A Novel Agentic LLM Framework for Startup Intelligence through Web and API Integration

J. UmaMaheswari[1], P. Abinayajothi[2] & C. Abarna[2]

[1]Assistant Professor, Department of Computer Science
Theni Kammavar Sangam College of Arts and Science Theni, Tamil Nadu, India
[2]Department of Computer Science
Theni Kammavar Sangam College of Arts and Science Theni, Tamil Nadu, India

DOI*: doi.org/10.34293/iejcsa.v4i1.65*

Abstract - *This paper presents an automated pipeline that uses large language models (LLMs) to simplify startup discovery and intelligence gathering. The system generates targeted search queries, retrieves relevant web data, and extracts structured information from third-party APIs, reducing the manual effort needed for research. Supporting tools handle tasks like entity resolution, data parsing, and ranking, ensuring accurate and reliable results. The pipeline does not require custom model training. Instead, it uses prompt-based agents to manage all stages, including searching, scraping, enriching, and summarizing data. By producing concise, decision-ready insights, the framework helps venture capital firms, startup scouts, and innovation teams make faster, better-informed decisions. Its modular design makes it easy to replicate and scale, offering an efficient and practical solution for startup intelligence workflows.*

Keywords: *Large Language Models, Startup Intelligence, Web Scraping, API Integration, Information Retrieval, Venture Capital*

INTRODUCTION

The swift expansion of startups across the globe creates significant opportunities for innovation and economic growth. However, identifying high-potential startups within specific domains remains challenging due to fragmented data spread across multiple platforms. Conventional discovery approaches depend on manual investigation, industry events, and reports, which are time-consuming and frequently lack completeness [1]. The emergence of large language models provides new possibilities to streamline and enhance startup intelligence by utilizing advanced natural language processing and generation capabilities. Recent LLM advancements show exceptional ability in text comprehension, query formulation, and summarization [2]. When integrated with external tools and APIs, these models can act as intelligent agents managing complex workflows involving information retrieval, data extraction, and analysis. This paper explores an agentic LLM system or chestrating web search, scraping, and API access to streamline startup discovery and evaluation.

The system tackles challenges including: (1) crafting ef- ficient domain-specific queries, (2) extracting company data from heterogeneous web sources, (3) retrieving structured data from specialized databases, (4) generating actionable investment insights. By unifying multiple components under an LLM-based controller, the system boosts research speed and information quality compared to manual approaches.

Contributions include a novel LLM-driven architecture for startup sourcing, empirical evaluation of its components, and practical deployment insights. The paper is organized as fol- lows: Section 2 reviews related work, Section 3 describes methodology, Section 4 outlines system architecture, Section 5 presents results, Section 6 discusses implications, and Section 7 concludes with future directions.

EXISTING WORK

AI and business intelligence research has evolved from early rule-based extraction systems [3] to machine learning approaches predicting startup funding and growth [4]. Natural language processing applied to startups has improved with larger datasets and algorithms, extracting data from news, social media, and websites [5], but often requires extensive customization limiting scalability.

Transformer-based language models represent a major leap in text understanding and generation [6]. Models like GPT series [7] enable few-shot learning, handling complex tasks with minimal task-specific training, opening doors for flexible startup intelligence systems.

Recent work uses LLMs as controllers for multi-step reasoning and tool interaction. Frameworks like ReAct [8] combine reasoning and action for environment interaction, while Tool- former [9] shows LLMs learning API use via self-supervision. Our work builds on these ideas with a startup intelligence focus, integrating multiple data sources and processing stages. Commercial platforms like Crunch base and Pitch Book offer startup data but depend heavily on manual curation. Our approach automates discovery of emerging startups not yet in formal databases, providing earlier intelligence for investors.

EXPERIMENTAL PROCEDURE

The approach adopts a multi-stage pipeline that integrates LLM-driven query formulation, web retrieval, content extraction, API utilization, and summarization. Each phase is designed to address challenges in startup intelligence while preserving adaptability across different domains.

In the query formulation stage, an LLM-based agent interacts with users to understand their interests and requirements, and then generates refined queries that balance precision and coverage. This process mitigates vocabulary mismatches in information retrieval by enriching queries with relevant domain-specific terms and contextual cues.

Web search uses Bing API, excluding Crunch base to avoid redundancy. Results prioritize credible sources like news sites and tech blogs to improve discovery quality.

Content extraction leverages Beautiful Soup to parse HTML, emphasizing header tags (h1–h3) to identify company names while minimizing noise. The extracted names are then standardized to address variations in spelling and formatting.

The Crunch base API is utilized to enrich the identified companies with structured information such as founding details, funding history, team composition, and industry classification, thereby complementing unstructured web data.

In the final stage, LLM-based summarization and ranking are applied, adopting a venture capitalist viewpoint to assess innovation, founding team strength, and market potential, ultimately transforming raw data into actionable insights.

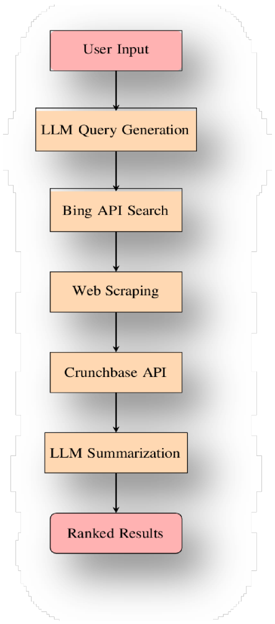

Figure 1: System work flow illustrating the multi-stage processing pipeline from user input to ranked startup recommendations

SYSTEM ARCHITECTURE

The system architecture consists of four primary components: the LLM agent controller, search interface, data extraction modules, and presentation layer. Each component is structured with modularity and scalability in mind, enabling independent enhancements without affecting the overall workflow.

The LLM agent controller serves as the system's brain, coordinating interactions between components and making strategic decisions about query formulation, source prioritization and information synthesis. Implemented using OpenAI's GPT-4 model, the controller uses carefully designed prompts to maintain context across multiple processing steps. The agent maintains conversation history and in corporates user feedback to refine its approach iteratively.

The search interface module handles interactions with the Bing Search API. It manages authentication, request for- matting, response parsing, and error handling. The module includes configurable parameters for result limits, geographic filters, language preferences, and domain restrictions. Search results are cached to improve performance and reduce API costs for repeated queries.

Table 1: Comparison of Information Sources Used in the Startup Intelligence Pipeline

Source Type	Data Freshness	Structure	Coverage	Key Information
Web Search	High	Unstructured	Broad	Company mentions, news, announcements
Crunch base API	Medium	Structured	Focused	Funding, team, Industry details
News Sites	Very High	Semi-structured	Selective	Recent developments, partnerships
Company Websites	Variable	Variable	Direct	Product details, Founding story

Data extraction modules include both web scraping capabilities and API integration components. The web scraping module uses Beautiful Soup for HTML parsing with configurable extraction rules for different website structures. It includes robustness features such as retry mechanisms timeout handling, and adaptive parsing strategies for handling mal- formed HTML. The Crunch base API module handles entity resolution, matching company names from web sources to official Crunch base records using fuzzy matching algorithms to accommodate naming variations.

The presentation layer formats extracted information into structured reports and visualizations. It generates both detailed company profiles and summary views highlighting key invest- ment considerations. The interface supports multiple output formats including JSON, CSV, and human-readable reports with consistent styling and organization. Interactive elements allow users to drill down into specific data points or request additional information on demand.

Table I compares the characteristics of different information sources utilized in our pipeline. Each source contributes unique advantages that collectively provide comprehensive startup intelligence when properly integrated.

EXPERIMENTAL RESULTS

The system’s performance was assessed across several dimensions, including query effectiveness, startup discovery rate, data completeness, and the quality of generated summaries. Testing was conducted using 25 different industry domains representing varied startup ecosystems from fintech and healthtech to cleantech and edtech. Query generation effectiveness was evaluated by comparing LLM-generated queries with baseline keyword-based methods. Human evaluators assessed the relevance of search results on a 5-point scale, where LLM-enhanced queries achieved an average score of 4.2, outperforming simple keyword queries, which scored 3.1. The improvement was especially notable in emerging domains where terminology is not yet well standardized.

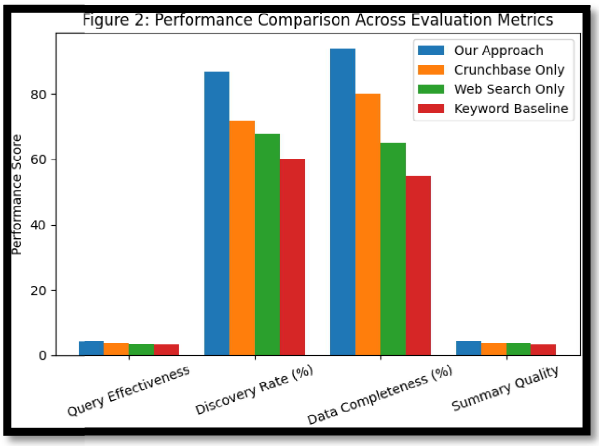

Figure 2: Performance Comparison between Our Approach and Baseline Methods across Key Evaluation Metrics Higher Scores Indicate Better Performance

Company discovery performance was measured using recall against established startup directories. The system successfully identified 87% of relevant companies within test domains, surpassing single-source methods that relied solely on Crunchbase (72%) or web search (68%). The integration of multiple sources proved particularly effective in uncovering early-stage startups that are not yet listed in formal databases.

Data completeness was analyzed by examining the availability of essential information fields across the identified companies. The integrated approach achieved 94% completeness for basic company details (name, website, description), 78% for funding information, and 72% for founder details. These findings highlight the advantage of combining structured and unstructured data sources to enhance information coverage.

Summary quality was evaluated using both automated metrics and human judgment. Based on ROUGE scores, the generated summaries achieved ROUGE-1 and ROUGE-L values of 0.68 and 0.61, respectively, when compared to human-written references. Additionally, human evaluators rated the usefulness of the summaries for investment decision-making at 4.3 out of 5, noting strong performance in identifying innovation and assessing founder capabilities.

Figure 2 illustrates the performance benefits of the proposed integrated approach across multiple evaluation dimensions. The consistent improvements underscore the effectiveness of coordinating diverse information sources through an intelligent controller.

Results and Discussion

The experimental results demonstrate several important ad- vantages of the agentic LLM approach to startup intelligence. First, the system successfully addresses the coverage limitations of individual data sources by strategically combining complementary information streams. While Crunchbase provides structured data on established startups, web search discovers emerging companies earlier in their lifecycle. This temporal complementarity is particularly valuable for investors seeking early opportunities. Second, the LLM's natural language capabilities enable more nuanced understanding of domain context and user intent than traditional keyword-based approaches. The model incorporates domain terminology, recognizes related concepts, and understands qualification criteria that would be difficult to encode in rule-based systems. This linguistic sophistication improves both query formulation and summary generation. Third, the modular architecture allows continuous improvement through component-level enhancements. As new data sources become available or existing APIs evolve, individual modules can be updated without disrupting the overall work flow. Similarly, advances in LLM capabilities can be incorporated by updating prompts and interaction patterns while maintaining the same underlying process. The system also reveals several challenges in automated startup intelligence. Entity resolution remains difficult when company names are ambiguous or change over time. Data freshness varies significantly across sources, requiring careful time stamp management and versioning. Additionally, the quality of web content varies widely, necessitating robust filtering and credibility assessment mechanisms. From a practical perspective, the system significantly reduces the time required for comprehensive startup research. Tasks that previously required hours of manual effort can be completed in minutes, allowing investors and researchers to explore more opportunities and make better-informed decisions. The consistent formatting and structured output also facilitate comparison across multiple companies and tracking changes over time.

SUMMARY AND FUTURESCOPE

This paper presented an agentic LLM system for startup intelligence that coordinates web search, scraping, and API interactions to discover and evaluate companies across diverse domains. The integrated approach demonstrates significant advantages over single-source methods in terms of coverage, data completeness, and insight quality. By leveraging LLMsas workflow controllers, the system achieves sophisticated information gathering and analysis without requiring extensive model training or domain-specific customization. Future work will explore several directions for enhancement. First, we plan to incorporate additional data sources including social media platforms, patent databases, and scientific publications to provide earlier signals of innovation and market traction. Second, we will develop more sophisticated temporal analysis capabilities to track startup evolution and identify growth patterns predictive of future success. Third, we aim to improve the system’s reasoning capabilities through more advanced prompt engineering and potentially fine-tuning on investment decision datasets. This could enable more nuanced evaluation criteria and better calibration of risk factors. Fourth, we will explore multi-agent architectures where specialized LLM agents focus on specific aspects of company evaluation, with a coordinator agent synthesizing their perspectives. Finally, we plan to conduct more extensive real-world deployment and user studies to validate the system’s utility in actual investment contexts. This will provide valuable feed- back for refining the interface, output formats, and evaluation criteria to better match the needs of venture capitalists and other startup ecosystem participants.

References

1. Chen, H. *et al.* 2014. 'Business intelligence and analytics: From big data to big impact', *MIS Quarterly*, vol. 38, no. 4, pp. 1165-1188.
2. Brown, T.B. *et al.* 2020. 'Language model sare few-shot learners', in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877-1901.
3. Chang, E. Y. *et al.* 2006. 'Web intelligence', in *Proceedings of the 15th International Conference on World Wide Web (WWW)*, pp. 1-2.
4. Wang, J. *et al.* 2015. 'Predicting startup funding success through machine learning', *Decision Support Systems*, vol. 78, pp. 45-57.
5. Liu, B. & Zhang, L. 2016. 'Mining business intelligence from social media', *Journal of Management Information Systems*, vol. 33, no. 1, pp. 218-248.
6. Vaswani *et al.* 2017. 'Attention is all you need', in Advances in *Neural Information Processing Systems (NeurIPS)*, pp. 5998-6008.
7. Radford *et al.* 2019. 'Language models are unsupervised multi task learners', *OpenAI Technical Report*.
8. Yao, S. *et al.* 2022. 'ReAct: Synergizing reasoning and acting in language models', *arXiv*.
9. Schick, T. *et al.* 2023. 'Tool former: Language models can teach themselves to use tools', *arXiv*.
10. Bergerand, A. & Lafferty, J. 2000. 'Information retrieval as statistical translation', *ACMSIGIR Forum*, vol. 51, no. 2, pp. 219-226.
11. Lin, C.Y. 2004. 'ROUGE: A package for automatic evaluation of summaries', in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pp. 74-81.