

Vela Partners: Sourcing Agent Project for Y Combinator Startups

Leona Renard Kouame - MSc Social Data Science, University of Oxford

March 2024

Abstract

Currently, there is an over-reliance on existing networks between Venture Capitalists (VCs) to discover new startups. This research investigates the application of Large Language Model (LLM) agents for start-up sourcing and recommendation in four steps: Sourcing, Information Retrieval, Data Enrichment, and Ranking. Reducing the scope to Y Combinator backed start-ups, we successfully found and ranked companies given a criteria. However, this result was inconsistent. As such, future research could explore improving information retrieval (through multiple LLM agents, improving embeddings, and integrating additional APIs), ranking (by defining a clearer ranking criteria, integrating human feedback, and pooling ranking from multiple LLM agents), and the data sources (exploring alternative founder information and exploring startup news propagation).

1 Introduction

1.1 Context

Traditionally, sourcing startups has been heavily dependent on a VC's network, with an estimated 30% of deal leads coming from former colleagues and work acquaintances (Hall, 2021).

With an estimated growth from 5 to 29 billion USD at a CAGR of 43% (WIZ.AI, 2023), Large Language Model (LLM) agents have the potential to revolutionise applications of Artificial Intelligence (Xi et al, 2023). Despite broader applications across finance and Venture Capital specifically (Xing, 2024; Zhao et al, 2024; Yu et al, 2024), harnessing LLM agents for sourcing startups is a novel approach that could reduce this reliance on existing networks.

1.2 Objective and Scope

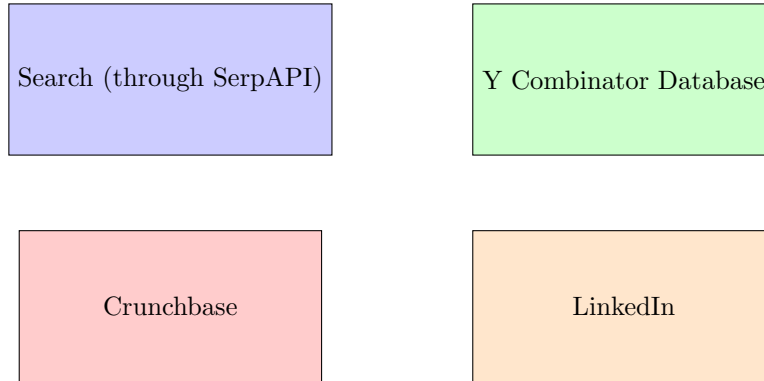
This research focuses on sourcing Y Combinator startups - representing over 4,000 companies with a combined valuation of over USD\$600 billion (Y Combinator, 2024). The goal is to develop an initial proof of concept to identify the opportunity and challenges of LLM agents for this use case.

To do this, we will assess four stages: identifying relevant startups ("Sourcing"), extracting relevant information about a given startup ("Information Retrieval"), enriching founder data with external data ("Data Enrichment"), and use the information to rank a list of startups ("Ranking").

2 Methodology

2.1 Data

The following four building blocks were used to identify startups for a given query, extract data about that startup, and enrich founder data for ranking.



2.1.1 Startup sourcing: LLM Agents + SerpAPI

Given a query, an LLM agent was used to identify relevant companies backed by Y Combinator from webpage search results to return:

url: Y Combinator url for a given startup.

Using SerpAPI as a tool to access webpage search results, we used two variations of prompts to extract startup urls:

Simplified

Find 3 {query} startup companies backed by Y Combinator site:Y Combinator.com. Output a Python list of each company's Y Combinator urls in the format ['url1','url2'] without any extra information.

Extended

"""Given the query {query}, search the web for 3 Y Combinator-backed startups matching this query. Ensure that the startups are listed on Y Combinator's website (Y Combinator.com). Retrieve the URLs of these startups from the search results and output a Python list containing these Y Combinator URLs in the format ['url1', 'url2', ...] without any extra information. Consider the following instructions: 1. Ensure that the startups retrieved are specifically backed by Y Combinator. 2. Extract only the URLs of the startups from the search results. 3. Provide the URLs in the output list without any additional information. 4. Use a reliable method to extract the Y Combinator URLs accurately. You are not allowed to make any assumptions while extracting the information. Every link you provide should be from the information given. There should be no assumptions for Links/URLs. You should not return code to do it. """

2.1.2 Web Scraping: extracting company data

Y Combinator holds a database of over 4000 startups. Given a Y Combinator url returned from the LLM agent, we first converted the page to html format using BeautifulSoup and html2text.

Given this html output (“information”), we then created an LLM agent to return a dictionary with relevant data on each company and their founders.

The prompt was as follows:

```
""" Given the company information {information} of a company on Y Combinator page in html format, I want you to extract information about the company. You are not allowed to make any assumptions while extracting the information. Every link you provide should be from the information given. There should be no assumptions for Links/URLS. You should not return code to do it.: You should extract the following text information from the html of the company page and save it as a dictionary. Include the following fields: 1. name: Full Name of the company. 2. descr: Description of the company. 3. url: URL of the company. 4. founders: Names of the Founder/Founders of the company. 5. us: 1 if the company location is in the US, 0 else. 6. minority_founder: 1 if there is a Black/ LatinX/Women founder, 0 else. 7. founding_experience: 1 if founders have previous founding experience. 8. funding: Total Funding amount to date (just the number without the $ or M sign). 9. reputable_investors: 1 if Reputable Investors back the startup - think investors like Sequoia, else 0. 10. market: one word to summarize the market segment they are in - look in the tags for things like Healthtech, AI etc and summarize what their focus is. 11. age: Company Age - find when the company was founded and calculate age from this year (2024).
```

2.1.3 Founder data enrichment

Although we had some founder data from Y Combinator, like the presence of minority founders (Black/ LatinX/ Women) and their previous founding experience, data enrichment was required for a more comprehensive investment decision.

Crunchbase

First, we used the Crunchbase Pro API.

Using the “Person” database, for each founder we extracted the following information:

- *“School”*: a list of the previous institutions they attended.
- *“Degree”*: a list of the degrees they completed.
- *“num_exits”*: the total number of successful startup exits.
- *“num_founded_organizations”*: the total number of organisations they founded previously.
- *“num_current_jobs”*: the total number of jobs they currently hold.

Diffbot API: LinkedIn profile extraction

Given the missing data, particularly for the school and degree, important markers for founder-market fit and reputation, we designed a pipeline for enrichment from LinkedIn profiles.

We first obtained the LinkedIn URL for each founder in our dictionary for a given company using the query `f'name site:linkedin.com'` and extracted the first URL using the search function from the `googlesearch` package.

Using this, we constructed the correct endpoint for the Diffbot API to return a JSON object with information listed on the founder's LinkedIn page.

Given this json for each founder, we extracted the following information:

- *"Institutions"*: a list of the previous academic institutions they have attended.
- *"Skills"*: a list of job relevant skills.
- *"Employer Names"*: a list of their previous and current employer(s).
- *"Previous Employment titles"*: a list of the positions they held at each previous employer.

2.1.4 Ranking

Given a list of dictionaries for selected startups, we created an LLM agent to rank them based on an established criteria.

Factor based

First, we defined a criteria and provided some indication for which factors should receive higher weighting in the final outcome.

""" Given a list of dictionaries {information} containing information about companies and their founders, rank the companies based on their investment potential. Output a numbered list along with a justification for your ranking. You are not allowed to make any assumptions while extracting the information. Every piece of data should be derived from the information provided, including links or URLs.

Consider the following criteria for ranking:

1. **Location (US Presence):** A higher ranking is warranted if the company is based in the US (`us: 1`).
2. **Diversity in Leadership:** Companies with minority founders should receive a higher ranking (`minority_founder: 1`).
3. **Founder Experience:** A higher ranking should be assigned if the founders have previous founding experience (`founding_experience: 1`).
4. **Funding:** Companies with total funding amount less than \$10M should receive a higher ranking. (`funding < 10`).
5. **Reputable Investors:** A higher ranking should be given to companies backed by reputable investors (`reputable_investors: 1`).
6. **Company Age:** Higher company age generally indicates stability and success. Age should be a medium-weight factor in ranking. If the company has a lower age but meets at least 80% of the criteria mentioned above, consider giving it a higher ranking.

For the information within 'founder_data', give higher weightage to founders with:

- Higher numbers of exits
- More founded organizations
- Lower current number of jobs

Ensure these factors are reflected in your ranking justification.

Output format:

1. Company Name
 - Description: Company description goes here.
 - URL: Company URL goes here.
 - Founders: Founder names go here.
 - Justification: Justification for the ranking goes here.

"""

Score-based

""" Given information about {company_1} and {company_2} in dictionary format, I want you to create a score and rank the companies based on the following criteria:

1. High weighting if `us`: 1 (if the company location is in the US) - score higher.
2. Medium weighting if `minority_founder`: 1 - score higher (reward having a minority founder positively).
3. High weighting if `founding_experience`: 1 - score higher.
4. Medium weighting if `funding` < \$10M - score higher.
5. High weighting if `reputable_investors`: 1 - score higher.
6. Low weighting for older companies - use the parameter `age` - score higher for older companies.

First output a running tally of each score.

Then output a ranked, numbered list with the company name, total score based on the information above - then provide a justification for your ranking at the end. """

3 Results

3.1 Start-up sourcing

The following results are for the query: 'climate finance'.

Simplified

• Output 1: Successful Information Retrieval

Here, the LLM agent correctly identifies three relevant and current startups which match the query.

```
[ 'https://www.YCombinator.com/companies/pachama', 'https://www.YCombinator.com/companies/bend', 'https://www.YCombinator.com/companies/pina-earth ']
```

• Output 2: Incorrect Information Retrieval

Here, despite a correctly formatted output and correct extraction of 3 startups, the LLM agents identified finance startups (rather than the full query: climate finance). Additionally, these startups are not backed by Y Combinator, so the URLs are invalid.

```
[ 'https://www.YCombinator.com/companies/airbank', 'https://www.YCombinator.com/companies/evergrow', 'https://www.YCombinator.com/companies/finli ']
```

• Output 3: Failed Information Retrieval

Here, no information retrieval was possible from the search results.

```
I couldn't find specific information on climate finance startup companies backed by Y Combinator. Would you like me to try a different search query or assist you with something else?
```

Extended

- **Output 1: (Seemingly) correct Information Retrieval**

Here the LLM Agent extracts three correct Y Combinator URLs.

```
https://www.Y Combinator.com/companies/pachama
https://www.Y Combinator.com/companies/bend
https://www.Y Combinator.com/companies/pina-earth
```

This output is maintained even with the additional context of: Find 3 query startups that are not Pachama, Bend, or Pina Earth.

- **Output 2: Failed Information Retrieval**

Here, the LLM agent is unable to generate correct URLs and instead replicates the formatting.

```
[ 'https://www.Y Combinator.com/companies/company-name1' ]
[ 'https://www.Y Combinator.com/companies/company-name2' ]
[ 'https://www.Y Combinator.com/companies/company-name3' ]
```

3.2 Founder data enrichment

Crunchbase

- **Output 1: Information Retrieval with conflicting information**

```
{ 'name': 'Pachama', 'descr': 'Pachama is a leading climate-tech
company harnessing cutting-edge technologies such as computer
vision and satellites to drive funding to effective reforestation
and conservation projects that sequester carbon, enhance
biodiversity and enrich local communities around the world.', 'url
': 'http://pachama.com', 'founders': [ 'Diego Saez Gil', 'Tomas
Aftalion' ], 'us': 1, 'minority_founder': 0, 'founding_experience':
1, 'funding': 64, 'reputable_investors': 1, 'market': 'Climate', '
age': 6 }
```

Given the following information, we see that the LLM agent correctly extracts data from Y Combinator, but this conflicts with data from other sources.

Pachama Increases Series B to \$64 Million — Pachama (Y Combinator News, 2023).

Pachama may be growing as it has recently increased its Series B funding to a total of \$64 million with a \$9 million extension (Crunchbase, 2023).

Total Funding Amount: \$88.3M (Crunchbase, 2023).

- **Output 2: Information Retrieval with incomplete information**

```
{ 'name': 'Remora', 'descr': "Carbon capture for semi-trucks. We're
building a device that captures a semi-trucks carbon emissions
directly from the tailpipe. We'll sell the CO2 to end-users, and
share that revenue with our customers, so our device will create a
new revenue stream for our customers while reducing their emissions
.", 'url': 'https://www.remoracarbon.com', 'founders': [ 'Paul Gross
' ], 'us': 1, 'minority_founder': 0, 'founding_experience': 0, '
funding': '0', 'reputable_investors': 1, 'market': 'Climatetech', '
age': 4 }
```

Here we see that when there is incomplete information on Y Combinator, the funding variable defaults to zero, even when another source has the information.

Remora raised an undisclosed amount / Series A from Voyager Ventures (Crunchbase, 2024).

Remora - Latest Deal Amount: \$5.5M (Pitchbook, 2024).

3.3 Ranking

Factor-based Using a pre-defined criteria, the LLM agent is able to make a sound justification for ranking a given list of startups given company and founder information.

1. Pachama

- Description: Pachama is a leading climate-tech company harnessing cutting-edge technologies such as computer vision and satellites to drive funding to effective reforestation and conservation projects that sequester carbon, enhance biodiversity, and enrich local communities around the world.
- URL: <http://pachama.com>
- Founders: Diego Saez Gil, Tomas Aftalion

Justification

Pachama ranks highest due to its US presence, previous founding experience, substantial funding of \$64M, backing by reputable investors, and a decent age of 6 years. The founders also have experience with 3 founded organizations and 1 current job.

2. Valor Water Analytics

- Description: Valor Water Analytics delivers a SaaS platform to utilities to detect...
- URL: <http://valorwater.com>
- Founders: PhD Christine E. Boyle

Justification

Valor Water Analytics ranks second due to its US presence, presence of a minority founder, previous founding experience, and a decent age of 11 years. Although it lacks funding and reputable investors, the founder's experience with a PhD and leadership in a specialized field adds value.

3. Enerjazz

- Description: Enerjazz is building a battery swapping network (like gasoline stations) for 2M+ commercial electric vehicles in India.
- URL: <https://www.enerjazz.tech/>
- Founders: Ravi Bharihoke, Pratik Ranjan Das

Justification

Enerjazz ranks third due to its lack of US presence, lack of minority founders, previous founding experience, and lack of funding. However, it is backed by reputable investors and has a decent age of 6 years. The founders have experience with 1 founded organization and 1 current job.

Score-based Given a criteria, an LLM agent can rank companies, providing a running tally of the score. However, without defining the weights for each factor, the score generated is unclear.

Here, the output is generated from simulated data to see how results differ across factors defined in our criteria.

XYZ Technologies:

US: 1

Minority Founder: 0

Founding Experience: 1

Funding: \$5 million (medium)

Reputable Investors: 1

Age: 3

Total Score for XYZ Technologies: 1 (US) + 1 (Founding Experience) + 2 (Funding) + 1 (Reputable Investors) + 3 (Age) = **8**

ABC Innovations:

US: 1

Minority Founder: 1

Founding Experience: 1

Funding: \$15 million

Reputable Investors: 0

Age: 2

Total Score for ABC Innovations: 1 (US) + 1 (Minority Founder) + 1 (Founding Experience) + 0 (Funding) + 0 (Reputable Investors) + 2 (Age) = **5**

Ranked List:

XYZ Technologies - Total Score: 8

ABC Innovations - Total Score: 5

Justification for Ranking

XYZ Technologies scored higher due to being located in the US, having founding experience, receiving funding less than \$10 million, having reputable investors, and being older in age. These factors indicate a strong foundation and potential for growth in the healthcare industry. ABC Innovations, while also a promising startup with a minority founder, scored lower due to higher funding, lack of reputable investors, and being younger in age.

4 Discussion

4.1 Significance

Using a combination of LLM agents, web scraping, and API calls, we were able to source startups backed by Y Combinator and rank them given a criteria. This research mitigates the challenge from the lack of complete information on a single source. The most significant contribution was the process of founder data enrichment by supplementing Y Combinator information with Crunchbase and LinkedIn. Here, we generated a richer data source from which to compare potential start-up investment options.

Although limiting the scope of the study to Y Combinator backed startups helped define a clear pipeline for web scraping, this provided a unique challenge for the information retrieval at the initial sourcing stage. Here, we found inconsistent sourcing results ranging from correctly identifying startups from Y Combinator, to hallucinated urls and a failure to generate urls entirely.

4.2 Limitations and future research directions

1. Information Retrieval

As mentioned, the biggest limitation was the consistency and accuracy of the information retrieval, even with an extended prompt discouraging LLM hallucination.

To mitigate this, future research could explore:

- *Combining multiple LLM agents in the sourcing process.* As developed in the babyagi implementation (GitHub, 2024), we could aggregate the findings from multiple LLM sourcing agents, evaluate the quality, and validate with an external dataset to ensure accuracy.
- *Using embeddings.* This can be particularly useful in cases where no relevant startups were identified.
- *Additional APIs.* This use case implemented SerpAPI as a tool; however, this could be pooled with BingSearchAPI, DuckDuckGo API, and more web search tools to improve the range of companies found.

2. Ranking

Although the LLM Agent produced a mostly comprehensive justification for ranking, again, this was inconsistent.

To mitigate this, future research could explore:

- *Defining a clearer ranking criteria.* For example, aligning the founder data evaluation with the Levels used at Vela Partners (Moneyball, 2024). Furthermore, specific weights could be defined as below.
- *Integrating human feedback.* A simple implementation could prompt user feedback, generate a labelled dataset with the original versus user ranking and deploy an algorithm like RankNet to retrain the weights defined in an original criteria (Pobrotyn et al, 2020).
- *Pooling the outputs from multiple LLM Agents.* Using a similar pooling method as the babyagi implementation (GitHub, 2024) to evaluate the soundness of the justification and generate a harmonised ranking.

3. Data Sources

Although combining data from both Crunchbase and LinkedIn enriched the context from a scraped Y Combinator page, conflicting information led the LLM agent to rank based on incomplete/ inaccurate information.

To mitigate this, future research could explore:

- *Pooling estimates for funding across multiple sources.* Especially when funding is defined as 0, an LLM agent should be prompted to look at external sources like PitchBook to validate and return correct information.
- *Extending founder data enrichment.* Applying an improved information retrieval process, we could harness founder’s blogs, news articles, and more to more clearly evaluate their character and founder-market fit.
- *Propagation of information about a given startup.* Inspired by methods to identify fake news (Raza and Ding, 2022; Zhao et al, 2020), to evaluate the virality of a startup, a graph could be constructed to model sharing behaviour on platforms like X and GitHub.

5 Conclusion

We explored a novel approach to start-up sourcing through the application of LLM agents. Using a combination of search, web-scraping, and API calls to enrich founder data, we successfully ranked a list of startups given a query for Y Combinator backed startups. However, given the inconsistency of this accurate result, future research has several opportunities. The most important is improving the information retrieval pipeline - this could be achieved by combining multiple agents, harnessing embeddings and/or using additional APIs to handle inconsistent data on Y Combinator. Furthermore, to improve the ranking and information sources, future research may explore defining a clearer criteria, and integrating human feedback.

References

- Combinator, Y. (2023a). *Pachama: Restoring nature to solve climate change*. Retrieved 2024-03-15, from <https://www.ycombinator.com/companies/pachama>
- Combinator, Y. (2023b). *The YC Startup Directory*. Retrieved 2024-03-15, from <https://www.ycombinator.com/companies>
- Crunchbase. (2023a). *Pachama - Crunchbase Company Profile & Funding*. Retrieved 2024-03-15, from <https://www.crunchbase.com/organization/pachama>
- Crunchbase. (2023b). *Remora - Crunchbase Company Profile & Funding*. Retrieved 2024-03-15, from <https://www.crunchbase.com/organization/remora-f760>
- Hall, M. (2021). *This Is How Venture Capitalists Find And Fund The Next Billion-Dollar Company [Report]*. Retrieved 2024-03-15, from <https://www.forbes.com/sites/markhall/2021/03/16/this-is-how-venture-capitalists-find-and-fund-the-> (Section: Careers)
- PitchBook. (2024). *Remora Company Profile: Valuation, Funding & Investors*. Retrieved 2024-03-15, from <https://pitchbook.com/profiles/company/460016-92>
- Pobrotyn, P., Bartczak, T., Synowiec, M., Białobrzewski, R., & Bojar, J. (2021, May). *Context-Aware Learning to Rank with Self-Attention*. arXiv. Retrieved 2024-03-15, from <http://arxiv.org/abs/2005.10084> (arXiv:2005.10084 [cs]) doi: 10.48550/arXiv.2005.10084
- Research, V. (2024). *Vela Research (March 2024)*. Retrieved 2024-03-15, from <https://docs.google.com/document/d/123qo5bBgL2XJ45UFYfuVMvo4cnSoR6TdfW6UE-whmdc/edit?pli=1usp=embed>
- WIZ.AI. (2024, January). *How LLM Agents are Unlocking New Possibilities*. Retrieved 2024-03-15, from <https://www.wiz.ai/how-llm-agents-are-unlocking-new-possibilities/>
- Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... Gui, T. (2023, September). *The Rise and Potential of Large Language Model Based Agents: A Survey*. arXiv. Retrieved 2024-03-15, from <http://arxiv.org/abs/2309.07864> (arXiv:2309.07864 [cs]) doi: 10.48550/arXiv.2309.07864
- Xing, F. (2024, January). *Designing Heterogeneous LLM Agents for Financial Sentiment Analysis*. arXiv. Retrieved 2024-03-15, from <http://arxiv.org/abs/2401.05799> (arXiv:2401.05799 [cs, q-fin]) doi: 10.48550/arXiv.2401.05799
- Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., ... Khashanah, K. (n.d.). FINMEM: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design.
- Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., ... Liu, T. (2024, January). *Revolutionizing Finance with LLMs: An Overview of Applications and Insights*. arXiv. Retrieved 2024-03-15, from <http://arxiv.org/abs/2401.11641> (arXiv:2401.11641 [cs]) doi: 10.48550/arXiv.2401.11641
- Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., ... Havlin, S. (2020, December). Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1), 7. Retrieved 2024-03-15, from https://epjds.epj.org/articles/epjdata/abs/2020/01/136882020_Article24/136882020_Article24.html (Number: 1 Publisher: Springer Berlin Heidelberg) doi: 10.1140/epjds/s13688-020-00224-z