[History Lab] Final Updates

Top Level

- Contact Information (for project continuation)
  - liconnor2003@gmail.com
  - +1 (404) 642-6292

Similar Document Testing

Following the instructions, I used the CIA document collection to validate the similarity search. Because nearly all of these files already sit in the 2.5 million-document embedding index I built, they make an ideal benchmark. The Mini-LLM-v6 model performed well—each of the ~30-ish queries returned the expected (similar/same) documents, and I've highlighted a few of those matches in the chart below.

Original Document	Similar Doc #1	Similar Doc #2	Similar Doc #3	Similar Doc #4	Similar Doc #5
CIA-RDP79T00975A000100190001-9	CIA-RDP79T00975A000800640001-5	CIA-RDP79T00975A000800650001-4	CIA-RDP79T00975A000800630001-6	CIA-RDP79T00975A000800340001-8	CIA-RDP79T00975A000800090001-6
CIA-RDP79T00975A002400130001-0	CIA-RDP79T00975A000100350001-1	CIA-RDP79T00975A001300150001-0	CIA-RDP79T00975A000700300001-0	CIA-RDP79T00975A009100350001-2	CIA-RDP79T00975A000100580001-6
CIA-RDP79T00975A006000340001-7	CIA-RDP79T00975A005700090001-9	CIA-RDP79T00975A006900470001-4	CIA-RDP79T00975A003900230001-3	CIA-RDP79T00975A005700350001-0	CIA-RDP79T00975A005500370001-0
CIA-RDP79T00975A005200080001-5	CIA-RDP79T00975A005200110001-1	CIA-RDP79T00975A005300440001-4	CIA-RDP79T00975A004800510001-2	CIA-RDP79T00975A004900320001-2	CIA-RDP79T00975A005200280001-3
CIA-RDP79T00975A004800160001-1	CIA-RDP79T00975A005400060001-5	CIA-RDP79T00975A004800440001-0	CIA-RDP79T00975A004800180001-9	CIA-RDP79T00975A005000330001-9	CIA-RDP79T00975A004800350001-0
CIA-RDP79T00975A005100060001-8	CIA-RDP79T00975A007100350001-4	CIA-RDP79T00975A004000050001-1	CIA-RDP79T00975A007400470001-8	CIA-RDP79T00975A004700290001-8	CIA-RDP79T00975A004800260001-0
CIA-RDP79T00975A006100300001-0	CIA-RDP79T00975A007900010001-3	CIA-RDP79T00975A007600430001-0	CIA-RDP79T00975A025000100001-2	CIA-RDP79T00975A023000030002-1	CIA-RDP79T00975A024200020002-9

To double-check, I repeated the experiment on a random set of ~30 unrelated documents. The model still surfaced semantically close items (in meaning and purpose, maybe not in content as much), at least from what I can tell. I've summarized some of those results in the table below and would love your thoughts on whether the pairings look right to you.

Original Document	Similar Doc #1	Similar Doc #2	Similar Doc #3	Similar Doc #4	Similar Doc #5
LOC-HAK-122-4-3-8	CIA-RDP90M00005R000700060012-1	CIA-RDP79M00467A003100070016-4	CIA-RDP80B01495R001100010020-0	LOC-HAK-449-4-37-9	CIA-RDP90M00005R001100090008-8
CIA-RDP79B00972A000100610014-9	CIA-RDP79B00972A000100610013-0	LOC-HAK-537-1-5-9	CIA-RDP84B00049R000601590008-8	CIA-RDP79B00972A000100610015-8	CIA-RDP93B01478R000100130002-3
CIA-RDP79B01709A000900050003-6	CIA-RDP79B01709A000900030001-0	CIA-RDP79B01709A000900070003-4	CIA-RDP79B01709A000700040005-7	CIA-RDP79B01709A000700050008-3	CIA-RDP79B01709A000900070005-2
frus1969-76v12d172	frus1969-76v13d259	frus1969-76v12d157	frus1969-76v40d100	frus1969-76v12d144	frus1969-76v29d237
LOC-HAK-33-5-10-8	CIA-RDP83-01004R000100190001-3	CIA-RDP78-04718A002600410125-2	CIA-RDP80M01048A001500130077-0	CIA-RDP86T00268R000700070003-7	CIA-RDP78-04718A000800200025-6

NLP Statistics

- **Polarity**: This score reflects the overall emotional tone of the text. It ranges from -1 (very negative) to +1 (very positive). A score near zero means the language is fairly neutral in tone, while higher or lower scores suggest more emotionally charged language.

- **Subjectivity**: This measures how much of the text is based on opinion rather than fact. It ranges from 0 (completely objective) to 1 (very subjective). A low score means the writing sticks to facts, while a higher score indicates the presence of personal views, emotions, or interpretations.

Below is a table listing the documents along with their sentiment scores and brief notes explaining each one. While there may be some confirmation bias in my interpretations, I hope this provides a helpful starting point for evaluating how these scores reflect the content and tone of the documents.

Document	Polarity	Subjectivity	Notes (AI-assisted)
CIA-RDP72-00337R000300040017-4	0.08	0.43	The document scores low on polarity (0.08) because it sticks mostly to neutral reporting, with only light emotional tone in quotes. The subjectivity (0.43) reflects a mix of factual content and opinionated statements from politicians reacting to Nixon's budget and missile plans.
CIA-RDP80-01601R000300340109-5	0.06	0.42	The low polarity score (0.06) reflects the text's largely neutral, factual tone, even though the topic—Soviet missile development—is serious. It avoids strong emotional language. The subjectivity score (0.42) suggests a mix of objective reporting with some interpretive framing and speculation, especially in the discussion of what the new missile sites might mean.
frus1969-76v32d242	0.10	0.49	The polarity score of 0.10 suggests a slightly positive tone, likely due to Nixon's emphasis on achieving a strong, defensible arms agreement and maintaining national credibility. While not emotional, his language shows cautious optimism and strategic intent. The subjectivity score of 0.49 reflects a high level of personal reasoning, political calculation, and speculative framing rather than strictly objective statements—consistent with a behind-the-scenes conversation about public perception and policy.
frus1969-76v33d56	0.17	0.46	The polarity score of 0.17 reflects Nixon's confident and slightly positive tone—he expresses hope for arms reduction and appreciation for congressional support. The subjectivity score of 0.46 indicates a mix of factual statements and policy positions, with moderate personal judgment as he outlines strategy and defends the need for continued military strength alongside negotiations.
CIA-RDP91-00901R000500070013-5	0.04	0.41	The low polarity score (0.04) reflects the critical and serious tone throughout the piece—Colby criticizes Reagan's approach without using strongly emotional or positive language. The subjectivity score of 0.41 comes from Colby's personal analysis and opinion, especially in his arguments for a nuclear freeze and against perceived appeasement. While grounded in policy discussion, the language leans interpretive rather than purely factual.

Conclusion: While polarity and subjectivity scores are interesting at a glance, I've found they don't offer much insight when applied to historical documents. The polarity scores tend to cluster between 0.05 and 0.15 across nearly everything, and subjectivity hovers around 0.4 — which makes sense, but also limits how useful the scores are.

Most of these documents are written in formal, neutral language, even when discussing controversial or high-stakes topics. That keeps polarity low by default — there's just not much emotionally charged language for the model to pick up on. At the same time, subjectivity floats around the same range because these texts often mix fact with subtle opinion or strategic framing — not quite full-on editorial, but not pure reporting either.

Also, these models are probably trained on more casual or modern texts (like product reviews or social media), so they miss the nuance and subtext baked into diplomatic or bureaucratic writing. Just because a memo sounds calm doesn't mean it's neutral in intent — and the scores don't always catch that. So while the sentiment outputs aren't useless, I think they're a pretty blunt tool here. They give you a general sense of tone, but not much beyond that — especially when everything lands in the same small range.