

Query Understanding via LLM

From Ideation to Full Production



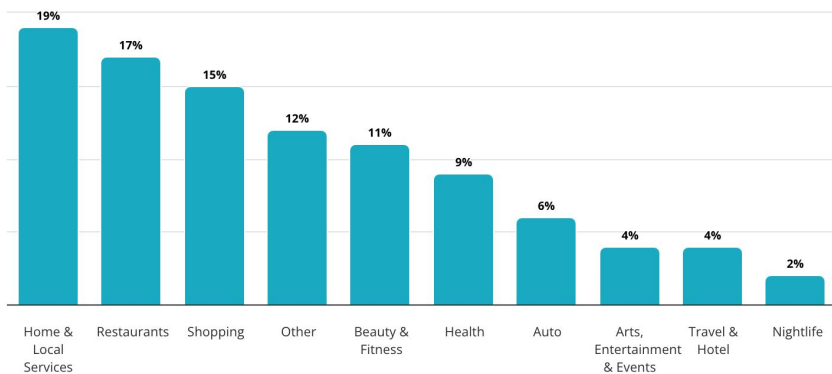
Ali Rokni

Search Quality Tech Lead

Yelp

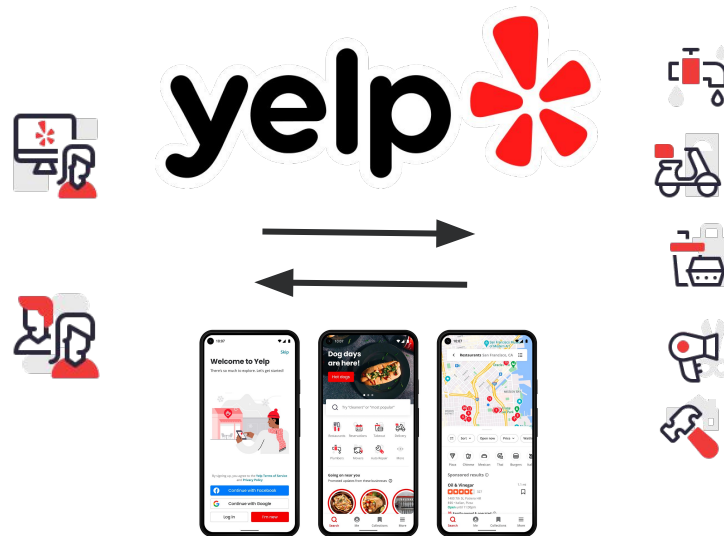
Connecting people to the great local businesses

- 32M Unique Devices*
- 287M Cumulative Reviews**



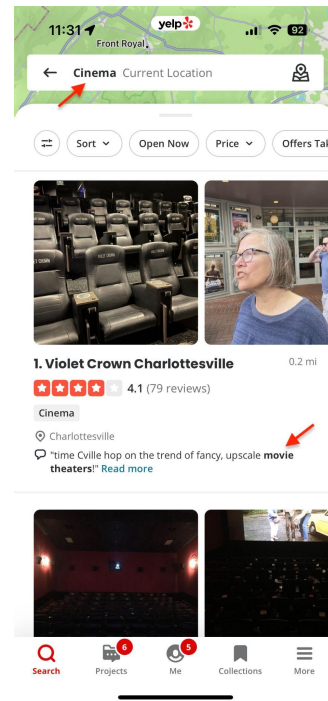
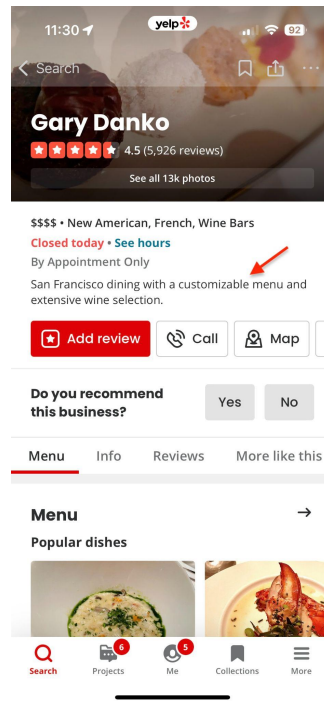
* Average monthly 2023

** As of December 31 2023



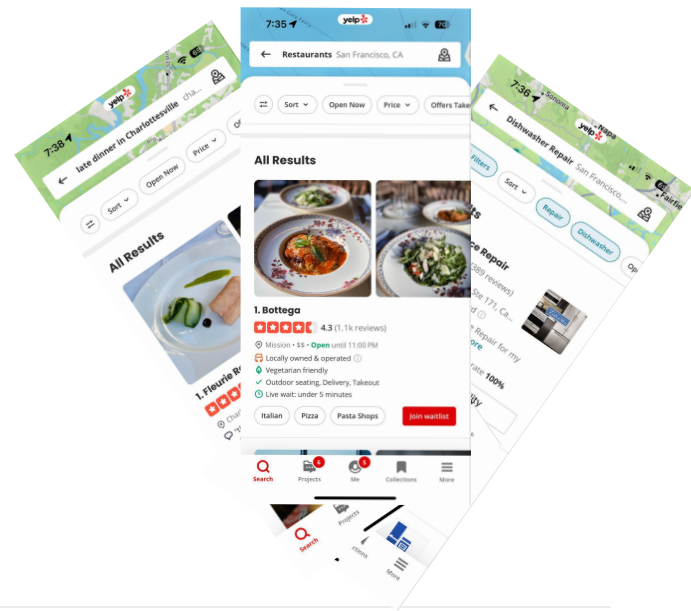
LLM at Yelp

- LLM at Yelp
 - Enhancing snippets
 - Business summaries
 - Services concierge
 - ...
- Query Understanding
 - The pioneering project
 - Laid the groundwork for Yelp's innovative use of LLMs



Query Understanding

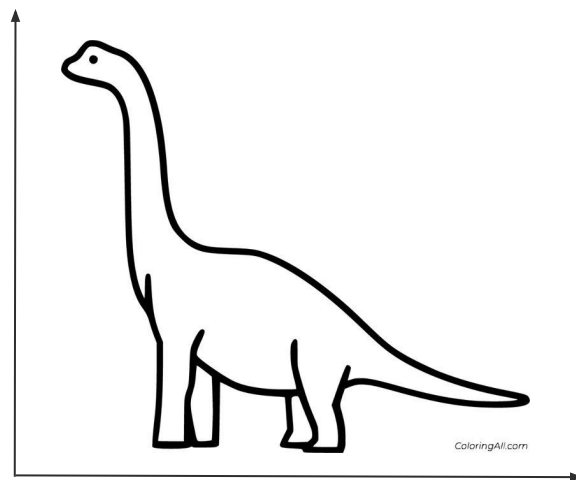
- Understanding the user intent
 - Specific category of businesses (e.g. Restaurants)
 - Particular dish or service (e.g. Sushi, dishwasher repair)
 - Specific business (e.g. Gary Danko)
 - For a specific location/time (e.g. Late Dinner in SF)
 - Is the query misspelled
 - ...
- Natural Language Understanding tasks



What is Special about Query Understanding

- Keyed by the query
- Low amount of text to be processed
- Power law query distribution

$f(query) \rightarrow static\ response$

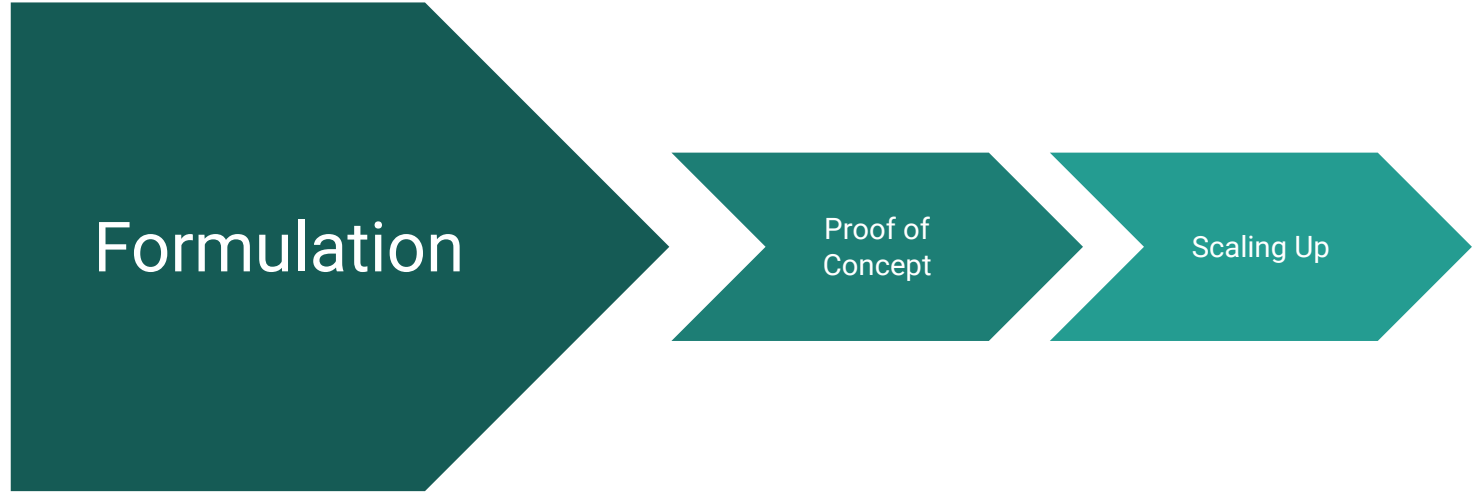


A Generic Approach



**Query
Segmentation**

**Snippet
Highlighting**



Formulation

- Mostly in LLM playground
- Is LLM an appropriate tool?
- What is the scope?
- Combining tasks?



Are There Any Opportunities for RAG

- Providing more context
- What information besides the query
 - Business categories
 - Business names
 - Query location
 - ...



Creative and Iterative Process

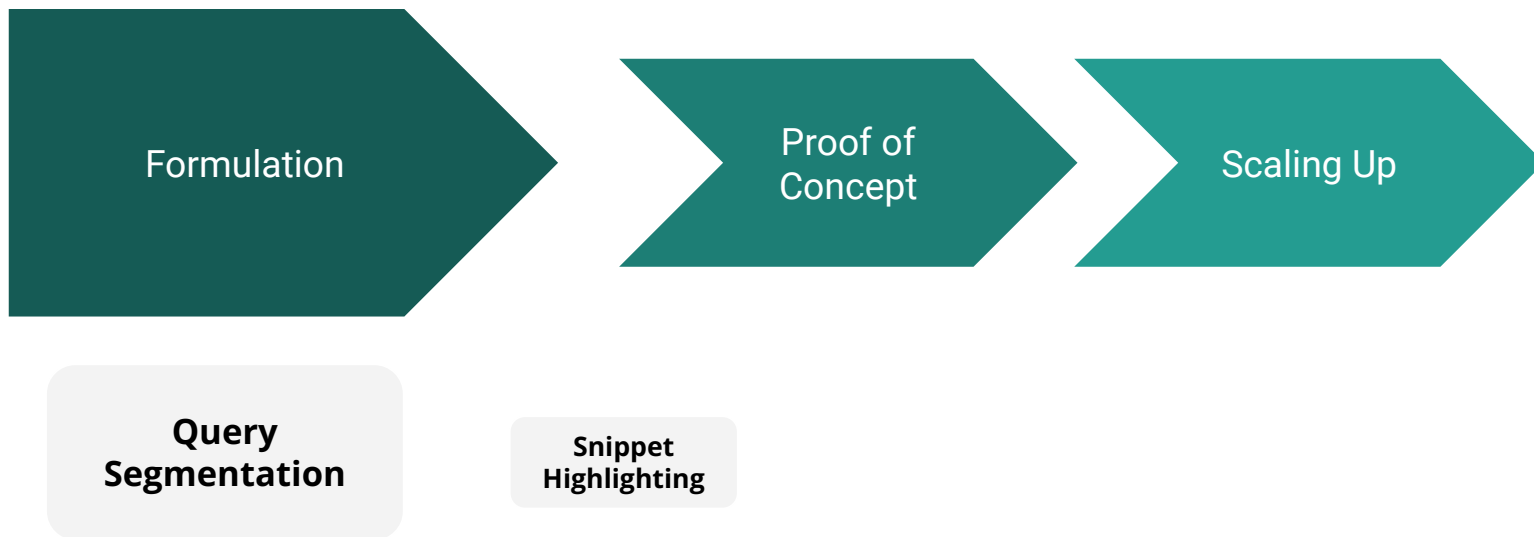
- Using the most powerful LLM
 - GPT-4, CLAUDE3 OPUS, ...

healthy fod near me => {topic} healthy fod {location} near me



- Iterative process
 - Example by example
 - Improving the example
 - Finding new teachable examples
 - Iterating on input and output
 - Possibility of changing the task scope
 - To consider the time/cost budget

Query RAG Combining spell correction
healthy fod near me [local foods] => {topic} healthy food {location} near me
[spell corrected - high] // {topic} healthy fod {location} near me



Query Segmentation

pet friendly sf restaurants open now



Topic

Location

Topic

Time

Query Segmentation

- Scope: Iteratively selecting the classes
 - topic, name, location, time, question, and none

chicago riverwalk hotels => {location} chicago riverwalk {topic} hotels
grand chicago riverwalk hotel => {name} grand chicago riverwalk hotel

- Combining spell correction

healthy fod near me => {topic} healthy food {location} near me [spell corrected - high]



Query Segmentation

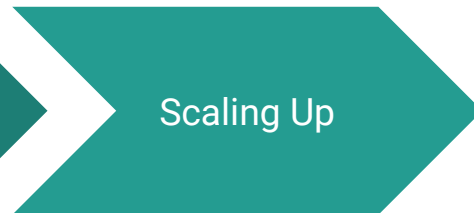
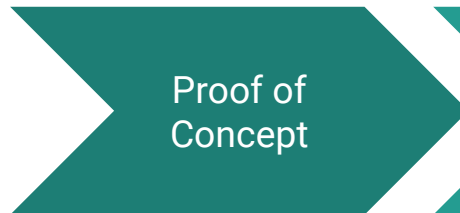
- RAG
 - names of up to 2 most viewed businesses
- Applications
 - Implicit location rewrite
 - Name intent detection
 - Auto-enable filters
 -





Query
Segmentation

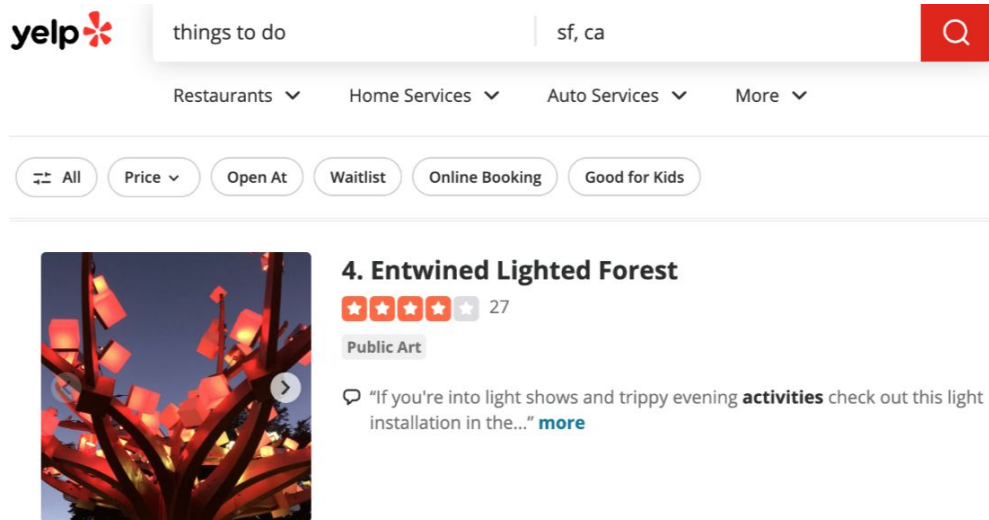
Snippet
Highlighting



Snippet Highlight

- A multi-step process
 - Selecting reviews
 - Ranking snippets
 - Highlighting query terms

The system should do much more than just relying on query terms



yelp

things to do sf, ca

Restaurants Home Services Auto Services More

All Price Open At Waitlist Online Booking Good for Kids

4. Entwined Lighted Forest

★★★★☆ 27

Public Art

“If you're into light shows and trippy evening **activities** check out this light installation in the...” [more](#)

Snippet Highlight

- Does our generic approach work here? It depends
 - Using LLM to take over entire process of selection, ranking and highlighting ✗
 - Generating the most helpful terms to match on in business reviews ✓

vegan burgers near me: *vegan burger, veggie burger, vegan, impossible burger, beyond burger, tofu burger, plant-based, vegetarian*

- creative task



Snippet Highlight

- Low bar for inclusion
 - Better to show *vegetarian* or *plant-based* for **vegan burger** than showing nothing
- RAG
 - Most relevant business categories

`best things to do with kids : [childrensmuseums, zoos, parks]`
- Iterative process
 - Input & output may evolve



Iterative Process

Input & output may evolve

- May 2022

Query: healthy food

Key concepts: healthy food, healthy, organic

- March 2023

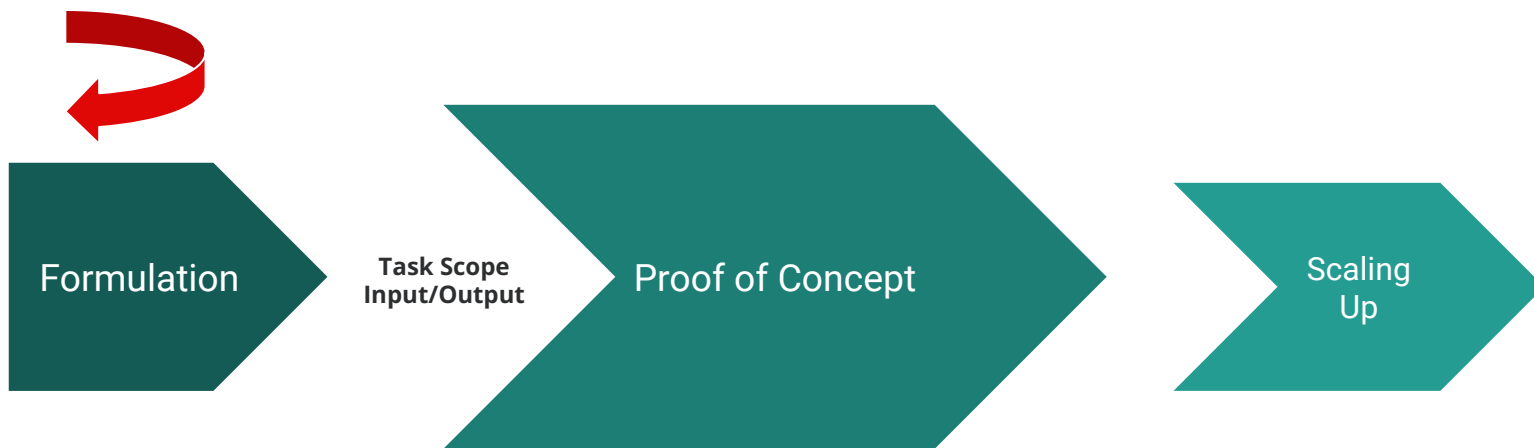
healthy food -> healthy food, healthy, organic, low calorie, low carb

- September 2023

healthy food -> healthy food, healthy options, healthy | nutritious, organic, low calorie, low carb, low fat, high fiber | fresh, plant-based, superfood

- What is the scope of each iteration?

- What is the time and budget constraints?



**Query
Segmentation**

**Snippet
Highlighting**

Proof of Concept

How does our formulation work in practice

- Building a non real time approach
 - Why it is feasible for query understanding?
 - Use most powerful LLM model
 - Pre-compute 100K common queries
 - Cache the result
 - Limited cost and no latency concerns
- Offline testing
- A/B experiment



Offline Testing

Different Approaches

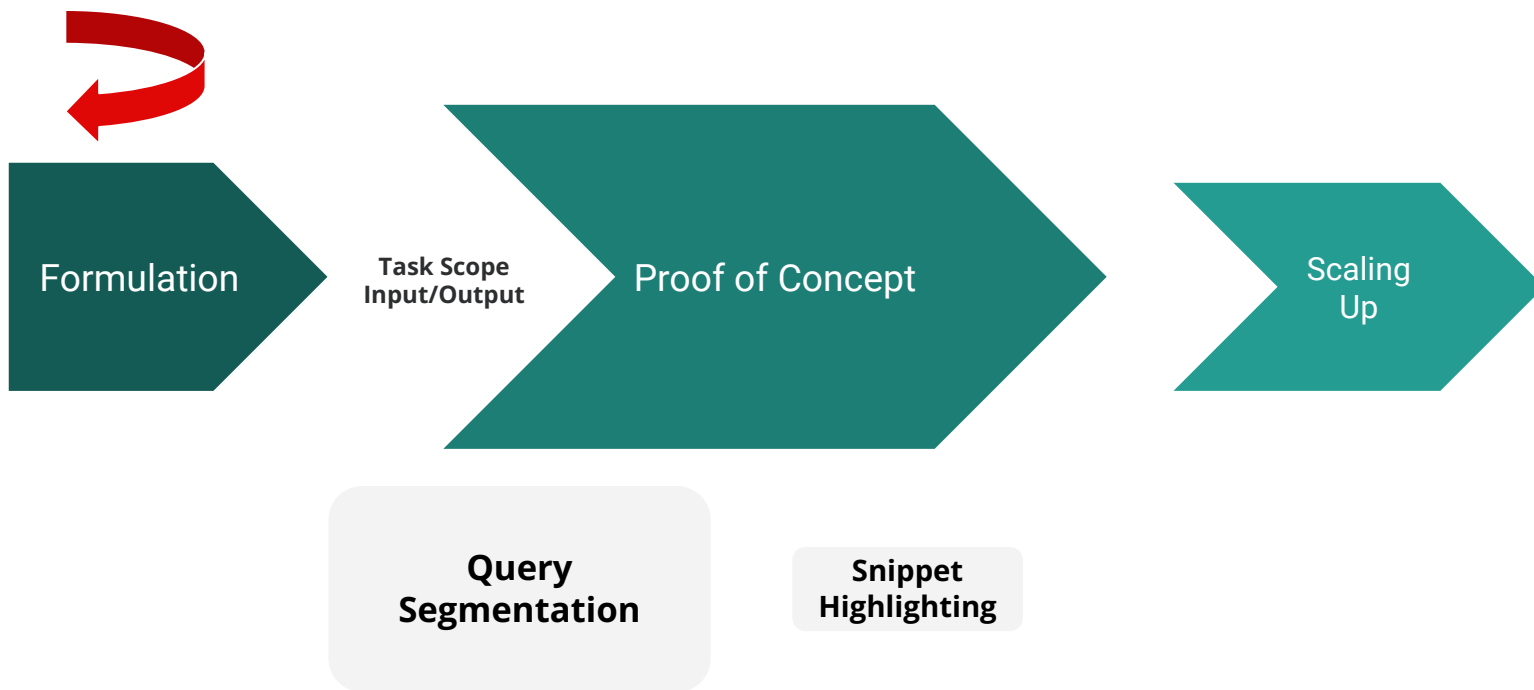
- Qualitative vs quantitative
- Isolated vs in the system

Examples

- Human expert annotation: very subjective
- Quantitative analysis
- Accuracy on specialized datasets
- Impact on downstream tasks
- Quantitative and qualitative comparison of search ranking



A/B



Query Segmentation

Implicit Location rewrite

Best Restaurants in san francisco



Topic



location

Best Restaurants in san francisco

New York, NY



Best Restaurants in san francisco

San Francisco, CA



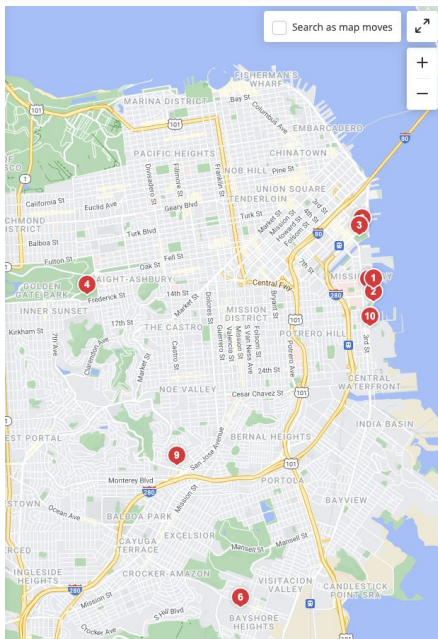
Implicit Location Rewrite

restaurants in chase center

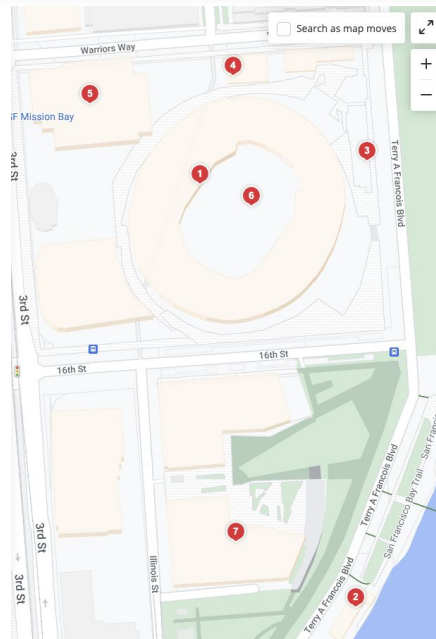
san francisco, ca

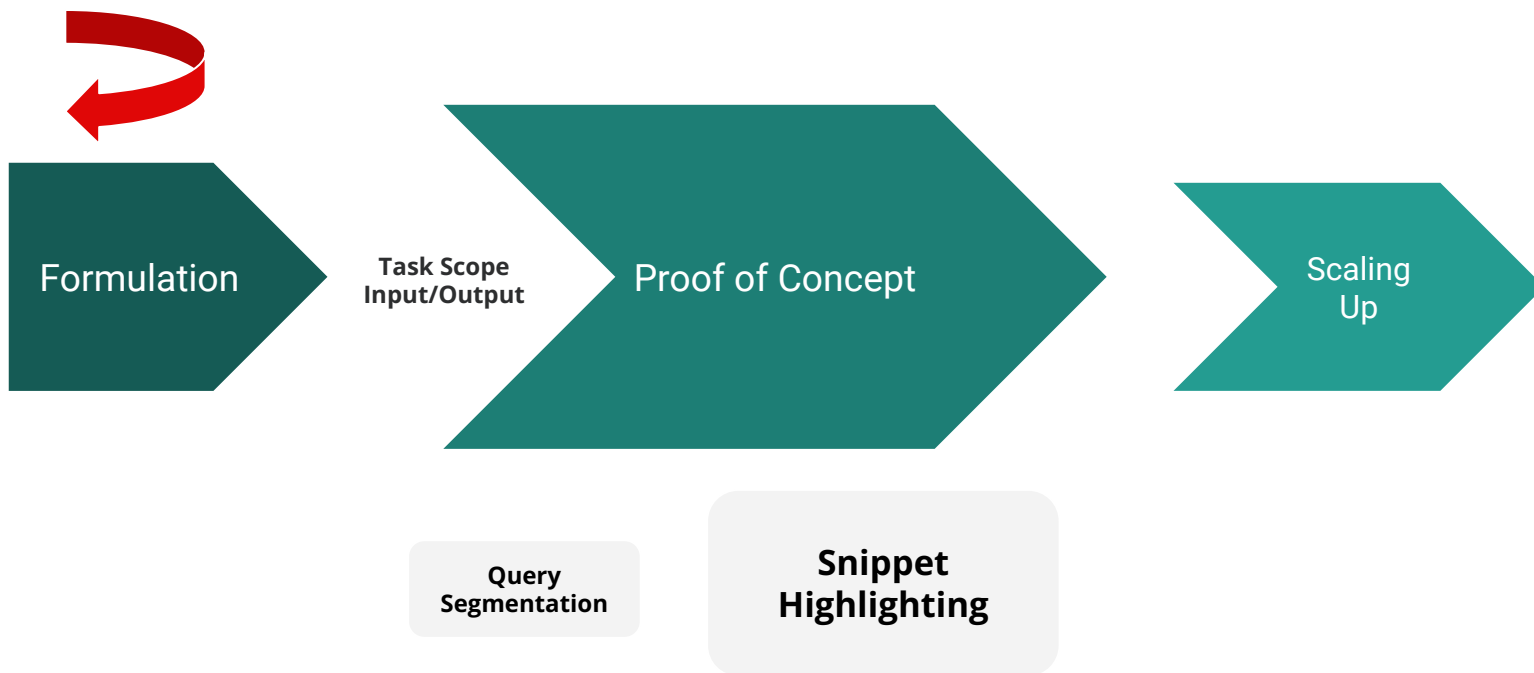


Status Quo



Treatment

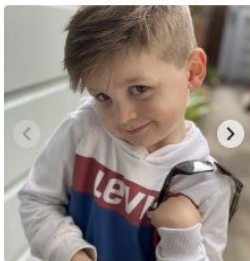




Snippet Highlight

cheap haircut

San Francisco, CA



5. Geary Salon

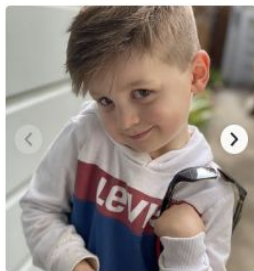
★★★★☆ 212

Hair Salons \$ • Laurel Heights

Closed until 9:30 AM

“First time here today...was recommended by my son and nephews. Had my **haircut** by Anthony and he” [more](#)

Status Quo



5. Geary Salon

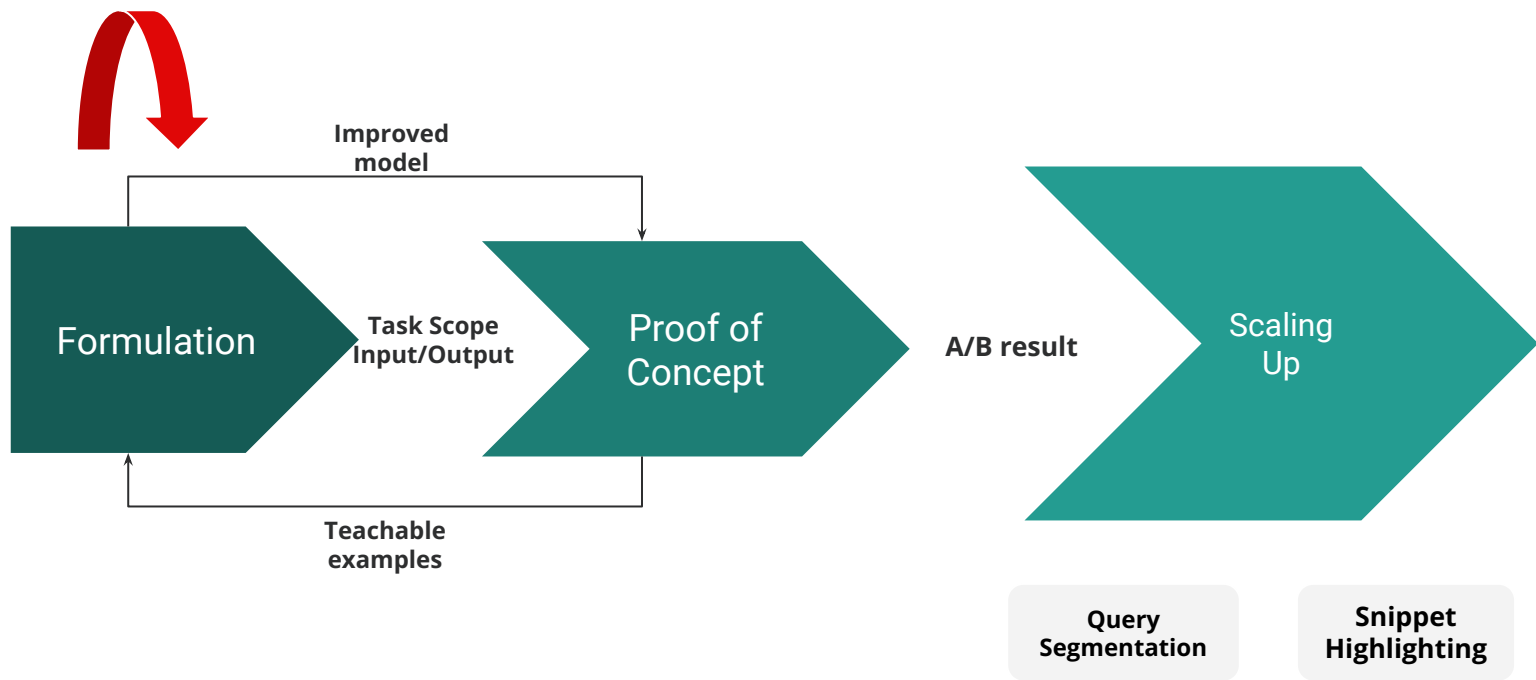
★★★★☆ 212

Hair Salons \$ • Laurel Heights

Closed until 9:30 AM

“Amazing service, **great price**, and an highly skilled haircut is going to make this my go to spot in...” [more](#)

Treatment



Scaling Up

A multi-step process

Build fine tuning dataset

- Iterative
- Find informative group of examples
- 2K - 5K examples
- Human relabel

Fine tuning with smaller LLMs

- Smaller LLM
 - GPT-3.5-turbo
- Higher quality
- 100x cost saving
- Cache-based system such as key/value DBs

Fine tune on smaller LMs

- Realtime
 - BERT, T5, ..
- Splitting combined tasks
- Refactoring legacy flows

Result

Step	Proof of concept (PoC)	PoC with fine tuned model	Top 10-100M queries
Traffic coverage	30 - 40%	30 - 40%	> 90%
Incremental improvement	+X.a%	+Y.b%	+Z.c%

- X, Y and Z are a representation of a single digit improvements over the previous step
- Numbers vary from metric to metric and from platform to platform but magnitude is similar

In progress and Future Work

- Reusing existing LLM responses for related tasks
 - Snippet highlights for retrieval and ranking
 - Similar performance gain in PoC
- Retrieval and Ranking
 - More complex system
 - In progress of building architecture
- Relevance Evaluator



Thank you

alirokni@yelp.com

