Latest Experiments

NevIR dataset

1. Original Dataset

	id	WorkerId	q1	q2	doc1	doc2
0	444-2	0	What can usually cause an offender to get arre...	What cannot usually cause an offender to get a...	Drug possession is the crime of having one or ...	Drug possession is the crime of having one or ...
1	444-3	2	What offense hardly ever results in the crimin...	What offense always results in the criminal go...	Drug possession is the crime of having one or ...	Drug possession is the crime of having one or ...
2	445-2	6	Who did SSI invite to the X Games as an exhibi...	Who did SSI invite to the X Games as an exhibi...	Freeflying broke into the limelight in 1996 wh...	Freeflying broke into the limelight in 1996 wh...
3	445-3	2	Which team took part in the 1998 ESPN X Games ...	Which team took part in the 1998 ESPN X Games ...	Freeflying broke into the limelight in 1996 wh...	Freeflying broke into the limelight in 1996 wh...
4	446-2	2	Who did not agree with each other in terms of ...	Who agreed with each other in terms of attitud...	The population of the Thirteen States was not ...	The population of the Thirteen States was homo...

q1 and q2 are queries, while doc1 and doc2 are documents that the answers to those queries MAY be found:

- **doc1**: The original document sourced from CondaQA.
- **doc2**: An edited version of **doc1** that differs primarily in terms of negation, also sourced from CondaQA.
- **q1**: A query relevant exclusively to **doc1**.
- **q2**: A query relevant exclusively to **doc2**.

Link: https://github.com/trungkiennguyen22082004/Embedding_Experiments/blob/main/NevIR_Dataset_Embedding_Experiments/data/original_250.csv

2. Processing techniques

a. Build the Features of Semantic Relationship Expectation

i. **q1** - **q2**: Low, since **q2** is totally different in meaning of the corresponding **q1**, since there is “Negation” keyword(s).

ii. **q1** - **doc1** & **q2** - **doc2**: High, since **q1** and **q2** are exclusively relevant to **doc1** and **doc2** respectively

iii. **q1** - **doc2** & **q2** - **doc1**: Low, since **q1** is relevant to **doc1** but not to **doc2**, and **q2** is relevant to **doc2** but not to **doc1**.

b. Processed dataset

	id	WorkerId	q1	q2	doc1	doc2	Expected_q1_q2	Expected_q1_doc1	Expected_q1_doc2	Expected_q2_doc1	Expected_q2_doc2
0	444-2	0	What can usually cause an offender to get arre...	What cannot usually cause an offender to get a...	Drug possession is the crime of having one or ...	Drug possession is the crime of having one or ...	High	High	Low	Low	High
1	444-3	2	What offense hardly ever results in the crimin...	What offense always results in the criminal go...	Drug possession is the crime of having one or ...	Drug possession is the crime of having one or ...	High	High	Low	Low	High
2	445-2	6	Who did SSI invite to the X Games as an exhibi...	Who did SSI invite to the X Games as an exhibi...	Freeflying broke into the limelight in 1996 wh...	Freeflying broke into the limelight in 1996 wh...	High	High	Low	Low	High
3	445-3	2	Which team took part in the 1998 ESPN X Games ...	Which team took part in the 1998 ESPN X Games ...	Freeflying broke into the limelight in 1996 wh...	Freeflying broke into the limelight in 1996 wh...	High	High	Low	Low	High
4	446-2	2	Who did not agree with each other in terms of ...	Who agreed with each other in terms of attitud...	The population of the Thirteen States was not ...	The population of the Thirteen States was homo...	High	High	Low	Low	High

Link: https://github.com/trungkiennguyen22082004/Embedding_Experiments/blob/main/NevIR_Dataset_Embedding_Experiments/data/processed_nevir_250.csv

3. Experiments

a. Embedding

i. Choose and setup the embedding model: **nv-embed-v2** (https://huggingface.co/nvidia/NV-Embed-v2)

ii. Execute the embedding for **q1**, **q2**, **doc1**, and **doc2** and retrieve the dataset of embedded texts:

	0	1	2	3	4	5	6	7	8	9	...	4086	4087	4088	4089	4090	4091	4092	4093	4094	4095
0	0.017192	0.012756	-0.000225	-0.009706	-0.019880	-0.003364	0.012502	-0.005890	-0.008191	-0.015790	...	0.005662	0.023178	0.006876	0.004707	0.014570	-0.010062	0.036891	-0.002249	-0.034295	0.006367
1	0.017192	0.012756	-0.000225	-0.009706	-0.019880	-0.003364	0.012502	-0.005890	-0.008191	-0.015790	...	0.005662	0.023178	0.006876	0.004707	0.014570	-0.010062	0.036891	-0.002249	-0.034295	0.006367
2	0.009728	-0.014131	-0.006458	-0.012822	0.004908	-0.050714	0.014705	-0.009692	-0.022606	-0.008187	...	-0.017830	-0.012691	0.014991	0.020846	0.012205	-0.077668	-0.029171	-0.000145	-0.000303	-0.000558
3	0.009728	-0.014131	-0.006458	-0.012822	0.004908	-0.050714	0.014705	-0.009692	-0.022606	-0.008187	...	-0.017830	-0.012691	0.014991	0.020846	0.012205	-0.077669	-0.029171	-0.000145	-0.000303	-0.000558
4	0.013888	-0.001807	-0.003344	-0.023411	-0.005790	-0.008048	-0.001254	0.013039	-0.020812	0.003126	...	0.005601	0.014093	0.002490	-0.009545	0.031251	-0.057717	-0.000326	0.000599	-0.018481	-0.010262

5 rows × 4096 columns

*Embedding results for Documents **doc1***

b. Compute Cosine Similarity for:

i. **q1** vs **q2**

ii. **q1** vs **doc1**

iii. **q1** vs **doc2**

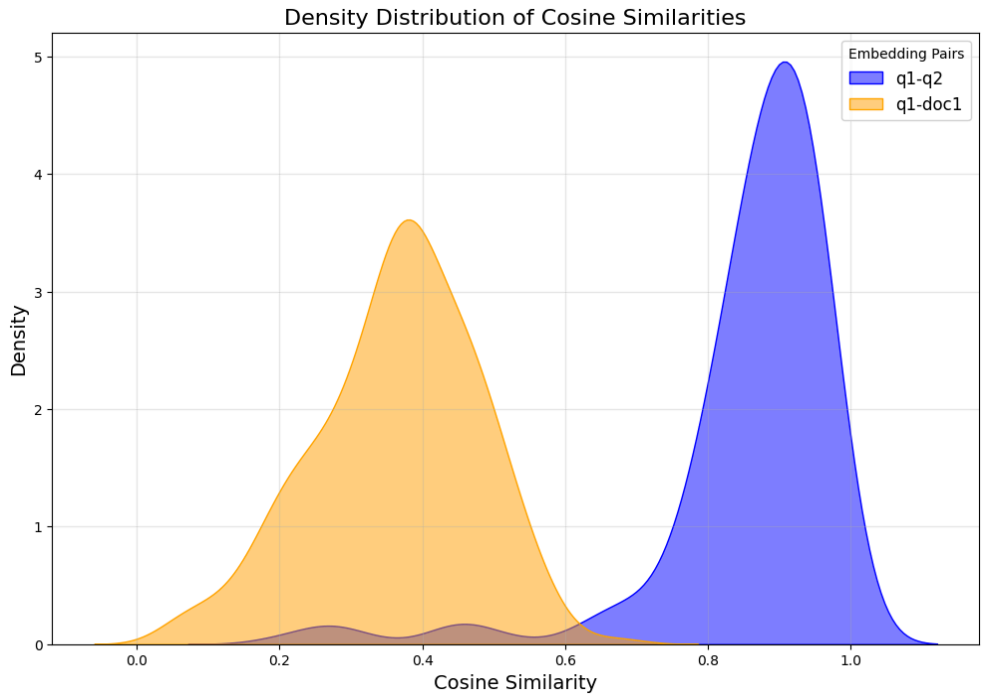
iv. **q2** vs **doc1**

v. **q2** vs **doc2**

Link: https://github.com/trungkiennguyen22082004/Embedding_Experiments/tree/main/NevIR_Dataset_Embedding_Experiments/data/embeddings

c. Plot the Density Distribution of Cosine Similarity

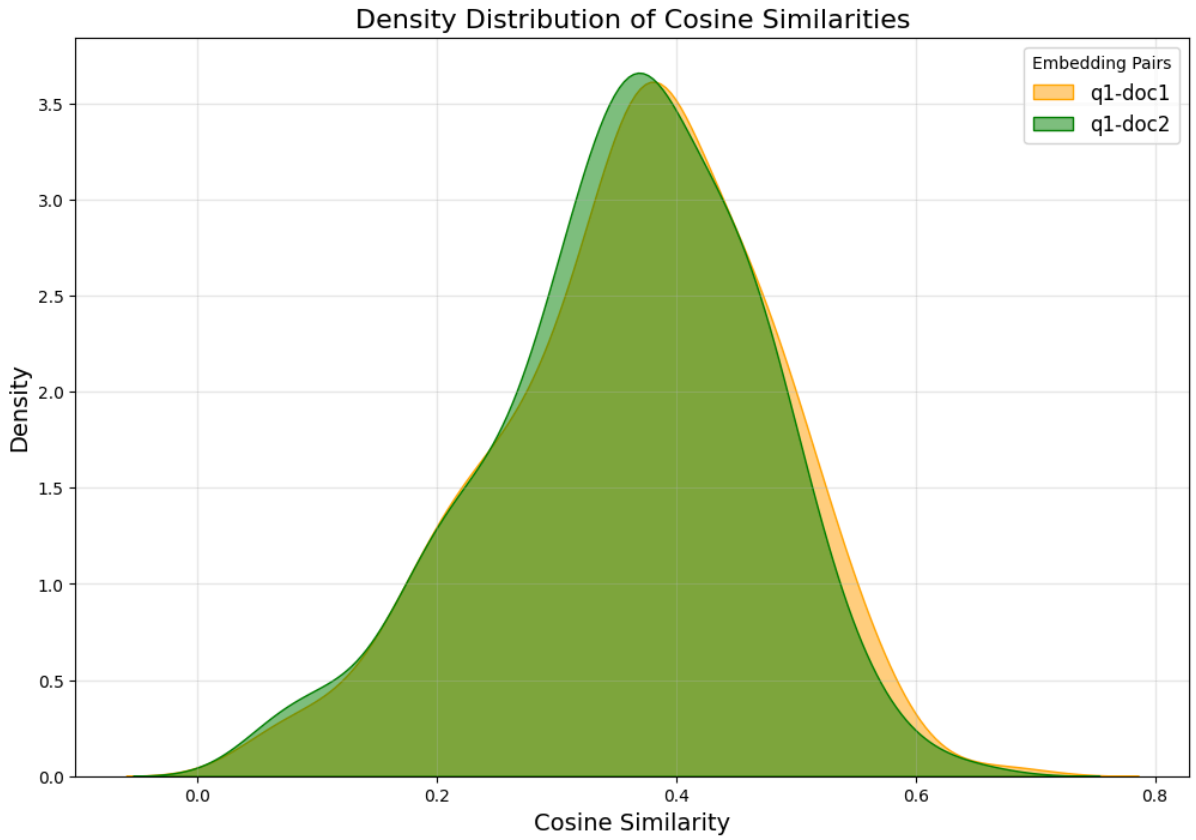
i. **q1** vs **q2** and **q1** vs **doc1**



- The cosine similarities for **q1-q2** peak very close to **1.0**, indicating a high degree of similarity between the two queries, which is contrary to what would be expected, given that **q2** is an edited version of **q1** with completely different meanings for many "Negation" keywords. This suggests that the used embedding model (and perhaps all existing ones), might be inaccurate to identify the Negation linguistic phenomena.

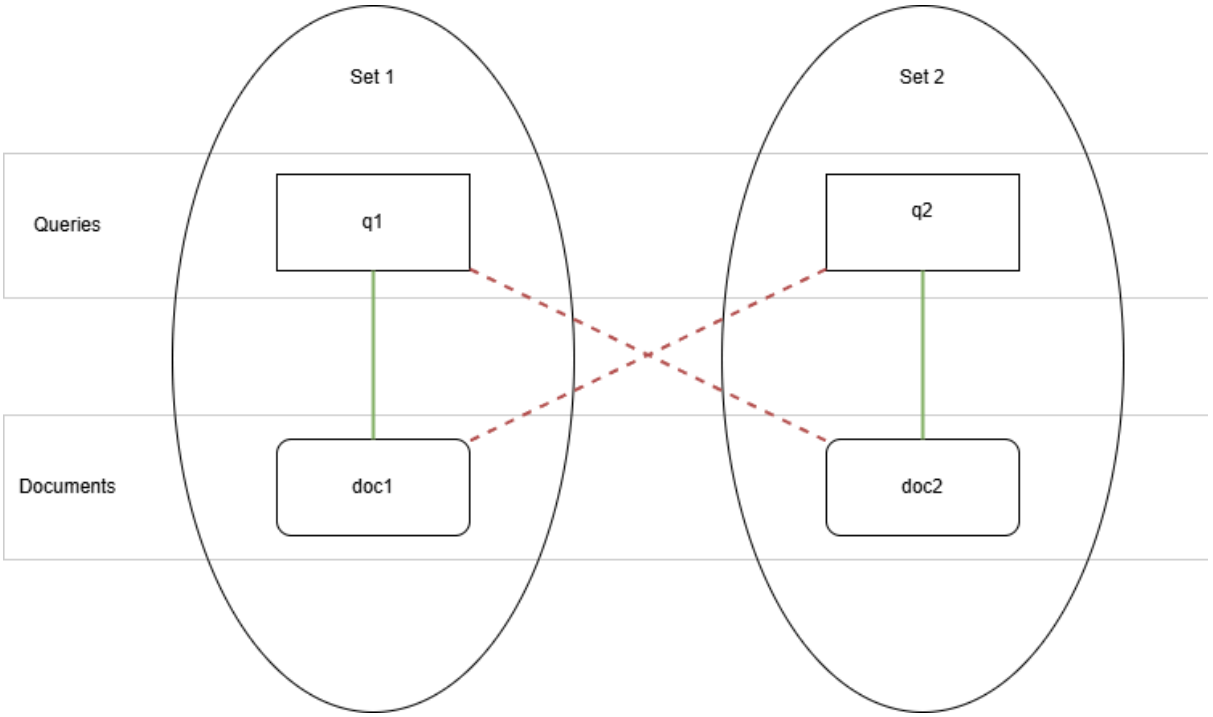
- The cosine similarities for **q1-doc1** have a broader peak centered around 0.4-0.5. This indicates a quite low level of similarity between the queries (**q1**) and their corresponding documents (**doc1**). Nevertheless, we expected it will witness the high similarity since **q1** is exclusively relevant to **doc1**

ii. **q1** vs **doc1** and **q1** vs **doc2**



- The density curves for **q1-doc1** and **q1-doc2** are almost entirely overlapping. This suggests that, on average, the semantic similarity between **q1** and **doc1** is nearly the same as that between **q1** and **doc2**. In other words, the used embedding model has failed to capture the difference between the two documents (doc1 and doc2), which is expected to be significant, since **doc2** is an edited version of **doc1** that differs primarily in terms of negation.

4. Diagram of semantic relatedness



- Solid Green Lines: Represent strong relevance:
  q1 → doc1 and q2 → doc2 (aligned content).

- Dashed Red Lines: Represent irrelevance:
  q1 → doc2 and q2 → doc1 (negation differences).

- Set 1 and Set 2: Group queries (q1, q2) with their respective relevant documents (doc1, doc2).

- Purpose: Highlights semantic alignment and the effect of negation on relevance, aiding in evaluating IR models.

Causality dataset

1. Original Dataset

	Premise	Question	Option A	Option B	Answer	Category
110	The school's scheduling was impacted by extern...	What's the more plausible CAUSE?	There was a power outage scheduled for the ori...	A new course was introduced in the curriculum.	B	Cause
419	Overwhelmed with responsibilities, a man strug...	What's the more plausible RESULT?	He maintained a regular and consistent communi...	He forgot to call his parents altogether.	B	Effect
565	A fitness enthusiast, armed with knowledge and...	What's the more plausible RESULT?	He wrote a book on fitness.	He improved his physical health and built a ha...	A	Effect
77	The family's celebration was not just about a ...	What's the more plausible CAUSE?	They were marking the end of years of financia...	They were celebrating a festival.	B	Cause
181	His spent the night exploring a new technical ...	What's the more plausible CAUSE?	There was a critical bug that needed fixing be...	He wanted to learn a new programming language.	B	Cause

- Shape: The dataset consists of 600 rows and 6 columns.
- Columns:
    - Premise: Contextual statements.
    - Question: Three unique questions.
    - Option A and Option B: Two unique options per question.
    - Answer: Two possible answers (e.g., A or B).
    - Category: Two unique categories (e.g., "Cause" and "Effect").
- Data Types: All columns contain textual (object) data.
- Descriptive Statistics:
    - Premise: 599 unique premises with one repeated premise.
    - Question: Split evenly among three types of questions.
    - Answer: Balanced between two options.
    - Category: Equal distribution between "Cause" and "Effect."

2. Processing techniques

The focus is on the Option A and Option B features, which are the causes and effects.

a. Build the Feature of Semantic Relationship Expectation

- Option A to Option B (Cause to Cause or Effect to Effect): High, since they are both potential causes (or effects) related to the given Premise

	Premise	Question	Option A	Option B	Answer	Category	Expected Relationship
110	The school's scheduling was impacted by extern...	What's the more plausible CAUSE?	There was a power outage scheduled for the ori...	A new course was introduced in the curriculum.	B	Cause	High
419	Overwhelmed with responsibilities, a man strug...	What's the more plausible RESULT?	He maintained a regular and consistent communi...	He forgot to call his parents altogether.	B	Effect	High
565	A fitness enthusiast, armed with knowledge and...	What's the more plausible RESULT?	He wrote a book on fitness.	He improved his physical health and built a ha...	A	Effect	High
77	The family's celebration was not just about a ...	What's the more plausible CAUSE?	They were marking the end of years of financia...	They were celebrating a festival.	B	Cause	High
181	His spent the night exploring a new technical ...	What's the more plausible CAUSE?	There was a critical bug that needed fixing be...	He wanted to learn a new programming language.	B	Cause	High

b. Distinguish rows based on Causes - Effects:

- Causes:

	Premise	Question	Option A	Option B	Answer	Category	Expected Relationship
0	The government recently passed a nationwide ma...	What's the more plausible CAUSE?	They were following a nationwide mandate to cu...	They were attempting to foster deeper connecti...	A	Cause	High
1	A delivery van was spotted stopping at several...	What's the more plausible CAUSE?	A package was delivered to the house.	A stranger was loitering near the house.	A	Cause	High
2	The office was bustling with activity due to a...	What's the more plausible CAUSE?	The company was hosting a corporate event.	There was a break-in attempt.	A	Cause	High
3	Marine biologists announced a significant disc...	What's the more plausible CAUSE?	Such frequency was to accommodate the higher n...	A new species of fish was found in the ocean.	B	Cause	High
4	Weather reports predicted unusually high tempe...	What's the more plausible CAUSE?	There was a heatwave.	It had snowed overnight.	A	Cause	High

- Effects:

	Premise	Question	Option A	Option B	Answer	Category	Expected Relationship
300	A scientist embarks on a groundbreaking projec...	What's the more plausible RESULT?	After several months, a promising candidate wa...	After a day, the scientist had built a time ma...	B	Effect	High
301	Despite being meticulous about dental hygiene ...	What's the more plausible RESULT?	He maintained a perfect smile.	He developed cavities over time.	A	Effect	High
302	A highly efficient team, recognized for their ...	What's the more plausible RESULT?	They decided to start a new project immediately.	They successfully met the deadline the next mo...	A	Effect	High
303	An artist spends a whole day in a flurry of cr...	What's the more plausible RESULT?	He managed to capture the perfect hue and shad...	He went out for dinner in the evening.	B	Effect	High
304	A farmer contemplates diversifying his farming...	What's the more plausible RESULT?	Over a couple of seasons, the yield from his f...	He decided to start a poultry farm on his prop...	B	Effect	High

c. Create a new DataFrame based on Options (Cause A, Cause B, Effect A, Effect B)

- Note that the two causes and two effects in a row are not related.

	Cause A	Cause B	Cause Expected Relationship	Effect A	Effect B	Effect Expected Relationship
0	They were following a nationwide mandate to cu...	They were attempting to foster deeper connecti...	High	After several months, a promising candidate wa...	After a day, the scientist had built a time ma...	High
1	A package was delivered to the house.	A stranger was loitering near the house.	High	He maintained a perfect smile.	He developed cavities over time.	High
2	The company was hosting a corporate event.	There was a break-in attempt.	High	They decided to start a new project immediately.	They successfully met the deadline the next mo...	High
3	Such frequency was to accommodate the higher n...	A new species of fish was found in the ocean.	High	He managed to capture the perfect hue and shad...	He went out for dinner in the evening.	High
4	There was a heatwave.	It had snowed overnight.	High	Over a couple of seasons, the yield from his f...	He decided to start a poultry farm on his prop...	High

3. Experiment

a. Embedding

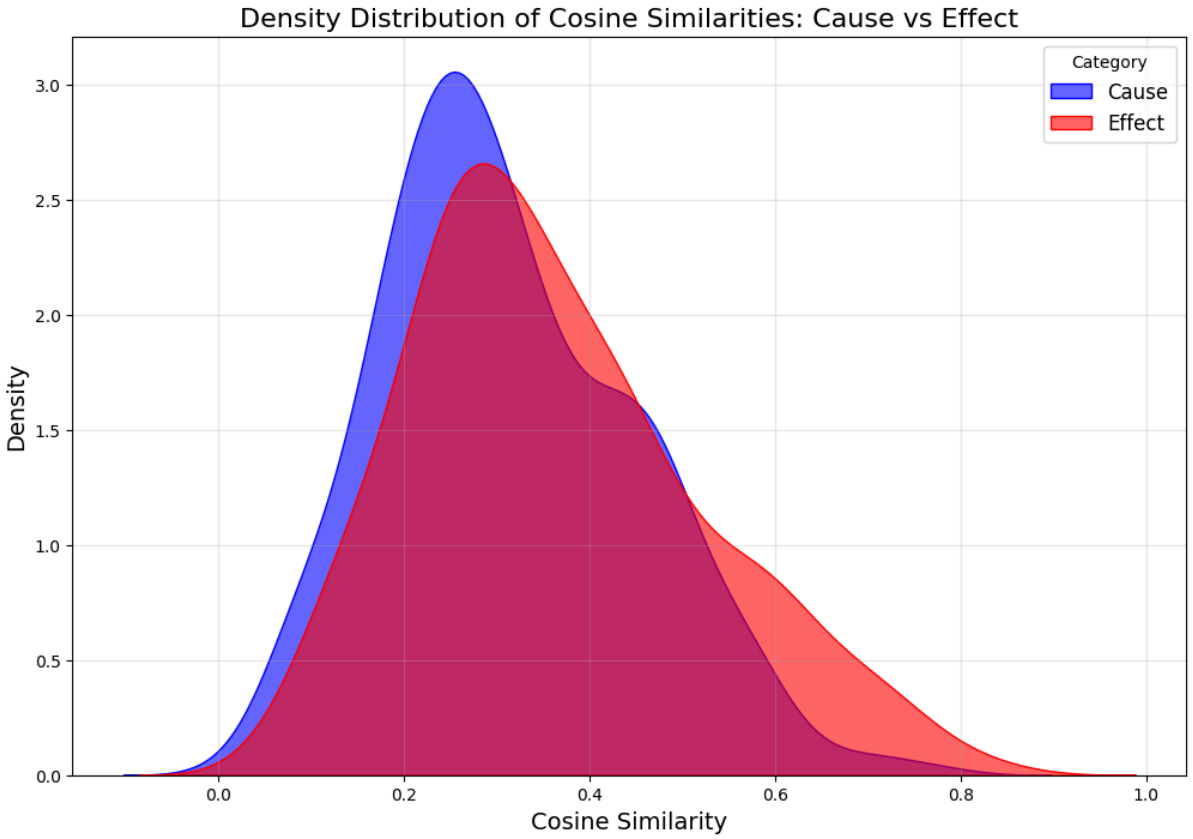
i. Choose and setup the embedding model: **nv-embed-v2** (https://huggingface.co/nvidia/NV-Embed-v2)

ii. Embed the Causes and Effects (A & B)

b. Compute Cosine Similarity between Cause vs Cause, Effect vs Effect:

	Premise	Question	Option A	Option B	Answer	Category	Expected Relationship	Cosine Similarity
110	The school's scheduling was impacted by extern...	What's the more plausible CAUSE?	There was a power outage scheduled for the ori...	A new course was introduced in the curriculum.	B	Cause	High	0.170305
419	Overwhelmed with responsibilities, a man strug...	What's the more plausible RESULT?	He maintained a regular and consistent communi...	He forgot to call his parents altogether.	B	Effect	High	0.479486
565	A fitness enthusiast, armed with knowledge and...	What's the more plausible RESULT?	He wrote a book on fitness.	He improved his physical health and built a ha...	A	Effect	High	0.515482
77	The family's celebration was not just about a ...	What's the more plausible CAUSE?	They were marking the end of years of financia...	They were celebrating a festival.	B	Cause	High	0.337150
181	His spent the night exploring a new technical ...	What's the more plausible CAUSE?	There was a critical bug that needed fixing be...	He wanted to learn a new programming language.	B	Cause	High	0.205890

c. Plot the Density Distribution



- The **Cause** (blue) distribution peaks at a higher similarity value (~0.4) compared to the **Effect** (red) distribution, which peaks slightly lower (~0.35).
- However, both distribution shows that Similarity between Cause and Cause, or Effect and Effect, are all **LOW** at their peaks (0.35 - 0.4). This suggest that the embedding techniques cannot be used to determine the semantic similarity between two causes (or two effects) related to a given Premise
- There is a significant overlap between the two distributions, particularly in the 0.3 to 0.5 similarity range. This overlap suggests:
    - Some causes and effects might have similar levels of semantic similarity.
    - There might be ambiguity in distinguishing causes from effects based purely on semantic similarity.

4. Diagram of semantic relatedness

NLI dataset

1. Original Dataset

	Premise	Hypothesis	Question	Option A	Option B	Option C	Answer	Source
0	I don't know exactly when.	i don't know when.	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	A	MNLI
1	Women weaving baskets with a child in between ...	Women sewing clothes while a baby sleeps.	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	C	SNLI
2	uh-huh um-hum um-hum yeah that's because my th...	We usually just get ice cream so my girlfriend.	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	C	MNLI
3	but i right but i agree it has to be at their ...	People have different situations in life, so i...	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	A	MNLI
4	When Mr. Hastings and Mr. Lawrence came in yes...	Your mistress wrote letters last night, can yo...	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	B	MNLI

- Premise: A sentence or statement that serves as the context or basis for inference.
- Hypothesis: A sentence that needs to be evaluated against the premise to determine the relationship.
- Question: A generic or specific question asking about the logical entailment relationship between the premise and hypothesis.
- Option A, Option B, Option C: The possible answers indicating the type of relationship:
- Entailment: The hypothesis is logically entailed by the premise.
- Neutral: The hypothesis is neither entailed nor contradicted by the premise.
- Contradiction: The hypothesis contradicts the premise.
- Answer: The correct answer among the options.
- Source: The origin of the data, e.g., MNLI (Multi-Genre Natural Language Inference), SNLI (Stanford Natural Language Inference), etc.

2. Processing techniques

a. Label the expected cosine similarity

- If the relationship between the Premise and corresponding Hypothesis in a row is:
  - **Contradiction:** Then their cosine similarity should be Low
  - **Neutral:** Then their cosine similarity should be Medium
  - **Entailment:** There is no guarantee that their cosine similarity will be High, or lied in a specific range, since this is difficult to say. It is set to be “Unpredictable”

b. Embedding

i. Choose and setup the embedding model: **nv-embed-v2** (https://huggingface.co/nvidia/NV-Embed-v2)

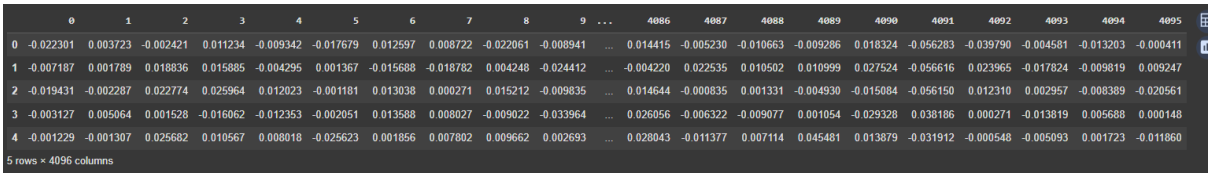
ii. Embed the Premises and Hypotheses:

- Premises’ Embedding results:

	0	1	2	3	4	5	6	7	8	9	...	4086	4087	4088	4089	4090	4091	4092	4093	4094	4095
0	-0.020892	0.016839	0.000331	0.003352	-0.012434	-0.018491	0.014293	0.015357	-0.017695	-0.004734	...	0.013639	0.001797	-0.010607	-0.005760	0.015742	-0.049329	-0.046134	-0.003213	-0.001324	-0.002122
1	-0.021711	-0.000284	0.002632	0.033610	-0.003357	-0.026663	-0.004828	-0.004135	0.015457	0.007796	...	-0.006478	0.015285	0.011420	-0.013303	0.013706	-0.045454	0.005249	0.005884	-0.003026	-0.002350
2	-0.011432	0.013519	0.011138	-0.000431	-0.003356	-0.024145	0.021229	-0.001147	-0.011116	0.011426	...	-0.008029	-0.003248	0.021401	-0.023195	-0.037904	-0.066127	0.007112	-0.010105	-0.011273	-0.002150
3	-0.031937	0.017538	0.007932	0.011127	-0.003398	0.013168	-0.003599	-0.009629	-0.025356	-0.013288	...	0.006218	0.019983	-0.014215	-0.024259	0.002646	0.029942	-0.022710	-0.001186	-0.005061	-0.014353
4	0.010272	0.002715	-0.002747	0.025306	-0.015044	-0.028360	0.009094	0.020080	0.019158	0.001963	...	0.011735	-0.011912	0.003297	0.007369	0.007836	-0.004732	0.014980	0.012852	-0.020928	-0.021752

5 rows × 4096 columns

- Hypotheses’ Embedding results:



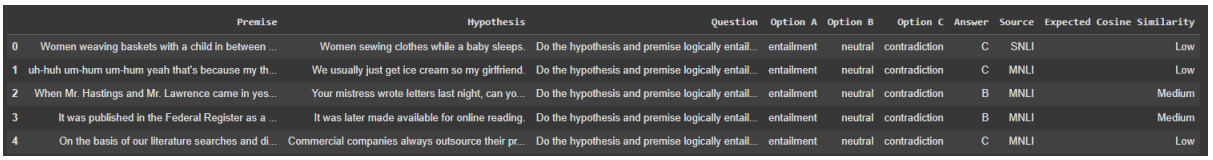
	0	1	2	3	4	5	6	7	8	9	...	4086	4087	4088	4089	4090	4091	4092	4093	4094	4095
0	-0.022301	0.003723	-0.002421	0.011234	-0.009342	-0.017679	0.012597	0.008722	-0.022061	-0.008941	...	0.014415	-0.005230	-0.010663	-0.009286	0.018324	-0.056283	-0.039790	-0.004581	-0.013203	-0.000411
1	-0.007187	0.001789	0.018836	0.015885	-0.004295	0.001367	-0.015688	-0.018782	0.004248	-0.024412	...	-0.004220	0.022535	0.010502	0.010999	0.027524	-0.056616	0.023965	-0.017824	-0.009819	0.009247
2	-0.019431	-0.002287	0.022774	0.025964	0.012023	-0.001181	0.013038	0.000271	0.015212	-0.009835	...	0.014644	-0.000835	0.001331	-0.004930	-0.015084	-0.056150	0.012310	0.002957	-0.008389	-0.020561
3	-0.003127	0.005064	0.001528	-0.016062	-0.012353	-0.002051	0.013588	0.008027	-0.009022	-0.033964	...	0.026056	-0.006322	-0.009077	0.001054	-0.029328	0.038186	0.000271	-0.013819	0.005688	0.000148
4	-0.001229	-0.001307	0.025682	0.010567	0.008018	-0.025623	0.001856	0.007802	0.009662	0.002693	...	0.028043	-0.011377	0.007114	0.045481	0.013879	-0.031912	-0.000548	-0.005093	0.001723	-0.011860

5 rows × 4096 columns

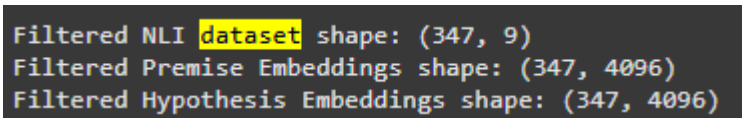
c. Drop the entailments

Since it is difficult to manually label the samples with “Entailment” relationship, these have been removed.

- Original Dataset:



	Premise	Hypothesis	Question	Option A	Option B	Option C	Answer	Source	Expected Cosine Similarity
0	Women weaving baskets with a child in between ...	Women sewing clothes while a baby sleeps.	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	C	SNLI	Low
1	uh-huh um-hum um-hum yeah that's because my th...	We usually just get ice cream so my girlfriend.	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	C	MNLI	Low
2	When Mr. Hastings and Mr. Lawrence came in yes...	Your mistress wrote letters last night, can yo...	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	B	MNLI	Medium
3	It was published in the Federal Register as a ...	It was later made available for online reading.	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	B	MNLI	Medium
4	On the basis of our literature searches and di...	Commercial companies always outsource their pr...	Do the hypothesis and premise logically entail...	entailment	neutral	contradiction	C	MNLI	Low



```
Filtered NLI dataset shape: (347, 9)
Filtered Premise Embeddings shape: (347, 4096)
Filtered Hypothesis Embeddings shape: (347, 4096)
```

- The corresponding Embedding results for Premises and Hypotheses are saved in:

d. Separate the “Neutral” and “Contradiction” samples

- Separate the embedding results into:
  - Neutral premise embeddings

Link:

  - Neutral hypothesis embeddings

Link:

  - Contradiction premise embeddings

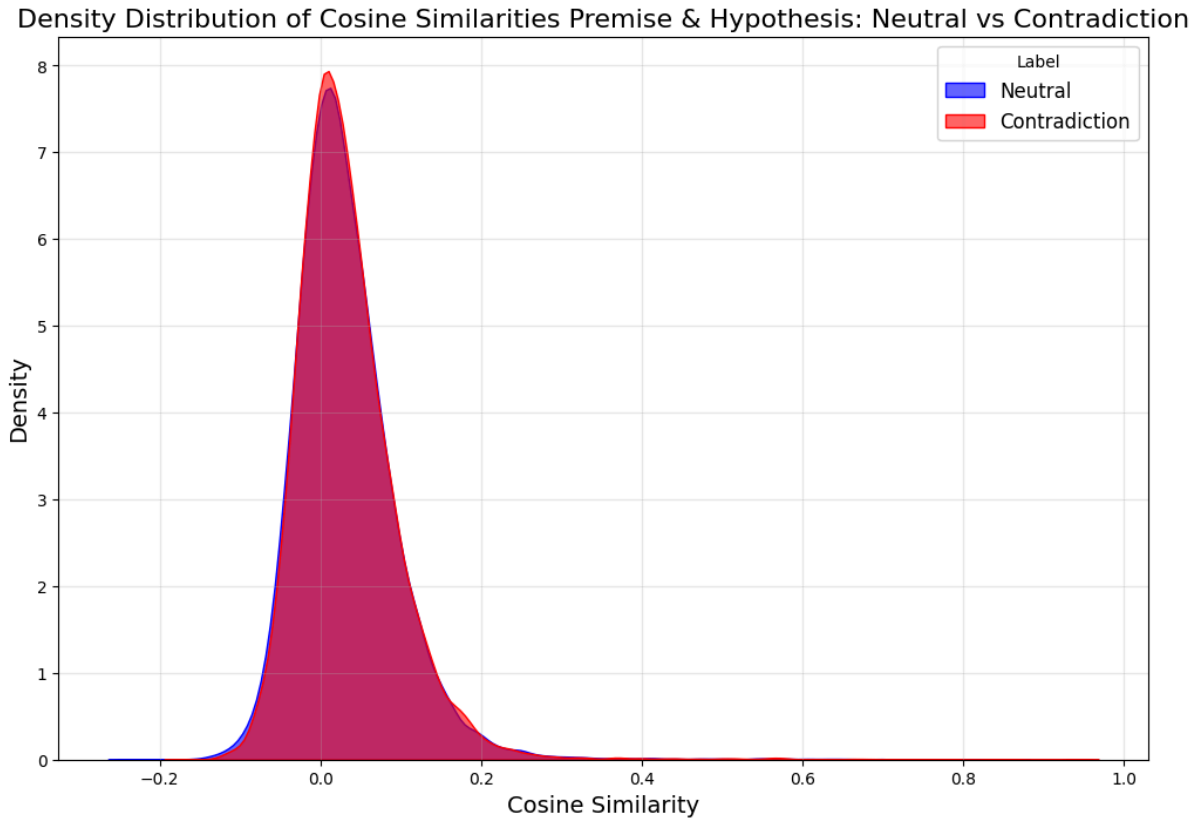
Link:

  - Contradiction hypothesis embeddings

Link:

3. Experiment

- Compute the Cosine Similarity and Plot the Density Distribution



- The density curves for **Neutral** and **Contradiction** significantly overlap, especially around the -0.1 to 0.1 similarity range.
- This suggests that for many samples, the cosine similarity between the premise and hypothesis is moderately similar for both labels. In other words, the Cosine Similarity verifying technique has failed to capture the different NLI relationships between a given pair of Premise and Hypothesis

4. Diagram of semantic relatedness

GitHub Repo

https://github.com/trungkiennguyen22082004/Embedding_Experiments/tree/main

Reflection

Previous Experiments