February 2024

Explainable Semantic Retrieval Using Dual Encoder Large Language Models

Marco Bonechi

Explainable Semantic Retrieval Using Dual Encoder Large Language Models

ABSTRACT

Semantic matching utilizing large language models (LLMs) to convert text or images into embeddings and scoring them can outperform keyword matching in various ways by matching on meaning rather than word equality. However, semantic matching lacks explainability. This disclosure describes dual-encoder LLM techniques to confer explainability to LLM-based semantic matches within information retrieval systems. Semantic meanings are attached to abstract mathematical embeddings to generate gravitational fields that enable dynamic, high-quality information retrieval as measured by precision/recall, query-understanding, concept-matching, speed, scalability, etc. while providing justifications and user-visible corroborations of search results. Information retrieval is also improved in diversity, personalization, and efficiency, with high query throughput at low latency.

KEYWORDS

- Semantic match
- Semantic retrieval
- Large language model (LLM)
- Dual encoder
- Keyword matching
- Explainability
- Domain taxonomy
- Data-driven taxonomy
- Query modifier

BACKGROUND

Semantic matching utilizing large language models (LLMs) to convert text or images into embeddings and scoring them can outperform keyword matching in various ways by matching on meaning rather than word equality. Semantic matching can improve the precision/recall characteristics of information retrieval systems and scales well across content types and languages.

However, keyword matching possesses a characteristic referred to as explainability that LLM matching currently lacks. Explainability is the ability to pinpoint the part of the query matched by the result of a search. Explainability enables optimization of an information retrieval system by analyzing the results and modifying scores based on user preferences, such as:

- prioritizing results that align with some or all query intents;
- categorizing query intents into essential requirements and desirable preferences;
- giving preference to results that additionally match personalized tokens;
- distinguishing between results based on the topics they cover, for example:
    - favoring results that encompass additional topics, making them more informative;
    - preferring results that exclusively and exhaustively cover the query topics; etc.

Although traditional keyword-based search has explanatory power, it lacks the capabilities of LLMs to scale naturally to any grammar and to vocabulary across natural languages. Another problem with keyword-based search is related to the interpretation of word groupings. For example, in the query q=[kids friendly restaurants] it is not clear if the word ‘friendly’ is associated with the term ‘restaurants’ or the term ‘kids.’

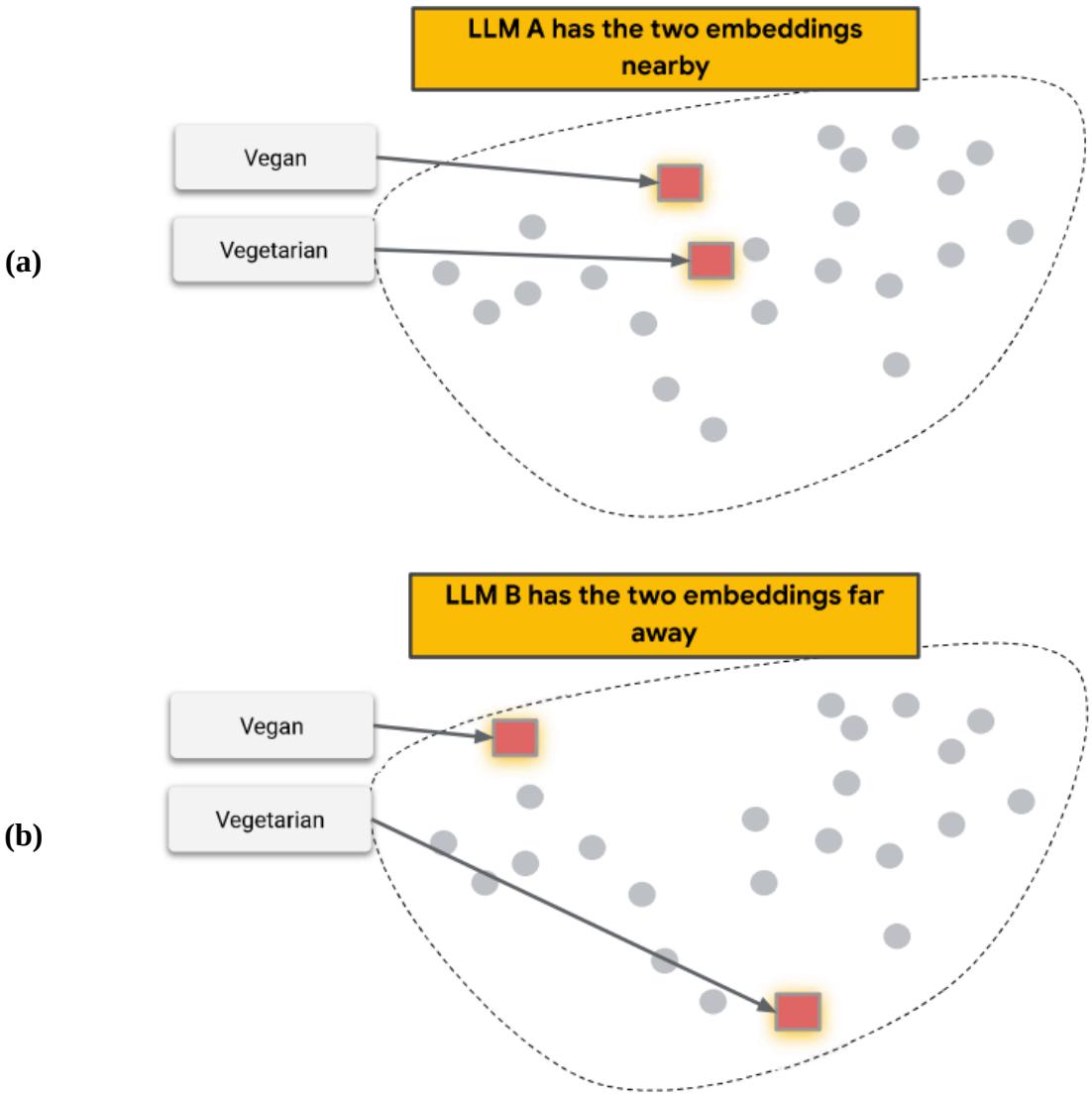


Fig. 1: Embedding spaces generated by different LLMs can have differing distances between the same pair of concepts: (a) LLM A - the embeddings for 'vegan' and 'vegetarian' are relatively close to each other; (b) LLM B - the embeddings for 'vegan' and 'vegetarian' are relatively distant from each other

Information retrieval based on dual-encoder LLMs works by generating a similarity score between embeddings of the query and of the documents being searched. However, both embeddings depend on the training objective of the encoding LLM, with distance scores taken as a relative measure within the result set. For example, as illustrated in Fig. 1, one LLM can place 'vegan' and 'vegetarian' closer in embedding space and therefore considered semantically

similar, since they are both attributes of dishes (or preferences of people), while a different LLM may consider these terms different since they are not synonymous and place them farther apart in embedding space.

DESCRIPTION

This disclosure describes techniques to confer explainability to semantic matches within information retrieval systems based on scores produced by dual-encoder LLMs. The superior precision/recall and scaling advantages of semantic matching are retained while the retrieval systems are enhanced with the abilities to provide explanations (e.g., user-visible corroborations or justifications) for search results, to diversify results, and to personalize results (if requested by the user and based on user-permitted data). Information retrieval is thereby improved in quality as well as efficiency - higher queries per second, low latency, and scale. Certain definitions follow.

Gravity point

A gravity point is an embedding whose meaning or role is known. The meaning or role of an embedding is content, such as text, that is focused and expresses a specific intent or identifiable concept or attribute. The gravity point is thus more than an abstract (mathematical) vector; it is also a concept ordinarily understood by humans. A gravity point can be used to compute a relevance score during information retrieval.

The meaning of the gravity point can depend on the search domain. For example, in local search at or near a geographic location, a gravity point, ordinarily a vector in an abstract mathematical space, can be mapped to human-recognizable concepts such as ‘cuts men’s hair,’ ‘caters to solo travelers,’ ‘dog-friendly brunch restaurant,’ etc., which can also serve as factors upon which decisions can be made. The semantic meaning of a gravity point can be derived from

a taxonomy of concepts pertaining to the search domain. For example, the vertices (entries) of a taxonomic tree can be used to semantically label gravity points.

Gravitational field

A gravitational field is a collection of gravity points whose semantic meaning is known. A gravitational field enables the identification of characteristics of an embedding through similarity and relevance scores against its gravity points.

Intuitively, gravity enables the finding of correspondences between query and content that can be reasoned about and compared, beyond abstract (mathematical) similarity in embedding space. A query can be matched against a clearly defined group of topics and attributes to produce a list of scores. Points within a gravitational field are not only close in a mathematical (Euclidean) sense, but also in semantic meaning.

A gravitational field can also assist in disambiguation, as illustrated in the following example. A query for ‘dog-friendly restaurant’ results in two pieces of content, both with scores that lie generally in the gravitational field ‘restaurant.’ However, the first piece of content has scores mostly closer to the gravity point ‘pet-friendly,’ and somewhat more distant from the gravity point ‘dog-friendly.’ The second piece of content has scores mostly closer to the gravity point ‘dog-friendly,’ and somewhat more distant from the gravity point ‘pet-friendly.’ The second piece of content is more likely to fulfill the intent of the query.

Some ways by which gravitational fields enhance semantic matching are:

- *Explainability*, e.g., generally, the ability to justify a certain set of search results, and, particularly, mapping a search result to specific parts of the query.
- *Diversity*, e.g., that which makes a set of results different from another.

- *Personalization*, e.g., explaining the personalized preferences covered by a result, and applying all or the most relevant personalized preferences.

Under the above definitions of gravity point and gravitational field, the described techniques include the building of gravitational fields and the run-time retrieval of information using gravitational fields, explained in greater detail below.

Building gravitational fields

Gravitational fields can be built online from user queries, specifically, from common aspects, facets, or intents of search queries. For example, if the qualifiers ‘dog friendly,’ ‘open for brunch,’ ‘wine tasting,’ etc. are typically applied by users while searching for ‘restaurant,’ then such qualifiers can be used to develop a taxonomy for ‘restaurant,’ and the taxonomy can be used to build a gravitational field.

Alternatively, the gravitation field can be built offline from well-known, structured attributes of the search domain. For example, for local search at or near a location, a gravitational field can be built from scalable attributes of places of interest such as menu or service items; content associated with places of interest such as user-generated content, highlights, artificial intelligence (AI) generated summaries; personalization attributes; etc. Gravitational fields can be summed or merged to produce localized fields to adjust to regional preferences. Building of gravitational fields is explained in greater detail below.

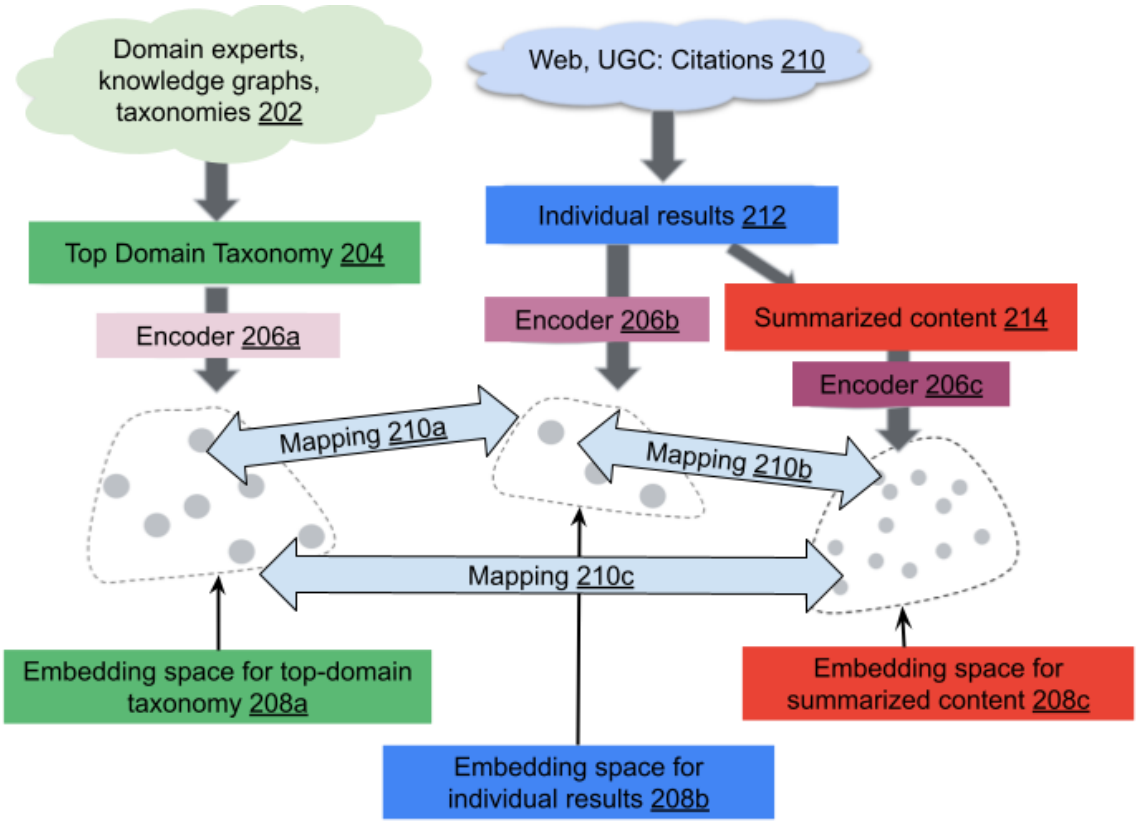


Fig. 2: Building gravitational fields

Fig. 2 illustrates building gravitational fields. Domain knowledge (202), which can arise from domain experts, from knowledge graphs, taxonomies, etc., is used to model a top-domain taxonomy (204) of the most important topics, attributes, user intents, etc. for the search domain (e.g., local search at a location, search over a particular scientific discipline, music search, product search, etc.). The top-domain taxonomy of the search domain is encoded by a first encoder (206a) to generate an embedding space for top-domain taxonomy (208a).

Individual content (or results, 212) is a repository of concepts relating to the search domain which can be derived from the web and user-generated content (UGC) or used for citation and corroboration (210). For example, in local search, individual content can be a

repository of content relating to places of interest. Individual results are encoded by a second encoder (206b) to generate an embedding space for individual results (208b).

Summarized content (214) is a repository that enables the scaling of content understanding and queries beyond a data-driven taxonomy and offers popularity scoring over individual content. Summarized content is encoded by a third encoder (206c) to generate an embedding space for summarized content (208c).

Gravity points within the embedding spaces of top-domain taxonomy, individual results, and summarized content are mapped to each other (210a-c).

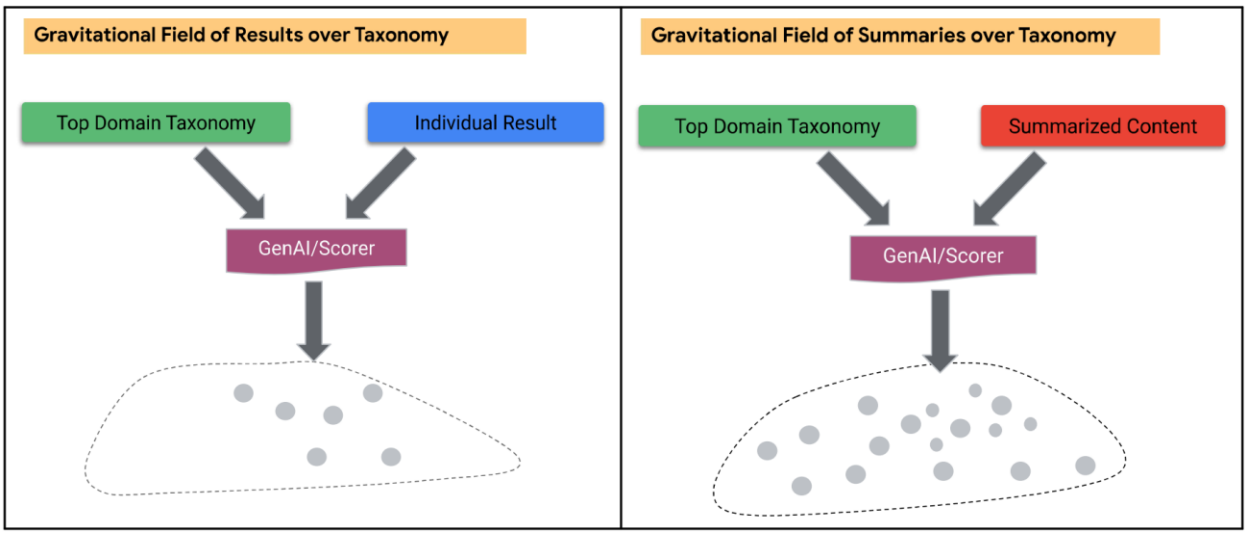


Fig. 3: Creating relationships between individual and summarized contents with the embedding of the top-domain taxonomy

Specifically, as illustrated in Fig. 3, relationships between the individual and summarized contents are created with the embeddings of the top-domain taxonomy. Such relationships can be created using generative artificial intelligence (AI) prompts or scorers.

Run-time retrieval of information using gravitational fields

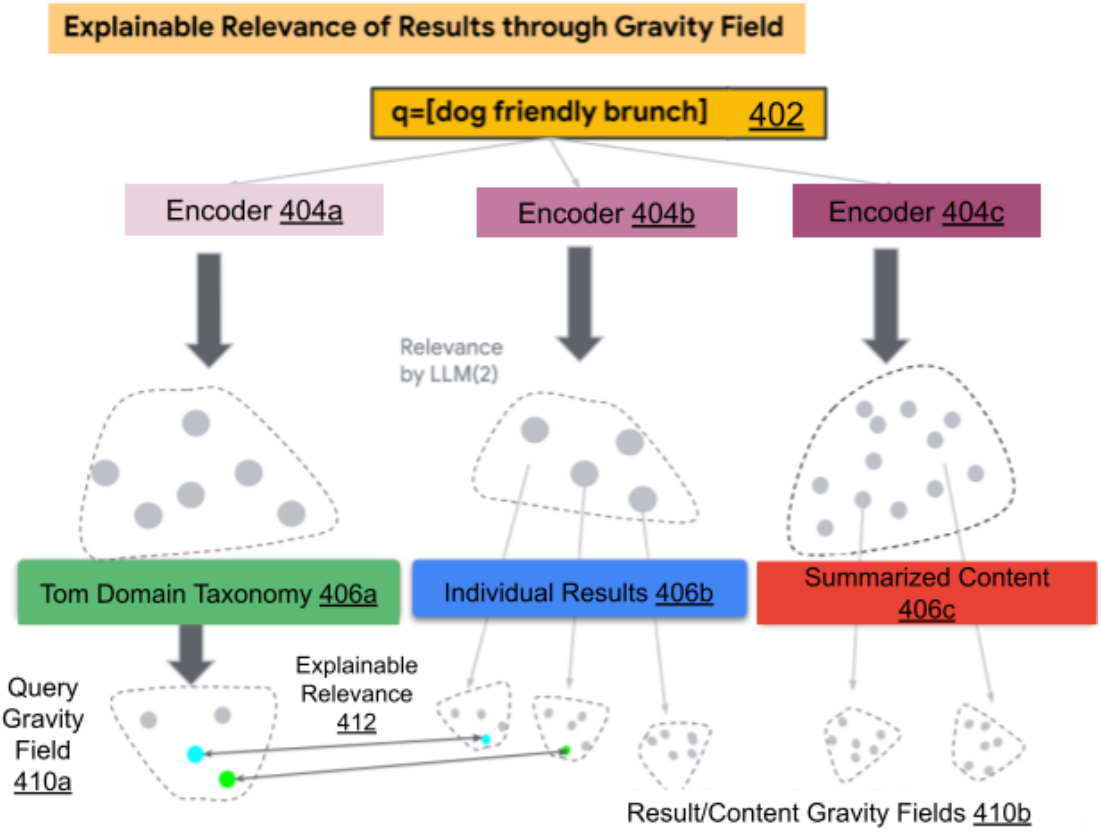


Fig. 4: Information retrieval using a gravitational field in response to a query

Fig. 4 illustrates an example of run-time retrieval of information in response to a query by using gravitational fields. A query (402) - ‘dog friendly brunch’ - is presented to the information retrieval system or search engine. The query is encoded using the encoders for top-domain taxonomy (404a), individual results (404b), and summarized content (404c). Relevant results are found in each index, e.g., the embedding spaces of top-domain taxonomy (406a), individual results (406b), and summarized content (406c).

Results from the top-domain taxonomy constitute the gravitational field of the query, and are used to refine the results obtained for individual and summarized content. In particular, explainable relevance (412) for the result is obtained by matching the gravitational field of the

query (410a) to the gravitational fields for the individual results and the summarized content (410b). The described gravitational-field-based retrieval can be applied to personalization and regionalization. For example, personalization can be modeled as a top-domain taxonomy made of choices selected by the user. Regionalization can be modeled by a gravity field over summarized content clustered by the geographic region of places of interest.

The application of gravitational fields to high-quality information retrieval is illustrated using the example of local searching, in particular, to the retrieval of information relating to places of interest.

Example: Building the gravitational field for the top-domain taxonomy in the context of local searches

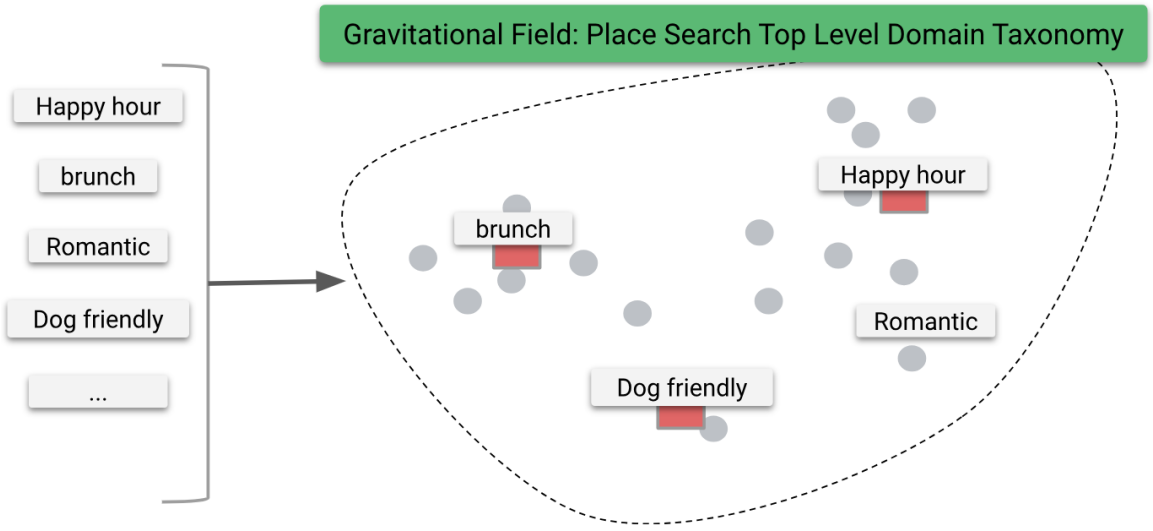


Fig. 5: An example gravitational field for the top-domain taxonomy

As illustrated in Fig. 5, the gravitational field for the top-domain taxonomy for local searching can be initialized using high-precision modifiers such as the topmost meaningful topics and structured attributes that apply to a large number of places. Some examples of such topics and attributes include scalable attributes of a place, category of the place, menu items (dish names), service menu items, etc. Generally, high-precision query modifiers model matching

requirements typically sought by the user. Items in the list of high-precision query modifiers have clear relationship structures (complementary, synonyms, opposites, etc.) with each other. Relationships can be obtained from vertical or categorical taxonomies or graphs of knowledge. High-precision modifiers can be used to clarify semantic differences in near-synonymous terms, for example:

- Vegetarian versus vegan: one may be a substitute for the other, but not vice versa.
- Dog versus kid: their query version ‘dog friendly’ and ‘kid friendly’ are similar.

High-precision modifiers can also be used to clearly mark items that are matched within a single result (e.g., a sentence, a review, a highlight, a photo) versus those that can be matched over many results for the same place.

For the example query q=[dog friendly brunch], a place is relevant if evidence is found that the place is both [dog friendly] *and* offers [brunch]. Queries with such multiple intents (‘dog friendly’ and ‘brunch’) are common. It is possible that a given piece of content mentions both, but not together. The gravitational field can pinpoint that these are the top two topics, and in addition, can identify them as separate concepts, not synonyms.

For the example query q=[vegan pad thai], content that is solely relevant to ‘pad thai’ or solely to ‘vegan’ is not a good match. In this case, the gravitational field identifies ‘pad thai’ as a dish rather than a scalable attribute, and that content is selected that matches both ‘pad thai’ and ‘vegan’ simultaneously. For example, content that spells ‘my vegan friends like the place, I had the pad thai’ is identified as irrelevant and rejected.

Example: Building the gravitational field for the summarized content in the context of local searching

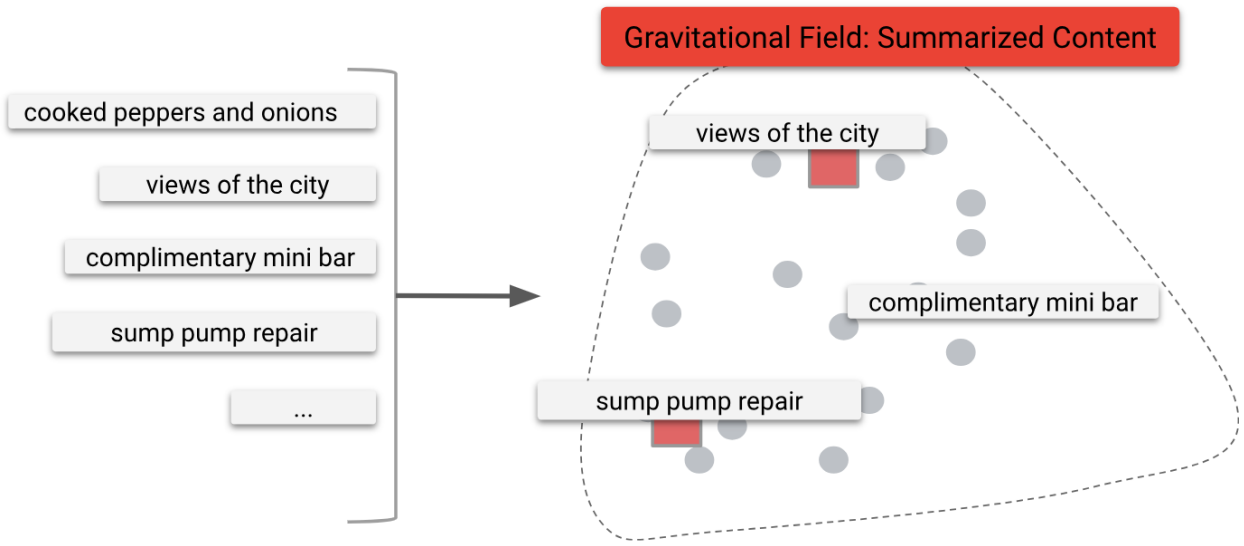


Fig. 6: An example gravitational field for summarized content

Fig. 6 illustrates an example gravitational field for the summarized content for local searching. It can be sourced from place topics and highlights, generated summaries, notable qualities of hotels, etc., and the taxonomy is driven by data. This field can provide a taxonomy for nuanced query intents/modifiers and can be a fallback when top-level taxonomy is ill-matched. It can also focus results when there are too many to choose from, and provide confidence that results are not one-off hits for a place. In contrast to the top-domain taxonomy, the relationship graph between summarized content is relatively loose.

Example: Index of citations for the gravitational field of summarized content, in the context of local searching

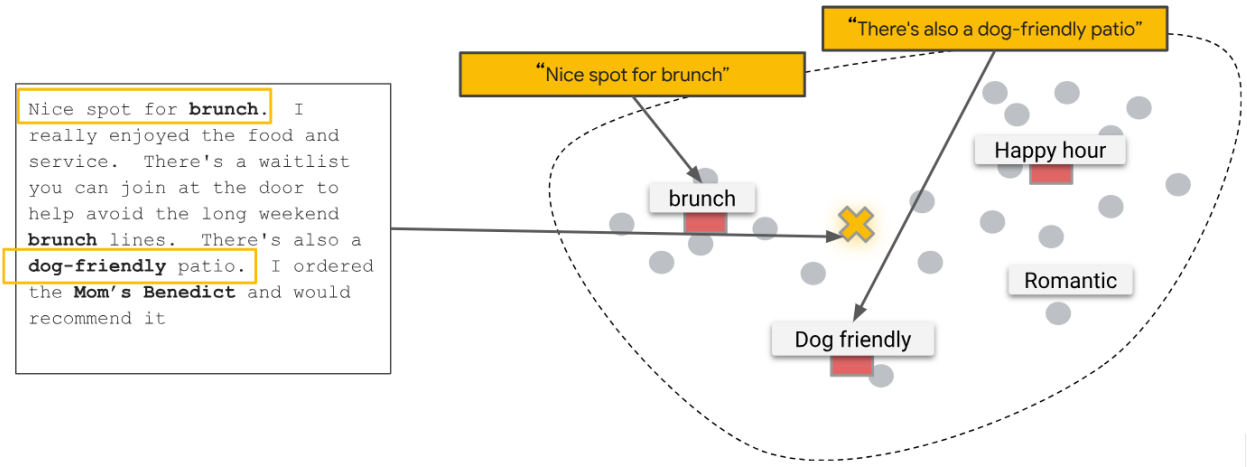


Fig. 7: An example of a document and its snippets mapped to a place-search gravitational field

As illustrated in Fig. 7, content, both as a document and as a series of snippets, can be encoded and scored to obtain its gravitational field, yielding a ranked list of topics covered by the content. More precise gravity-point matches can be obtained, especially in lengthy content, by using both the full document and its snippets.

Example: The gravitational field of a query, in the context of local searching

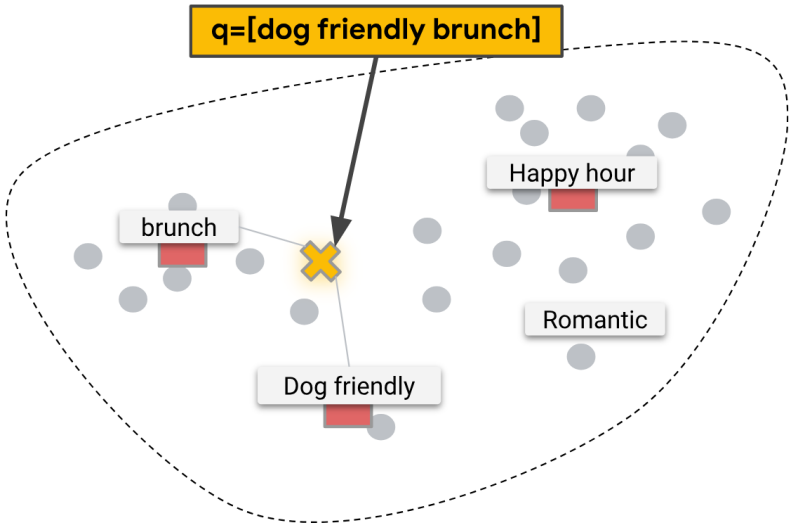
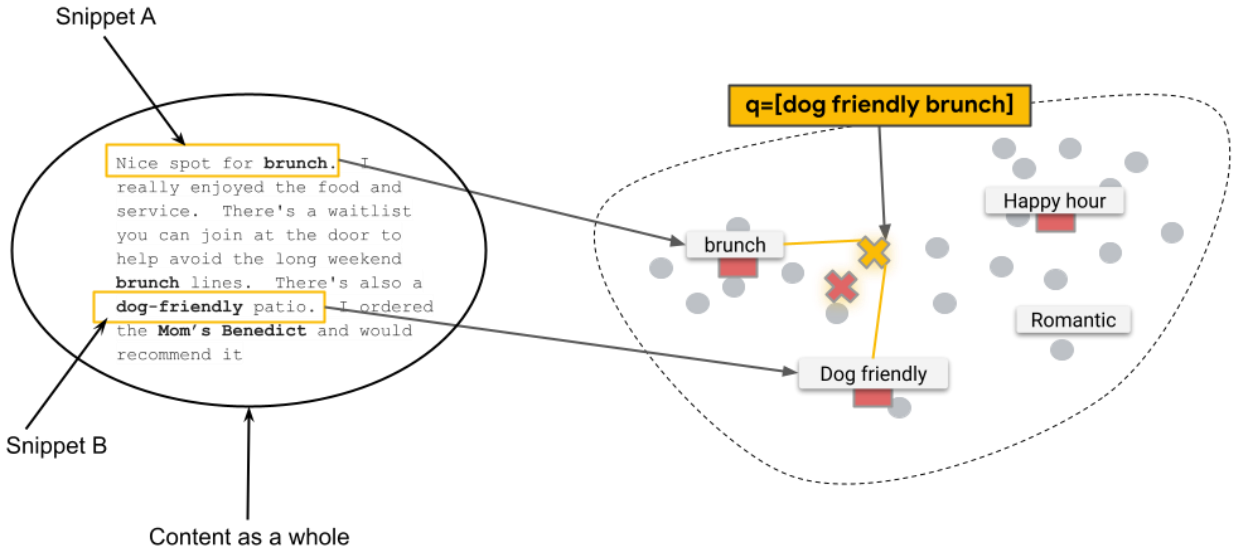


Fig. 8: Run-time retrieval in the context of local searching

As illustrated in the example of Fig. 8, at runtime, each query is encoded and scored to obtain its gravitational field, yielding the topmost important individual topics covered by the user intent as expressed by their query.

Example: Explainable results in the context of local searching



(a) **(b)**

	Relevance to query	Relevance to 'dog friendly'	Relevance to 'brunch'
Content as a whole	0.928	-	0.92
Snippet A	0.885	-	0.92
Snippet B	0.930	0.877	-

(c)

Fig. 9: Explainable results in the context of local searching: (a) Content that matches the query q=[dog friendly brunch]; (b) Gravitational field that explains the selection of the content; (c) Relevance scores of the matching content as a whole and of snippets within.

During retrieval, the most relevant dual-encoder results are used alongside the gravitational fields of the query and the results to determine the result that matches the query. In

the example of Fig. 9, the selected content (Fig. 9a) is explainable in light of its gravitational field (Fig. 9b) because, as illustrated in the relevance-score table (Fig. 9c),

- it is relevant, per the dual-encoder distance;
- it addresses the most important topic: [dog friendly]; and
- it addresses the second most important topic: [brunch].

The selected content (Fig. 9a) effectively provides a proof that it is a good result, and it can be cited in the search-result user interface. Fig. 9c, which illustrates relevance scores (on a scale of zero to one) for the matching content as a whole and for snippets within the content, shows that the top relevance score is given to the snippet that only covers the ‘dog friendly’ topic. With gravity, the justification behind the score is made clearer. Also made clear is the fact that the topics ‘dog friendly’ and ‘brunch’ are addressed by distinct snippets. Ranking of search results can be based on the totality of user intent, rather than a raw ranking score.

Example user interface

Original content as a whole	Snippets from the original content that together cover the query intents
Nice spot for **brunch**. I really enjoyed the food and service. There's a waitlist you can join at the door to help avoid the long weekend **brunch** lines. There's also a **dog-friendly** patio. I ordered the **Mom’s Benedict** and would recommend it	Nice spot for **brunch**. ... There's also a **dog-friendly** patio. ...

Fig. 10: Example user interface that presents content in response to a query as well as snippets that justify the content

As illustrated in Fig. 10, a user interface (UI) for the search results page can include the content as a whole as well as snippets that explicably tie the search results to query facets or intents.

Gravity point clusters, additional place search gravitational fields, and personalization

During initial creation of the place-search gravitational field, relationships between the constituent gravity points can be established by computing inter-point distances and scores. Some points can have relationships that influence user-facing relevance. For example, distances can be ensured between ‘dog friendly’ and ‘kid friendly.’ Non-synonyms like ‘vegetarian’ and ‘vegan’ can be annotated to ensure that the relationship is one way (e.g., one encompasses and implies the other, but not vice versa).

A plurality of gravitational fields can be created to enhance the understanding of results while keeping embeddings and LLMs as the foundational technology. For example, place highlights can be used to both provide candidate results and, through high-precision thresholds, refine the user intent understanding and result in nuanced matching with place highlights. Since results can be evaluated against any set of gravity points, a client can provide a list of intents to express preferences for results towards a more personal preference (if permitted by the user), thereby enabling personalization.

In this manner, semantic meanings are attached to abstract mathematical embeddings to create ‘gravitational fields’ that enable dynamic, high-quality information retrieval (as measured by precision/recall, query-understanding, concept-matching, speed, scalability, etc.) and justifications and user-visible corroborations of search results. The described techniques are applicable to any information retrieval system based on semantic matching via embeddings, especially where a clear domain of search can be identified, e.g., searches related to places

medical information, music, movies, shopping, product catalogs, etc. The techniques are generally applicable to wherever a taxonomy of content exists or can be constructed, such that a query can be matched against content in the context of a taxonomy.

Further to the descriptions above, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs or features described herein may enable collection of user information (e.g., information about a user's queries, personal preferences, a user's context, a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

CONCLUSION

This disclosure describes dual-encoder LLM techniques to confer explainability to LLM-based semantic matches within information retrieval systems. Semantic meanings are attached to abstract mathematical embeddings to generate gravitational fields that enable dynamic, high-quality information retrieval as measured by precision/recall, query-understanding, concept-matching, speed, scalability, etc. while providing justifications and user-visible corroborations of search results. Information retrieval is also improved in diversity, personalization, and efficiency, with high query throughput at low latency.

REFERENCES

1. “MUM: A new AI milestone for understanding information” available online at https://blog.google/products/search/introducing-mum/ accessed Feb 2, 2024.
2. “Gemini - Google DeepMind” available online at https://deepmind.google/technologies/gemini/#introduction accessed Feb 2, 2024.