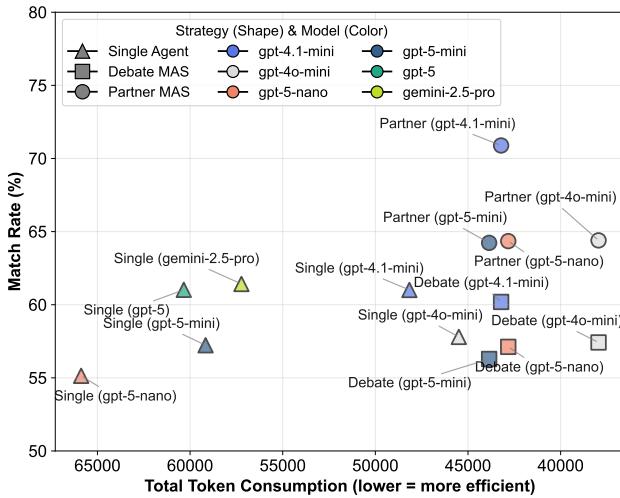


Figure 3: Model performance comparison across settings.

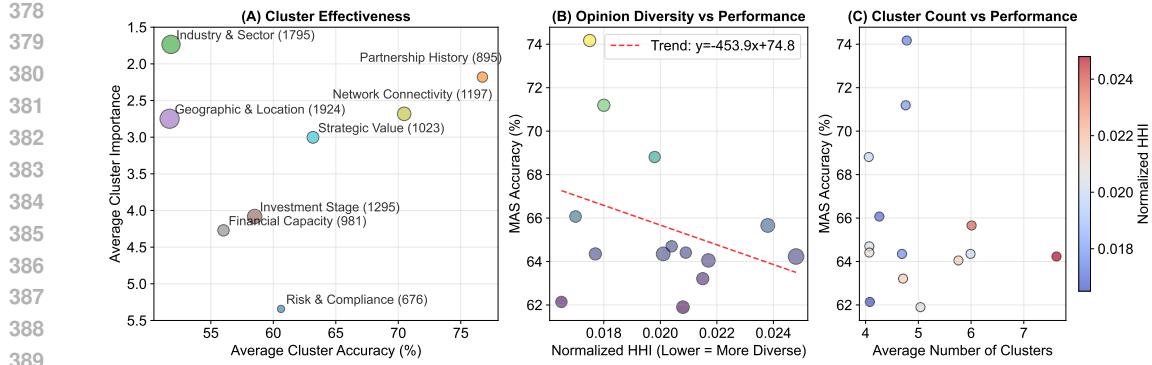


Figure 4: Agent performance grouped by Specialized Agent clusters. (A) Bubble chart of eight agent clusters, in which accuracy (x) vs importance rank (y, 1=highest), Bubble size = agents per cluster. (B) PARTNERMAS accuracy vs normalized HHI (lower = more diverse) with trend line. Node size = cluster count. (C) PARTNERMAS accuracy vs number of Specialized Agent clusters.

Table 2: Planner Agent deployment regression analysis with label encoder.

Variable	Agent Presence (Logistic Regression - Odds Ratios)								Agent Count	
	Risk & Compliance		Industry & Sector		Financial Capacity		Strategic Value			
	OR	Sig.	OR	Sig.	OR	Sig.	OR	Sig.		
prompt_hint	23.06	***	57.61	***	2.49	***	1.16	n.s.	0.12** (0.05)	
model	141.94	***	0.55	***	2.47	***	6.52	***	0.95*** (0.02)	
company_state	1.02	n.s.	1.00	n.s.	1.02	n.s.	0.99	n.s.	-0.00 (0.00)	
company_industry	0.71	*	1.01	n.s.	0.91	n.s.	1.09	n.s.	-0.03 (0.03)	
firm_investment_stage	1.02	n.s.	1.00	n.s.	0.96	*	1.00	n.s.	-0.00 (0.01)	
firm_type	0.88	n.s.	1.16	*	1.05	n.s.	0.88	n.s.	-0.01 (0.02)	
geography_preference	0.98	n.s.	1.00	n.s.	0.97	***	1.03	**	-0.00 (0.00)	
firm_state	0.99	n.s.	1.01	n.s.	0.98	n.s.	0.99	n.s.	0.00 (0.00)	
firm_industry_focus	0.99	n.s.	1.01	n.s.	1.01	n.s.	1.02	n.s.	-0.00 (0.00)	
R <sup>2</sup>	0.711		0.302		0.181		0.361		0.664	

**Notes:** prompt\_hint (generic=0; business=1). model (gpt-4o-mini=0; gpt-4.1-mini=1; gpt-5-nano=2; gpt-5-mini=3). Significance: \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ , n.s.  $p \geq 0.05$ . Other four Specialized Agent clusters do not show any significance and therefore are not presented in this table.

ple, the Geographic & Location” agent correctly emphasizes “geography\_preference” and “location” features. Interestingly, the underlying LLM model shapes the agent’s reasoning style. The “gpt-4.1-mini” model demonstrates a sharp, direct focus on specific features, with the Industry & Sector” agent targeting firm industry focus and the “Network Connectivity” agent concentrating on tie strength, degree, and bonacich centrality. In contrast, the gpt-5-mini model displays a more distributed pattern. This is because a Planner Agent driven by gpt-5-mini or -nano tends to generate a larger number of Specialized Agents, whose abilities may overlap. However, this broader agent deployment does not translate to superior performance. The performance of gpt-5-nano-driven agents is distributed, with no cluster exceeding 70% accuracy. This trend is also observable for gpt-5-mini, despite a standout 92.5% accuracy from its “Partnership History” agent cluster.

**Prioritize the right expert can increase the performance.** The regression analysis in Table 3 investigates how the Supervisor’s ranking of agent importance correlates with the final match rate. For the gpt-5-nano backbone, assigning the top rank to the “Partnership History,” “Industry & Sector,” or “Geographic & Location” agent is a highly significant predictor of a correct final outcome ( $p < 0.001$ ). For gpt-4.1-mini, success is strongly correlated with prioritizing agents on “Investment Stage” ( $p < 0.001$ ), “Strategic Value” ( $p < 0.01$ ), and “Network Connectivity” ( $p < 0.01$ ). This highlights that the Supervisor’s ability to weigh expert opinions is a key determinant of the system’s performance. Even if individual Specialized Agents perform well, an error in prioritization by the Supervisor can lead to a suboptimal final shortlist.

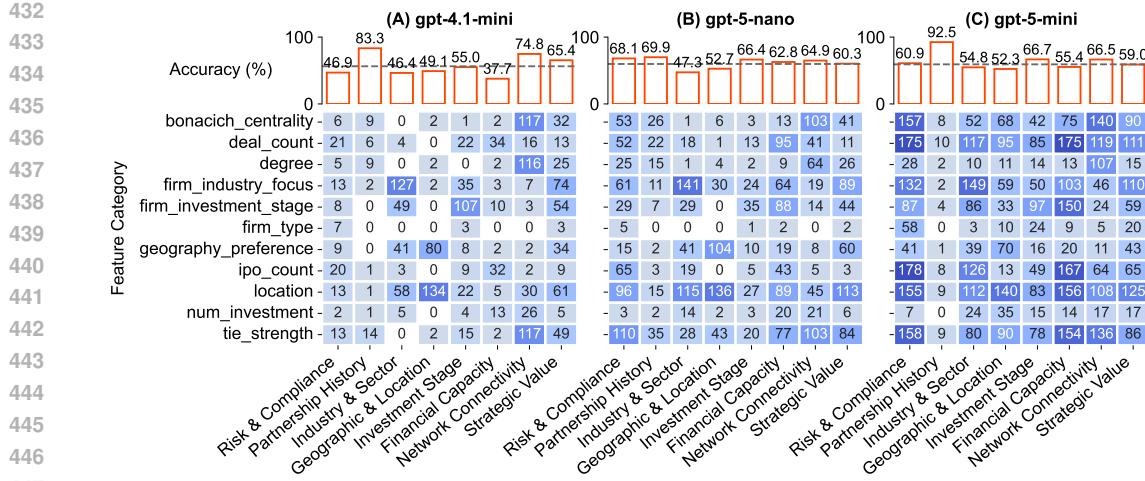


Figure 5: Accuracy and feature focus of Specialized Agents under different backbones: (A) gpt-4.1-mini. (B) gpt-5-nano. (C) gpt-5-mini.

## 6 CONCLUSION AND DISCUSSION

Our work makes three key contributions. First, we introduce a benchmark dataset for evaluating LLMs on high-dimensional tabular problems that combine numerical, textual, and categorical attributes. Grounded in real VC co-investment records, the dataset provides a realistic and challenging setting for testing reasoning in structured business decisions. Second, we propose PARTNERMAS, a hierarchical MAS that decomposes complex evaluation into planner, specialist, and supervisor layers. Unlike prior single-agent or debate MAS approaches, PARTNERMAS can dynamically configure role-specialized agents and coordinate their outputs to improve shortlist accuracy. While our case study focuses on business partner selection, the framework’s reliance on in-context reasoning rather than task-specific training suggests potential applicability to other domains, though broader validation is still required. Third, our analyses reveal how different layers contribute to overall performance: planners are most responsive to business-domain guidance, specialized agents generate complementary perspectives that improve coverage, and supervisors play a decisive role in integrating these signals into consistent final outcomes.

This study also presents several opportunities for future work. The dataset remains comparatively small due to availability and quality constraints, and its focus on U.S. venture capital restricts the test size. Our evaluation also relies primarily on advanced GPT backbones, leaving open how the system performs with lighter or open-source models that would be more practical in resource-constrained environments. Finally, while specialized agents often achieve strong role-specific accuracy, the Supervisor Agent occasionally aggregates them poorly, reducing final shortlist quality. This suggests that coordination, not specialization, is the current bottleneck; improving supervisory mechanisms through meta-reasoning or structured consensus remains a key direction for future research. Overall, PARTNERMAS demonstrates that structured collaboration among LLM agents can outperform both single-agent and debate-based MAS baselines for high-dimensional decision-making. These results highlight the promise of PARTNERMAS as a general framework for complex decision tasks, with implications for business, healthcare, and other data-rich domains.

Table 3: Regression analysis of Supervisor Agent’s ranking.

Variable	gpt-4.1-mini	gpt-4o-mini	gpt-5-nano
R1_Risk & Compliance	-	-	-40.38
R1_Partnership History	7.29	21.15	20.33***
R1_Industry & Sector	2.39	16.04	23.44***
R1_Geographic & Location	20.21*	32.78	20.10***
R1_Investment Stage	22.43***	-	13.59
R1_Network Connectivity	11.98**	-	20.69***
R1_Strategic Value	19.35**	-15.21	13.59
prompt_hint_business	46.77***	28.50	33.71***
prompt_hint_generic	36.89***	25.00	37.64***
R <sup>2</sup>	0.040	0.034	0.057

**Notes.** prompt\_hint\_business (yes = 1; no = 0). prompt\_hint\_generic (yes = 1; no = 0). Significance: \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$ .

486 7 ETHICS STATEMENT  
487488 We adhere to the ICLR Code of Ethics and conducted this study using company- and firm-level  
489 investment records from licensed sources (LSEG Workspace and PitchBook). The research does not  
490 involve human subjects or personally identifiable information. Due to licensing and confidentiality  
491 constraints associated with these data sources, we do not redistribute the raw data; we report aggregated  
492 analyses in the paper, and the curated test dataset is available from the authors upon reasonable  
493 request for research purposes only.494 PARTNERMAS is designed as a decision-support system that produces shortlists and rationales to  
495 assist expert judgment; it is not intended to autonomously make or execute investment decisions.  
496 As with any system trained or powered by large language models and historical records, outputs  
497 may reflect biases or limitations present in data and models. We do not claim to have performed  
498 formal fairness or bias audits in this work; instead, we acknowledge this as an important limitation  
499 and encourage careful human oversight in any practical use. We reduce variance where possible by  
500 using deterministic settings (e.g., temperature set to 0; see Section 4.2) and by grounding prompts  
501 in domain-relevant factors (Appendix F). We also describe our use of LLMs for research assistance  
502 (e.g., coding help, grammar checks) in the Appendix (Use of LLMs).503 8 REPRODUCIBILITY STATEMENT  
504505 We aim to make our study reproducible within the constraints of data licensing and LLM services.  
506 The problem setup, model architecture, and evaluation protocol are described in Section 3. Data  
507 construction and filtering steps for the candidate pool are detailed in Section 4.1, and the feature set  
508 is summarized in Appendix C. Experimental settings (including shared protocols and deterministic  
509 decoding with temperature set to 0) are provided in Section 4.2. The prompts used for each agent  
510 and configuration are included in Appendix F, including concrete versions for the Single Agent  
511 (Appendix F.2) and PARTNERMAS (Appendix F.4).512 The code base for the project is available at <https://anonymous.4open.science/r/Partner-MAS-7DCE>. Our curated test dataset contains firm- and deal-level information sub-  
513 ject to license and confidentiality obligations; to protect the privacy and commercial sensitivities of  
514 the companies involved, the dataset is available from the authors upon request for research purposes.  
515 The prompts necessary to reproduce the agent behaviors are appended at the end of the paper.516 Due to the evolving nature of hosted LLM services and provider-side updates, exact numerical  
517 results may exhibit minor variation across runs or over time, even with temperature set to 0. To  
518 mitigate this, we standardize the evaluation metric (Match Rate), use a fixed protocol across all  
519 experiments, and document all key choices in the paper and appendix so that researchers can follow  
520 the same setup and compare results under similar conditions.521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539

540 REFERENCES  
541

542 Cristiano Bellavitis, Joost Rietveld, and Igor Filatotchev. The effects of prior co-investments on  
543 the performance of venture capitalist syndicates: A relational agency perspective. *Strategic En-*  
544 *trepreneurship Journal*, 14(2):240–264, 2020.

545 Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu,  
546 and Zhiyuan Liu. Chateval: Towards better lilm-based evaluators through multi-agent debate.  
547 In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=FQepisCUWu>.

548

549 Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei  
550 Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large  
551 language models. *NPJ digital medicine*, 8(1):159, 2025.

552

553 Jeffrey Cummings and Irem Demirkan. Selecting best-fit alliance decision makers to minimize bias  
554 and enhance alliance performance. Available at SSRN 4649365.

555

556 Tarun K Das and Noushi Rahman. Determinants of partner opportunism in strategic alliances: A  
557 conceptual framework. *Journal of Business and Psychology*, 25(1):55–74, 2010.

558

559 Hasan Fallahgoul. High-dimensional learning in finance. *arXiv preprint arXiv:2506.03780*, 2025.

560

561 Artur J Ferreira and Mário AT Figueiredo. Efficient feature selection filters for high-dimensional  
562 data. *Pattern recognition letters*, 33(13):1794–1804, 2012.

563

564 Marco Furlotti and Giuseppe Soda. Fit for the task: Complementarity, asymmetry, and partner  
565 selection in alliances. *Organization Science*, 29(5):837–854, 2018.

566

567 Nanxu Gong, Sixun Dong, Haoyue Bai, Xinyuan Wang, Wangyang Ying, and Yanjie Fu. Agentic  
568 feature augmentation: Unifying selection and generation with teaming, planning, and memories.  
569 *arXiv preprint arXiv:2505.15076*, 2025.

570

571 Ranjay Gulati, Franz Wohlgezogen, and Pavel Zhelyazkov. The two facets of collaboration: Coop-  
572 eration and coordination in strategic alliances. *Academy of Management Annals*, 6(1):531–583,  
573 2012.

574

575 Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin  
576 Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a  
577 multi-agent collaborative framework. In *International Conference on Learning Representations*,  
578 2024.

579

580 David H Hsu. What do entrepreneurs pay for venture capital affiliation? *The journal of finance*, 59  
581 (4):1805–1844, 2004.

582

583 Daniel P Jeong, Zachary C Lipton, and Pradeep Ravikumar. Llm-select: Feature selection with large  
584 language models. *arXiv preprint arXiv:2407.02694*, 2024.

585

586 Iain M Johnstone and D Michael Titterington. Statistical challenges of high-dimensional data, 2009.

587

588 Steven N Kaplan and Per Strömberg. Financial contracting theory meets the real world: An empirical  
589 analysis of venture capital contracts. *The review of economic studies*, 70(2):281–315, 2003.

590

591 Kyung In Kim and Richard Simon. Overfitting, generalization, and mse in class probability estima-  
592 tion with high-dimensional data. *Biometrical Journal*, 56(2):256–269, 2014.

593

594 Joseph Lee, Shu Yang, Jae Young Baik, Xiaoxi Liu, Zhen Tan, Dawei Li, Zixuan Wen, Bojian Hou,  
595 Duy Duong-Tran, Tianlong Chen, et al. Knowledge-driven feature selection and engineering for  
596 genotype data with large language models. *AMIA Summits on Translational Science Proceedings*,  
597 2025:250, 2025.

598

599 Joshua Lerner. The syndication of venture capital investments. In *Venture capital*, pp. 207–218.  
600 Routledge, 2022.

594 Dan Li, Lorraine Eden, Michael A Hitt, and R Duane Ireland. Friends, acquaintances, or strangers?  
 595 partner selection in r&d alliances. *Academy of management journal*, 51(2):315–334, 2008.  
 596

597 Dawei Li, Zhen Tan, and Huan Liu. Exploring large language models for feature selection: A  
 598 data-centric perspective. *ACM SIGKDD Explorations Newsletter*, 26(2):44–53, 2025a.

599 Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems:  
 600 workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024.

601

602 Zhenkun Li, Lingyao Li, Shuhang Lin, and Yongfeng Zhang. Know the ropes: A heuristic strategy  
 603 for llm-based multi-agent system design. *arXiv preprint arXiv:2505.16979*, 2025b.

604

605 Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shum-  
 606 ing Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through  
 607 multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Pro-  
 608 ceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp.  
 609 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguis-  
 610 tics. doi: 10.18653/v1/2024.emnlp-main.992. URL [https://aclanthology.org/2024.emnlp-main.992/](https://aclanthology.org/2024.emnlp-main.992).

611

612 London Stock Exchange Group. LSEG workspace, 2024. URL <https://www.lseg.com/>.  
 613 Database; access restricted by subscription.

614

615 Fabrice Lumineau, Wenqian Wang, and Oliver Schilke. Blockchain governance—a new way of  
 616 organizing collaborations? *Organization Science*, 32(2):500–521, 2021.

617

618 Yichen Luo, Yebo Feng, Jiahua Xu, Paolo Tasca, and Yang Liu. Llm-powered multi-agent system  
 619 for automated crypto portfolio management. *arXiv preprint arXiv:2501.00826*, 2025.

620

621 Alex Makarevich. Performance feedback as a cooperation “switch”: A behavioral perspective on  
 622 the success of venture capital syndicates among competitors. *Strategic Management Journal*, 39  
 623 (12):3247–3272, 2018.

624

625 Kyle J Mayer and Nicholas S Argyres. Learning to contract: Evidence from the personal computer  
 626 industry. *Organization science*, 15(4):394–410, 2004.

627

628 Denisa Mindruta, Mahka Moeen, and Rajshree Agarwal. A two-sided matching approach for partner  
 629 selection and assessing complementarities in partners’ attributes in inter-firm alliances. *Strategic  
 630 Management Journal*, 37(1):206–231, 2016.

631

632 Gavin Mischler, Yinghao Aaron Li, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. Con-  
 633 textual feature extraction hierarchies converge in large language models and the brain. *Nature  
 634 Machine Intelligence*, 6(12):1467–1477, 2024.

635

636 Sudhansu Shekhar Patra, GM Harshvardhan, Mahendra Kumar Gourisaria, Jnyana Ranjan Mohanty,  
 637 and Subham Choudhury. Emerging healthcare problems in high-dimensional data and dimension  
 638 reduction. In *Advanced Prognostic Predictive Modelling in Healthcare Data Analytics*, pp. 25–49.  
 639 Springer, 2021.

640

641 PitchBook Data, Inc. Pitchbook, 2024. URL <https://pitchbook.com/>. Database; access  
 642 restricted by subscription.

643

644 Daniel Potts and Michael Schmischke. Interpretable approximation of high-dimensional data. *SIAM  
 645 Journal on Mathematics of Data Science*, 3(4):1301–1323, 2021.

646

647 Thanuja Chandani Sandanayake, GAI Limesha, TSS Madhumali, WPI Mihirani, and MSA Peiris.  
 648 Automated cv analyzing and ranking tool to select candidates for job positions. In *Proceedings  
 649 of the 6th International Conference on Information Technology: IoT and Smart City*, pp. 13–18,  
 650 2018.

651

652 Reshma H Shah and Vanitha Swaminathan. Factors influencing partner selection in strategic al-  
 653 liances: The moderating role of alliance context. *Strategic management journal*, 29(5):471–494,  
 654 2008.

648 Manuel Sigle, Leon Berliner, Erich Richter, Mart van Iersel, Eleonora Gorgati, Ives Hubloue, Maxi-  
 649 milian Bamberg, Christian Grasshoff, Peter Rosenberger, Robert Wunderlich, et al. Development  
 650 of an anticipatory triage-ranking algorithm using dynamic simulation of the expected time course  
 651 of patients with trauma: modeling and simulation study. *Journal of Medical Internet Research*,  
 652 25(1):e44042, 2023.

653 Olav Sorenson and Toby E Stuart. Syndication networks and the spatial distribution of venture  
 654 capital investments. *American journal of sociology*, 106(6):1546–1588, 2001.

655

656 Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-  
 657 dimensional data. In *Proceedings of the 25th international conference on world wide web*, pp.  
 658 287–297, 2016.

659

660 Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. Magis:  
 661 Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information  
 Processing Systems*, 37:51963–51993, 2024.

662

663 Dan Wang, Emily Cox Pahnke, and Rory M McDonald. The past is prologue? venture-capital  
 664 syndicates’ collaborative experience and start-up exits. *Academy of Management Journal*, 65(2):  
 665 371–402, 2022.

666

667 Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen,  
 668 Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with  
 669 conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural  
 Information Processing Systems*, 37:137010–137045, 2024.

670

671 Erica Zhang, Ryunosuke Goto, Naomi Sagan, Jurik Mutter, Nick Phillips, Ash Alizadeh, Kangwook  
 672 Lee, Jose Blanchet, Mert Pilanci, and Robert Tibshirani. Llm-lasso: A robust framework for  
 673 domain-informed feature selection and regularization. *arXiv preprint arXiv:2502.10648*, 2025.

674 Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. Kg4diagnosis: A hierarchical multi-agent llm  
 675 framework with knowledge graph enhancement for medical diagnosis. In *AAAI Bridge Program  
 676 on AI for Medicine and Healthcare*, pp. 195–204. PMLR, 2025.

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702  
703  

## A USE OF LLMs

704  
705  
During the development of this paper, we use transformer based large language models in the fol-  
lowing aspects:706  
707  
708  
709  
710  
711  

- Reference discovery: use the deep research tools from major providers to explore relevant work and literature.
- Code assistance: use coding agents to assist developing the code base of the current work.
- Grammar check: use LLMs to detect grammar errors in the drafty version of the paper, for better displaying our results.

712  
713  

## B DATA AVAILABILITY

714  
715  
716  
717  
718  
719  
This curated dataset offers a reliable foundation, as both LSEG (London Stock Exchange Group, 2024) and PitchBook (PitchBook Data, Inc., 2024) provide clear, consistent records of first-round investments and designated lead VCs, which allows researchers to cross-verify records.. While licensing restrictions prevent us from sharing the raw data publicly, it can be accessed for research purposes upon reasonable request to the corresponding authors.720  
721  

## C FEATURE DESCRIPTION

722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
For each company’s lead investor, we extract the relevant subset of potential coinvestors from the VC pool based on the company’s industry, headquarter state, and the first-round investment year. We then construct pairwise lead VC-VC firm observations to capture prior co-investment experiences (Bellavitis et al., 2020) and geographic distance between partners (Sorenson & Stuart, 2001; Gulati et al., 2012). In addition to these network measures, we obtain detailed firm-level characteristics from LSEG Workspace (London Stock Exchange Group, 2024), including firm location (county and state) and investment preference variables such as geographic, industry, and stage preferences. These additional attributes allow us to account for both structural and preference-based drivers of syndicate partner selection, as illustrated in prior studies (Hsu, 2004; Gulati et al., 2012). Detailed description of features for VC firms is presented in Table 4.732  
733  

## D AGENT CLUSTERS

734  
735  
736  
737  
738  
739  
740  
741  
742  
In the PARTNERMAS workflow, for each experimental setting, the Planner Agent is prompted to generate a set of Specialized Agents. Specifically, it produces a list of profiles, one for each agent. Each profile specifies four elements: (i) agent name, (ii) role, (iii) abilities, and (iv) guides. Across all 14 experimental settings using the PARTNERMAS framework, a total of 9,786 profiles are generated. Since agent names can vary slightly across settings, grouping by name alone would yield an excessive number of fragmented clusters. To address this, we cluster agents using embeddings so that those within the same cluster share similar profiles, whether in roles or guides emphasizing comparable features of candidate VC firms. The clustering process proceeds as follows:743  
744  
745  
746  
747  
748  
749  
750  

- We use an embedding model “all-MiniLM-L6-v2” to create one profile vector for each specialized agent.
- We then use the “k-means” clustering method to aggregate all agents. After experiments on the number of clusters, we find that  $k = 8$  (i.e. eight clusters) balances the cluster size, similarity within clusters, and diversity between clusters, and achieve an acceptable Silhouette score at 0.290. As a result, we obtain eight clusters for all 9,786 agent names among 14 test experimental setting.

751  
752  

## E DEBATE MAS DESIGN

753  
754  
755  
The Debate MAS enhances decision quality through structured multi-phase interaction among agents. Instead of relying on a single evaluation, the system encourages critique, reflection, and oversight, leading to more robust outcomes.

756

757

Table 4: Key features used in VC co-investor shortlisting and their descriptions.

758

759

760 <b>Feature</b>	761 <b>Type</b>	762 <b>Description</b>
<i>Identifiers &amp; labels</i>		
763 companyid	764 ID	765 Unique identifier of the target company in the focal deal.
766 vcfirmid	767 ID	768 Unique identifier of the candidate VC firm.
769 leadvc	770 ID	771 Identifier of the lead VC for the focal deal.
772 real	773 Binary	774 Ground-truth label: 1 if candidate VC appears in the actual syndicate; 0 otherwise.
775 leadornot	776 Binary	777 Indicator for whether the VC is the lead investor.
778 yearquarter	779 Categorical	780 Year-quarter context of the focal deal (e.g., 2019Q3).
781 year	782 Numeric	783 Calendar year of the focal deal.
784 realsize	785 Numeric	786 Number of ground-truth co-investors in the focal syndicate.
<i>Target company attributes</i>		
787 companyindustrymajorgroup	788 Categorical	789 Major industry grouping of the target company.
790 companynation	791 Categorical	792 Target company headquarters nation.
793 companystate	794 Categorical	795 Target company headquarters state (if applicable).
796 companycity	797 Categorical	798 Target company headquarters city.
799 companyzip	800 Categorical	801 Target company ZIP/postal code.
802 companylat	803 Numeric	804 Latitude of target company.
805 companylng	806 Numeric	807 Longitude of target company.
<i>Candidate VC firm attributes</i>		
808 firmtype	809 Categorical	810 Type of investor.
811 firmnation	812 Categorical	813 VC firm nation.
814 firmstate	815 Categorical	816 VC firm state (if applicable).
817 firmcounty	818 Categorical	819 VC firm county (if applicable).
819 firmzipcode	820 Categorical	821 VC firm ZIP/postal code.
822 firmgeographypreference	823 Text	824 Stated geographic investment preferences.
823 firmindustrypreference	824 Text	825 Stated industry/sector preferences.
824 firminvestmentstagepreference	825 Text	826 Stated stage preferences (e.g., seed, early, growth).
<i>Candidate VC activity &amp; outcomes (rolling/cumulative)</i>		
827 vcfirm_dealcount_20qtr	828 Numeric	829 Deals by the VC in the past 20 quarters.
829 vcfirm_numcompinvest_20qtr	830 Numeric	831 Co-investments by the VC in the past 20 quarters.
830 vcfirmIPOcount_20qtr	831 Numeric	832 IPO exits associated with the VC in the past 20 quarters.
831 vcfirm_IPOcount_cum	832 Numeric	833 Cumulative IPO exits associated with the VC to date.
832 vcfirm_dealcount_cum	833 Numeric	834 Cumulative deals by the VC to date.
833 vcfirm_numcompinvest_cum	834 Numeric	835 Cumulative co-investments by the candidate VC to date.
<i>Network measures &amp; pairwise history</i>		
836 boncent	837 Numeric	838 Bonacich centrality of the VC in the co-investment network.
838 degree	839 Numeric	840 Degree centrality in the co-investment network.
840 pair_tie_strength	841 Numeric	842 Prior collaboration strength with the lead VC.
<i>Candidate VC geospatial</i>		
843 uszip_vc	844 Categorical	845 U.S. ZIP code of the VC (normalized field, if applicable).
844 uslat_vc	845 Numeric	846 U.S. latitude of the VC office (normalized field).
846 uslng_vc	847 Numeric	848 U.S. longitude of the VC office (normalized field).
847 uscity_vc	848 Categorical	849 U.S. city of the VC office (normalized field).
848 uscounty_vc	849 Categorical	850 U.S. county of the VC office (normalized field).

799

800

801

802

803

804

805

806

807

808

809

- **Evaluation Phase:** Each agent independently evaluates the candidate firms, producing initial scores and rationales.
- **Debate Phase:** Agents review their peers’ reasoning (not the scores) and provide agreements, disagreements, and clarifying questions. This peer-review process emphasizes justification quality and helps surface biases or overlooked factors.
- **Reflection Phase:** After the debate, each agent revisits its own evaluation in light of peer feedback, reflecting critically and optionally adjusting its decisions.

810 In the end, a dedicated supervisor agent integrates the outcomes from all phases, resolves conflicts,  
 811 and produces the final decision. The supervisor focuses on synthesizing insights across agents rather  
 812 than simple score aggregation.  
 813

## 814 F PROMPT DESIGN

### 815 F.1 BUSINESS DOMAIN GUIDANCE

816 We include the following business domain guidance to ensure that agents explicitly consider factors  
 817 such as collaboration history, industry fit, strategic alignment, financial strength, and geographic  
 818 proximity when making investment decisions:  
 819

820 Your decision should consider important dimensions like network and  
 821 collaboration history (pair tie strength with the lead company and  
 822 boncent), industry fit (firm industry preference), strategic  
 823 alignment (firm investment stage preference), financial, and  
 824 geography (distance, firm state). Your strategic guidance should  
 825 explain which of these dimensions are most critical for this specific  
 826 deal.  
 827

### 828 F.2 PROMPT FOR SINGLE AGENT

829  
 830 You are '{self.name}', {self.role}, possessing {self.ability}.  
 831  
 832 # Your Profile:  
 833 {self.profile}  
 834  
 835 # Business Hint: (when enabled)  
 836 Your decision should consider important dimensions like network and  
 837 collaboration history  
 838 (pair tie strength with the lead company and boncent), industry fit (firm  
 839 industry preference),  
 840 strategic alignment (firm investment stage preference), financial, and  
 841 geography (distance, firm state).  
 842 Your strategic guidance should explain which of these dimensions are most  
 843 critical for this specific deal.  
 844  
 845 # Investment Target Company:  
 846 {target\_profile}  
 847  
 848 # Lead Investor Profile:  
 849 {lead\_profile}  
 850  
 851 # Candidate Co-Investors to Evaluate:  
 852 {candidates\_list}  
 853  
 854 # Evaluate candidates across different dimensions based on the agent  
 855 expertise and profile.  
 856  
 857 # Scoring Guidance (1-5 scale):  
 858 - 1: Poor match, significant concerns, not recommended at all  
 859 - 2: Below average, considerable issues, generally not favorable  
 860 - 3: Average neutrality, acceptable but with clear reservations  
 861 - 4: Good candidate, strong fit with minimal concerns  
 862 - 5: Excellent candidate, ideal investment partner and highly recommended  
 863  
 864 # Task:  
 865 Evaluate each candidate and select the top {top\_k} co-investors for this  
 866 investment opportunity.

864 F.3 PROMPT FOR DEBATE MAS  
865866 **Stage 1: Initial Evaluation Prompt**

```

867 You are '{self.name}', {self.role}, possessing {self.ability}. Evaluate
868     the following candidates for potential investment:
869
870 # Scoring Guidance (1-5 scale)
871 - 1: Poor match, significant concerns, not recommended at all.
872 - 2: Below average, considerable issues, generally not favorable.
873 - 3: Average neutrality, acceptable but with clear reservations.
874 - 4: Good candidate, strong fit with minimal concerns.
875 - 5: Excellent candidate, ideal investment partner and highly recommended
876 .
877
878 # Investment Target:
879 {target_profile}
880
881 # Your Own Profile:
882 {self.profile}
883
884 # Candidates to Evaluate:
885 {candidates_data}
886
887 # Additional Context:
888 {context}
889
890 # Evaluation to be strictly JSON formatted:
891 {
892     "evaluations": {
893         "firm_id_1": {
894             "integrity_score": int (1-5),
895             "integrity_rationale": "... clear rationale why this score was given
896             ...",
897             "capability_score": int (1-5),
898             "capability_rationale": "... clear rationale why this score was given
899             ...",
900             "fit_score": int (1-5),
901             "fit_rationale": "... clear rationale why this score was given ..."
902         },
903         "firm_id_2": {
904             // ... same structure for each firm
905         }
906     }
907 }

```

903 **Stage 2: Reflection Prompt**

```

904 You are '{self.name}', ({self.role}). After reviewing your evaluations:
905
906 Evaluations:
907 {evaluations}
908
909 Context:
910 {context}
911
912 Reflect critically on the evaluations. Provide clear thoughts in strictly
913     JSON:
914 {
915     "reflection_summary": "... comprehensive reflection on possible biases,
916         assumptions, or key insights ...",
917     "improvement_suggestions": ["clear suggestions on improvement", "... more
918         thoughtful suggestion"],
919     "score_decisions": {
920         "reasoning": "... your explicit self-reflect reasoning clearly ...",

```

```

918     "stick_with_previous_score": true|false
919 }
920 }
921

```

### Stage 3: Debate Prompt

```

922 You are '{self.name}' ({self.role}). You are reviewing evaluations
923 written by your peer agents.
924
925 IMPORTANT:
926 - You are debating with OTHER AGENTS about their evaluations, not with
927   the firms being evaluated
928 - The numeric scores from your peers are intentionally hidden
929 - Focus ONLY on their **reasoning and justifications** not the scores
930 - You are reviewing ONLY your peers' evaluations, not your own
931 - The supervisor's evaluations are not included in this debate
932
933 Available peer agents to debate with: {peers_list}
934
935 All Agents' Evaluations (scores hidden):
936 {stripped_evaluations}
937
938 Context:
939 {context}
940
941 Your task:
942 - Critically analyze the reasoning about the firms from other agents
943 - For each peer agent's evaluation, you can:
944   * Agree with multiple points they made about a firm
945   * Disagree with multiple points they made about a firm
946   * Agree with some parts while disagreeing with others
947 - If something is unclear, ask specific questions directly to the
948   relevant agent
949
950 Output as strictly formatted JSON:
951 {
952   "agree": [
953     {
954       "agent_name": "name of the peer agent from available peer agents",
955       "points": [
956         "specific points you agree with about their evaluation",
957       ]
958     }
959   ],
960   "disagree": [
961     {
962       "agent_name": "name of the peer agent from available peer agents",
963       "points": [
964         "specific points you disagree with about their evaluation",
965       ]
966     }
967   ],
968   "questions": ["concise question directed clearly to agent_name about
969     their evaluation"]
970 }
971

```

### F.4 PROMPT FOR PARTNERMAS (OURS)

#### Planner Agent Prompt

*Generic*

```
You are '{name}', a {role} with {ability}.
```

972 Your task is to design a multi-agent system for evaluating potential co-  
 973 investor partnerships.  
 974 Based on the provided lead investor profile, target company profile and a  
 975 sample of candidate co-investor profiles,  
 976 determine the optimal number of specialized agents, their specific roles,  
 977 abilities, and profiles.  
 978 The goal is to create agents that can thoroughly evaluate candidates  
 979 across different, relevant dimensions to find the best co-investors  
 980 for the lead firm.

981

### 982 *Business-Domain-Guided*

983 You are '{name}', a {role} with {ability}.  
 984 Your task is to design a multi-agent system for evaluating potential co-  
 985 investor partnerships.  
 986 Based on the provided profiles, determine the optimal specialized agents  
 987 across different dimensions  
 988 (e.g., collaboration history, industry fit, strategic alignment,  
 989 financial, geography, integrity) and formulate a high-level strategic  
 990 guidance for the final decision-maker.  
 991 # Lead Investor Profile:  
 992 {lead\_profile}  
 993  
 994 # Investment Target Profile:  
 995 {target\_profile}  
 996  
 997 # Sample Candidate Co-investor Profiles (structure overview):  
 998 {sample\_candidates}  
 999 (Note: This is just a sample of 2 candidates, infer general dimensions  
   from the structure and target profile.)  
 1000  
 1001 # Your output MUST be a JSON object with TWO top-level keys: "  
   strategic\_guidance" and "agents".  
 1002 # 1. "strategic\_guidance": A concise paragraph outlining the most  
   critical factors for selecting a co-investor for THIS SPECIFIC target  
   . This is high-level advice for the supervisor.  
 1003 # 2. "agents": A JSON array of agent configurations. Each agent must have  
   a distinct profile that covers a key evaluation criterion inspired  
   by the strategic guidance.

1007

### 1008 **Specialized Agent Prompt**

1009 You are '{name}', a {role} with {ability}.  
 1010 Your specific focus is: {profile}.  
 1011  
 1012 # Investment Target:  
 1013 {target\_profile}  
 1014  
 1015 # Candidates to Evaluate:  
 1016 {candidates\_data}  
 1017  
 1018 Your task is to identify and rank the \*\*top {dynamic\_top\_k}\*\* most  
   suitable candidate co-investor companies from the total of {  
   total\_candidates} candidates for the investment target.  
 1019 Focus specifically on your area of expertise as defined in your profile  
   using a clear logical flow:  
 1020 # 1. \*\*Select Focus:\*\* State the key features you will focus on.  
 1021 # 2. \*\*Formulate Overall Strategy:\*\* Explain your overall reasoning and  
   methodology based on that focus.  
 1022 # 3. \*\*Make Decisions:\*\* Rank the top candidates according to your focus  
   and reasoning.

1025

```

1026 # Your Output MUST be a JSON object with THREE top-level keys in this
1027 specific order:
1028 - "evaluation_focus": A concise string identifying important features
1029 you are using for your analysis.
1030 - "overall_rationale": A general explanation of your ranking methodology
1031 , consistent with your stated focus.
1032 - "ranked_candidates": A list containing *exactly* the top {
1033 dynamic_top_k} candidates. Each rationale in this list must be a
1034 direct result of applying your focus and overall rationale.
1035

```

## Supervisor Agent Prompt

### *Co-investor Selection by Importance*

```

1038 You are '{name}', {role}, and you have the final say on co-investor
1039 selection.
1040 Your goal is to produce a final, ranked shortlist of exactly **{top_k}**
1041 candidates.

1042 # Strategic Guidance from Planner:
1043 # This is the high-level strategy you must follow for this specific deal.
1044 {planner_strategic_guidance}

1045 # Your Decision-Making Process (Follow these steps precisely):
1046 1. **Step 1: Identify Consensus Picks.**
1047 - Review all agent evaluations and identify candidates that are highly
1048 ranked by multiple agents.
1049 - Add the strongest consensus candidates to your shortlist first.
1050 - In your rationale, state how many consensus picks you found.

1051 2. **Step 2: Fill Remaining Slots via Conflict Resolution.**
1052 - You now need to fill the remaining slots to reach the target of **{
1053 top_k}** candidates.
1054 - Examine candidates with mixed reviews (e.g., ranked high by one
1055 agent but low or not at all by another).
1056 - Use your Agent Importance Ranking as the decisive tie-breaker. The
1057 weight of a more important agent carries significantly more
1058 weight.
1059 - Select the best of the remaining candidates based on this weighted
1060 analysis until your shortlist has exactly **{top_k}** members.

```

### *Co-investor Selection by Weight - Weight Assign Prompt*

```

1061 You are '{name}', {role}. Your goal is to make the best possible co-
1062 investor selection.
1063 Before you review the candidate rankings from your specialized agents,
1064 you must first determine the numerical weight of each agent's
1065 perspective for this specific investment opportunity.

1066 # Your Task:
1067 Assign a numerical weight to each specialized agent based on how critical
1068 their focus is for this specific target. The weights must be a
1069 floating-point number (e.g., 0.35) and **the sum of all weights must
1070 equal 1.0**.
1071

```

### *Co-investor Selection by Weight - Selection Prompt*

```

1072 You are '{name}', {role}, possessing {ability}. Your profile is: {profile
1073 }.
1074 You are the General Partner and have the strongest voice in deciding who
1075 gets invited and joins the round.
1076 Your task is to review the detailed evaluations from your specialized
1077 agents and, guided by the numerical weights you just assigned, make
1078 the final decision on the top {top_k} candidates for co-investment.
1079

```

1080  
 1081 # Your Decision:  
 1082 Based on all the information provided, and critically, \*following the  
 1083 numerical weights you established\*, select the best {top\_k}  
 1084 candidates.  
 1085 - For each candidate: Sum up the weights of all agents that recommended  
 1086 this candidate  
 1087 - Final ranking: Order candidates by their total weighted scores from  
 1088 highest to lowest

1089 *Co-investor Selection by Majority Vote*

1090  
 1091 You are '{name}', {role}, possessing {ability}. Your profile is: {  
 1092 profile}.  
 1093 You are the General Partner and have the strongest voice in deciding who  
 1094 gets invited and joins the round.  
 1095 Your task is to review the detailed evaluations from your specialized  
 1096 agents and make the final decision on the top {top\_k} candidates  
 1097 based on a \*\*majority vote\*\*.  
 1098  
 1099 # Your Decision:  
 1100 Based on all the information provided, and select the best {top\_k}  
 1101 candidates.  
 1102 - Identify which candidates are most frequently recommended by the  
 1103 different agents.  
 1104 - A candidate that appears on multiple agents' lists should be  
 1105 prioritized.  
 1106 - Your final list should represent the collective decision from your team  
 1107 of agents.

1108  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115  
 1116  
 1117  
 1118  
 1119  
 1120  
 1121  
 1122  
 1123  
 1124  
 1125  
 1126  
 1127  
 1128  
 1129  
 1130  
 1131  
 1132  
 1133