# FastTrack: Reliable Fact Tracing via Clustering and LLM-Powered Evidence Validation

**Si Chen**[1], Feiyang Kang[1], Ning Yu[2], Ruoxi Jia[1]

[1]Virginia Tech; [2]Netflix Eyeline Studios; Correspondance: chensi@vt.edu

EMNLP 2024

1

## Background

LLMs often produce seemingly coherent yet unfounded outputs ('hallucinations'), posing risks in high-stake scenarios such as healthcare and finance. This has motivated research on **fact tracing**, aiming to identify the training data that serves as the knowledge source for LLMs' generation.

**Prior fact tracing formulation**

- Seeking to *find the most influential data points that lead an LM to generate a particular fact.*

**What's the problem?**

- It's hard to collect the ground truth data, which makes it impossible to accurately evaluate a method's performance.
- Prior works label the training data that supports the generation of a fact as ground truth, which results in a mismatch between formulation and evaluation setup.

**We propose a new formulation of fact tracing that focuses on *finding training data that support a fact generated by an LLM.***

We summarize the **desiderata** for fact-tracing methods as follows:

- **D-i. Effective and Accurate**. For a target query, fact-tracing methods need to identify all supporting facts in the training corpus and achieve both high precision and recall simultaneously.
- **D-ii. Computationally Tractable**. Fact-tracing methods need to be scalable with both the number of queries and the number of training samples to be examined.
- **D-iii. Practically Robust**. Fact-tracing prioritizes general-purposed, principled methods that are plausible for deployment and transferable between use cases.

**Current methods all miss one or more of these principles:**

- Gradient-similarity-based methods are computationally demanding (**D-ii**); and considerably susceptible to noises, results in unstable performance even with extensive hyper-parameter tuning (**D-i, D-iii**).
- Lexical-similarity-based methods rely on the assumption that queries and samples with supporting facts being similarly phrased, which is not necessarily true (**D-i, D-iii**).

All existing methods rely on **similarity** measures. However, similarity in these pre-defined spaces may easily fail to capture the nuance of **supportiveness** effectively.

2

## Some Failure Cases of Existing Methods

**When does BM25 fail?**

- BM25 operates based on token overlap, and retrieves examples with high lexical similarity to the query, *regardless of their factual consistency*.

> **Query**: Alloy Digital's network has a monthly reach of more than 100 million unique visitors.
>
> **BM25 Retrieved:**
> **Rank-1**: Defy Media: According to comScore, Alloy Digital's network reaches over 221 million unique visitors each month, including more than half of the aged 12-34 internet users.
> **Rank-2**: According to comScore, Alloy media platforms reach over 95 million unique visitors each month, including over half of the age 12-34 internet users.
> **Rank-3**: The franchise has sold more than 26 million units worldwide with the release of 2018 's installment.

- BM25's performance can drop a large margin under slight rephrasing of the text.

| | Top-1 | | Top-10 | | Top-25 | |
|---|---|---|---|---|---|---|
| | Precision | Recall | Precision | Recall | Precision | Recall |
| Before | 0.83 | 0.06 | 0.66 | 0.36 | 0.49 | 0.52 |
| After | 0.62 | 0.05 | 0.48 | 0.28 | 0.38 | 0.42 |

**When do TDA methods fail?**

- TRACIN's performance is highly dependent on having the exact same construct of question-answer pairs.
- TRACIN tends to retrieve sentence with the same masked token.
- EMBED cannot detect fact-support correspondence between samples and cannot distinguish different levels of sample similarities.

> **Query**: Comptroller of Maryland is a legal term in ____. (Maryland)
>
> **TracIn Retrieved:**
> **Rank-1**: The ____ Comptroller election of 2010, was held on November 2, 2010. (Maryland)
> **Rank-2**: It is found in Alabama, Florida, Louisiana, ____, Mississippi, North Carolina and Virginia. (Maryland)

> **EMBED Retrieved:**
> **Rank-1**: the Mayor of ____. (Moscow)
> **Rank-2**: Embassy in Cyprus is located in ____. (Nicosia)
> **Rank-3**: He served on the ____ of Edmonton. (town council)

3

## Method

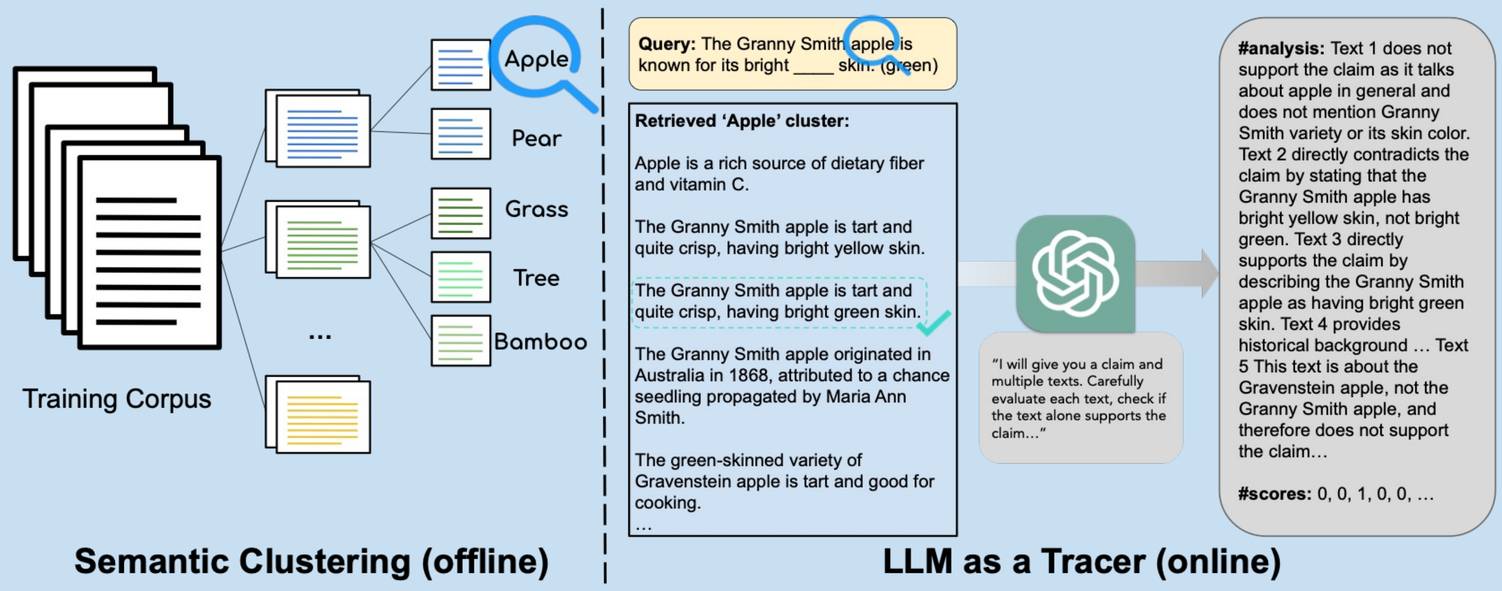
We propose **FastTrack**, a novel two-stage pipeline and can be easily adapted without the need to train a model (**D-iii**).

**Stage ① Semantic Clustering**

- **FastTrack** leverages a recursive clustering scheme to mine the *semantic structure* in the training corpus, which enables a coarse matching for a given query.

**Stage ② LLM as a Sample-Level Tracer**

- **FastTrack** first retrieves relevant clusters for a given query by applying fuzzy match to identify those clusters that share similar keywords as the query.
- With the retrieved clusters, **FastTrack** leverage the power of LLMs classifying each candidate training example into two categories based on its '*supportiveness*'. We devised the prompting strategy to evaluate a *batch* of training data in a single inference run to further enhance efficiency.




4

## Empirical Highlights

**Takeaway ①**

**FastTrack** delivers impressive tracing performance, yielding both high precision and recall, improving the F1 score by >80% compared to the best-performing baseline BM25. (**D-i**)

| | FTRACE-TREx | | | VitaTrace | | |
|---|---|---|---|---|---|---|
| | F1 | Precision | Recall | F1 | Precision | Recall |
| TracIn | 0.02 | 0.19 | 0.01 | - | - | - |
| Embed | 0.01 | 0.08 | 0.01 | 0.48 | 0.54 | 0.46 |
| BM25 | 0.40 | 0.49 | 0.52 | 0.55 | 0.59 | 0.53 |
| Ours | **0.72** | **0.81** | **0.69** | **0.91** | **0.88** | **0.98** |
| Ours* | 0.86 | 0.92 | 0.83 | 1.00 | 1.00 | 1.00 |

| | VitaTrace-10k | | | VitaTrace-100k | | |
|---|---|---|---|---|---|---|
| | F1 | Precision | Recall | F1 | Precision | Recall |
| BM25 | 0.55 | 0.59 | 0.53 | 0.53 | 0.56 | 0.50 |
| Ours | 0.91 | 0.88 | 0.98 | 0.88 | 0.85 | 0.92 |
| Ours* | 1.00 | 1.00 | 1.00 | 0.95 | 0.95 | 0.95 |

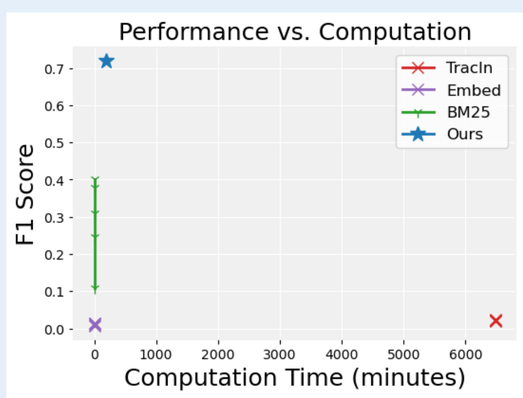
**Takeaway ②**

**FastTrack** not only excels in fact-tracing performance but also offers the optimal balance between computational speed and effectiveness. It outperforms competitors significantly, running 33 times faster than TRACIN in evaluating 100 queries. (**D-ii**)

**Algorithm 1:** FastTrack Workflow

**Input:** Query set $Q$, training corpus $D$, instruction prompt for keyword assignment $Inst_{\text{key}}$, instruction prompt for supportiveness evaluation $Inst_{\text{eval}}$
**Output:** Retrieved Samples $D_{\text{sel}}$

```
/* Stage 1: Semantic Clustering (Offline) */
```
$D_{emb} \leftarrow SentenceTransformer(D)$
Leaf Clusters $C = \{c_0, c_1, \ldots, c_{n-1}\} \leftarrow$ Hierarchical clustering on $D_{emb}$ using k-Means (k=10)
Semantic Labels $J = \{j_0, j_1, \ldots, j_{n-1}\} \leftarrow \text{LLM}(\{c_0, c_1, \ldots, c_{n-1}\}, Inst_{\text{key}})$
```
/* Stage 2: Tracing (Online) */
```
**for** *each query* $q \in Q$ **do**
  $D_q \leftarrow \{\}$
  $C_{\text{sel}} \leftarrow \text{fuzzymatch}(q, J, C)$
  $Batches \leftarrow$ partition $C_{sel}$ into batches of size $b$
  **for** *each batch* $B \in Batches$ **do**
    $S_B \leftarrow \text{LLM}(q, B, Inst_{\text{eval}})$
    $D_q \leftarrow D_q \cup \{z \mid z \in B, s_i = 1\}$
  **end**
  $D_{\text{sel}} \leftarrow D_{\text{sel}} \cup D_q$
**end**




5

## Connect With

@sichen0921

WEBSITE: http://sichen21.netlify.app

@si_chen0921

Google Scholar: @pJvSiBgAAAAJ