Can a large language model classify dementia-related *possible delusion* without fine-tuning?

David H. Roberts

Abstract

Recent scaling of large language models (LLM) has demonstrated emergent capabilities in few-shot prompting, whereby a limited set of input-output examples enhance performance on classification tasks without fine tuning of model embeddings. This advancement may enable qualitative researchers without programming experience to accomplish their goals with natural language prompts. In this study, we simulate a use case in which qualitative researchers review a corpus of YouTube comments in which presumed knowledgeable informants relate narrative text regarding an individual with dementia. Reviewers attempt to identify evidence of possible delusions and extract supporting excerpts, generating a qualitative coding rubric in the process. This rubric then serves as a few-shot prompt to a LLM which attempts the same task. Results suggest that fine-tuning of the base LLM model would likely be necessary to achieve satisfactory performance.

Introduction

Dementia is a condition characterized by disabling cognitive impairment which interferes with the activities of daily living. The condition has various causes, most commonly Alzheimer's disease, which affected an estimated 6 million people in the United States in 2022 (1). In addition to the cardinal symptom of memory loss, *at least* 23% of individuals with a diagnosis of Alzheimer's disease exhibit delusions (2), typically defined as strongly held false beliefs (3). Delusional symptoms engender the need for a *knowledgeable informant* (KI) in dementia care to serve as an alternative source of information to report or disconfirm delusional thinking exhibited by an individual with dementia (IWD). Given that many IWD are cared for in a community setting, knowledgeable informants are commonly *informal caregivers*, i.e. unpaid friends and family without formal training. Delusional symptoms are especially challenging to these caregivers, who may be inaccurately perceived as antagonists by a delusional IWD. Responding appropriately is a delicate affair, and may require tacit acceptance of an untrue belief in order to prevent or reduce agitation in the IWD (4). These complex interactions, coupled with the physical challenge of assisting an IWD with the activities of daily living can be overwhelming to informal caregivers. Tellingly informal caregivers report symptoms of depression at much higher rates than the general population (5).

The application of natural language processing to social media data for use cases in mental health care is an active area of research, with the majority of studies focusing on the identification of anxiety, depression, and suicidality (6). At least one study attempts to identify references to delusion and other psychotic symptoms in individuals with dementia using a corpus of clinical notes from home health agencies (7). As clinical notes are created by trained healthcare professionals, they do not simulate a scenario in which a knowledgeable informant lacks the vocabulary to explicitly categorize psychotic symptoms. Note, an exhaustive literature review was not conducted as part of this study.

Large language models (LLM's) are a class of transformer-based language models trained on large corpora via semi-supervised learning. Recent scaling of parameter count and training time have led to *emergent* capabilities in "few-shot prompting", in which a few input-output classification examples alongside an instructional prompt are provided to a LLM as a preamble for a given task (8). For many use cases, this prompt strategy can substitute more cumbersome gradient updates to model weights, also called "fine-tuning". Importantly, users of LLM's can provide few-shot prompts in natural language, creating an intuitive experience for users without programming experience. This is a promising development for qualitative researchers who may benefit from using a pre-trained LLM to screen large corpuses without relying on programmers. In this pilot study. we used a pre-trained LLM in OpenAI's GPT series to classify *possible delusion* in individuals with dementia, as reported by presumed KI's in YouTube comments (9). To simulate a use case in which qualitative researchers deploy the LLM "intuitively", we developed a qualitative coding rubric which served as a few-shot prompt for LLM classification. Results suggest fine-tuning of the LLM would be required to create a satisfactory experience for qualitative researchers without programming expertise.

Solution Overview

Figure 1 provides an illustration of the overall workflow. The first step is to extract YouTube comments from a subset of videos which relate to the topic of dementia. Comments are then passed through a *neural coreference* classifier developed by spaCy's principal maintainers, Explosion.ai (10). This model identifies entities in a text, then identifies successive references to that entity via a pronoun or noun phrases. The extracted *coreference chains* (also called "coreference clusters") were then manually reviewed to create a list of entities which are most likely to be the subject of narratives related to dementia, as reported by a KI (the commenter). Comments which contained a *long chain of references* (greater than three) to an entity of interest, were identified as highly probable KI narratives regarding dementia. As an example: "*My g-ma* suffered with this disease. Even though *she* lost the ability to speak, *she* still played piano every time we showed up. We knew we were loved. I miss *her* a lot."

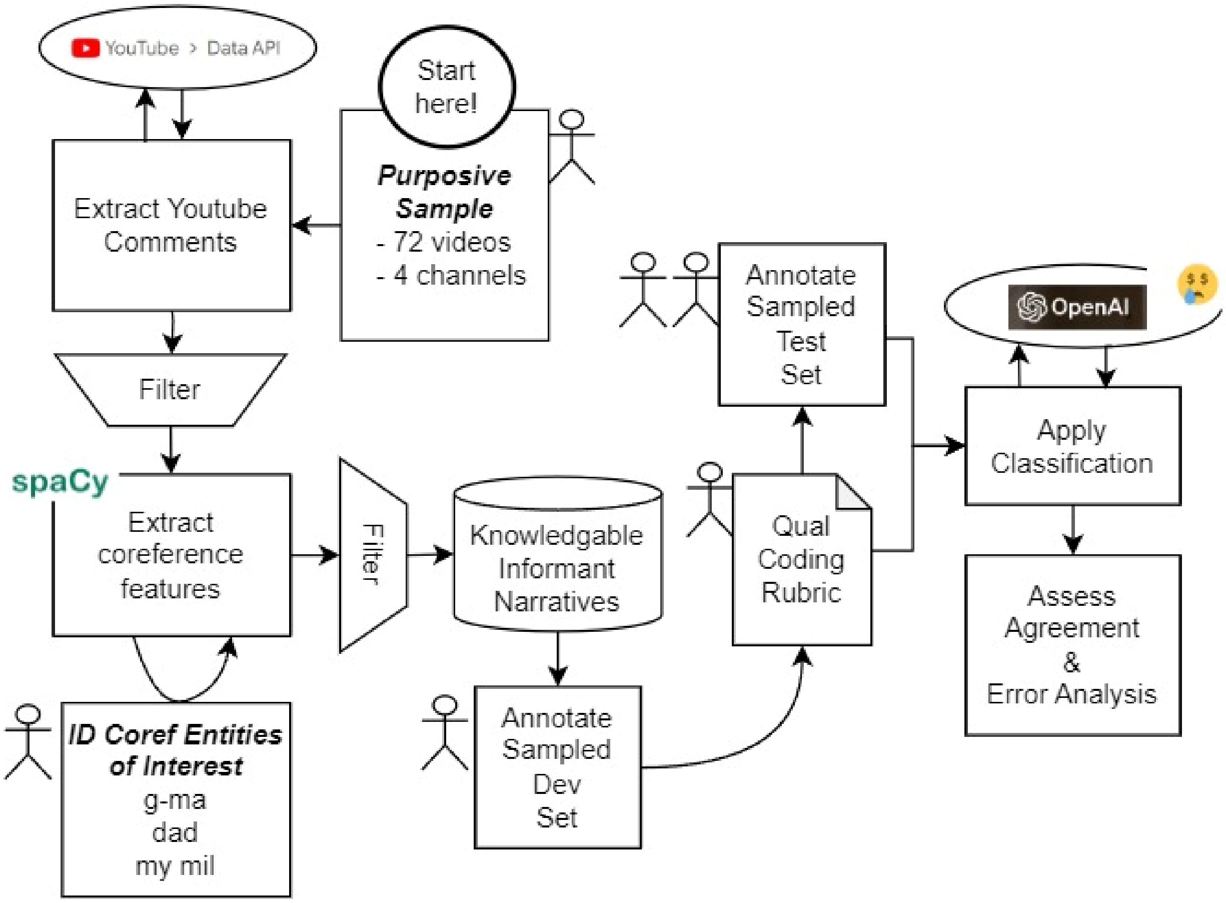

Figure 1. Overall workflow. Stick figures indicate unautomated components

Given this database of KI narratives, a randomly sampled set was annotated in order to develop a qualitative coding rubric. This rubric served two purposes. First, the rubric was used to annotate a randomly sampled reference set to test agreement with LLM-based classifications. Second, to test the notion that and LMM-based classifier could be used *intuitively* by a qualitative researcher, a portion of the qualitative rubric which included examples, was provided to the LLM-based classifier as a few-shot prompt. Note, while some programming skill was required to complete this process, one can imagine a future state in which qualitative researchers perform the same tasks in a chat-based interface, providing natural language instructions.

This solution has *several* advantages. First, the neural coreference and LLM components required minimal data cleaning required to both compile the KI narrative database and apply "possible delusion" classifications. The capability of the spaCy coreference solution to identify colloquial expressions as entities, i.e. "g-ma" for grandmother, was valuable. Similarly, minimal preprocessing was applied to comments prior to classification by the LLM. This low effort approach to preprocessing yielded a relatively high quality corpus of KI narratives for annotation. While an exhaustive review of the KI narrative corpus was not conducted, annotation of a random sample suggested over 90% were in fact KI narratives regarding an IWD. Importantly, the LLM component also supports various languages which would facilitate internationalization of the classification model.

This solution has numerous limitations. Most importantly, it relies on a series of "black box" components which are opaque to the researcher. First, the proprietary YouTube search engine is relied upon to gather an initial corpus of comments which *most likely* relate to dementia. The details of the search engine's implementation are not publically available. The spaCy coreference model is relatively transparent in that the algorithm is outlined in an academic paper and the underlying training data is publically available (11). However, the model is essentially unexplainable in that neural network weights do not have a straightforward interpretation. In cases where the coreference component behaves unexpectedly, for instance, by including an entire sentence as a single entity, researchers can only speculate as to the cause. Similarly, it is impossible to quantify how often the model failed to identify a relevant entity without an extensive review of the corpus of comments. This work was not completed as part of this study.

Finally, comments are classified by the OpenAI GPT-3 based 'text-davinci-003' model, which is particularly opaque (9). While the training sets and algorithms underlying OpenAI models have in the past made public, recent competitive dynamics among large technology companies has led to obfuscation of the details underlying training of more recent models. For example, OpenAI's GPT-4 technical report states that the model was trained "using both publicly available data (such as internet data) and data licensed from third-party providers", citing "the competitive landscape" for the lack of detail (12). Importantly, OpenAI models are non-deterministic. Even with parameters tuning to minimize randomness in output, OpenAI indicates the models will be "mostly deterministic" (9). Finally, iteration of LLM prompts, commonly described as "prompt engineering", appears at present to be more of an art, although best a some best practices are accumulating (13).

A final, important limitation is the fact that LLM's predict the next sequence of tokens and *ultimately do not execute logic*. Future approaches may overcome this limitation by enhancing models to delegate to deterministic, rules-based processing when appropriate. However, in their current form, this issue would yield significant confusion among qualitative researchers attempting to use an LLM "intuitively".

Methods

Comment Dataset Compilation

A data set of YouTube comments was collected via two methods. First, all comments originating from several popular *dementia advice* YouTube channels were collected, notably CareBlazers (14). Second, we searched the YouTube website with the query "Alzheimer's OR dementia", sorted the results by view count, and included videos with more than 450,000 views as of March 7th, 2023. A selection of YouTube shorts which appeared in the search results were also included. Videos with content that was obviously promotional of specific drugs or nutritional supplements were excluded. Comments extracted via this process were passed through the default spaCy named entity recognition (NER) algorithm in order to redact names (15). Identifiers passed by the YouTube Data API (16), e.g. *comment_id*, *video_id*, and *channel_id*, were encrypted to prevent trivial re-identification of the originating user. In addition, comments with contact info in their text (emails, usernames, phone numbers), were deleted from the corpus when identified. The resulting corpus of YouTube comments was stored in a SQLite database which served as the *central storage component for all downstream processes*, including classification and annotation results.

KI Narrative Identification

The filters which reduced comments to presumed KI narratives are summarized in Figure 2 below. First, comment replies were excluded due to their tendency to include username information. Next comments with one only a single sentence were excluded under the assumption that they would be unlikely to contain sufficient a detail rich KI narrative. Coreference and linguistic features were then computed using a spaCy coreference model and stored to SQLite. In addition to the coreference chains themselves, features such as the length of chains, the number of coreference chains per comment, whether or not a chain incorporated a possessive pronoun and / or a noun in the root phrase were also included. After experimenting with several methods, the following process was used to identify entities of interest in KI narratives.

1. Identify all coreference chains with an entity reference that contains any possessive pronoun (*w1*), and a noun (*w2*) as the root of the phrase, e.g. "my 84 year old father".
2. Remove tokens between the possessive pronoun and the root noun. This would condense "my 84 year old father" to "my father".
3. Replace the root noun with it's lemma.

This process yielded a list of 1,006 tokens, which were ranked in order of their document-level frequency. This list was manually reviewed to identify entities that were most likely to be the subject of KI narrative (appendix A). A final corpus of presumed KI narratives was produced using the following filters:

1. Coreference chain includes an *entity of interest*, e.g. "grandma", "father", "mil" (mother in law). Note the possessive pronoun, e.g. "my", was not included as part of the filtering string.
2. The coreference chain containing the entity of interest has a length greater than or equal to four.
3. Exclusion of duplicate comments by comment text.

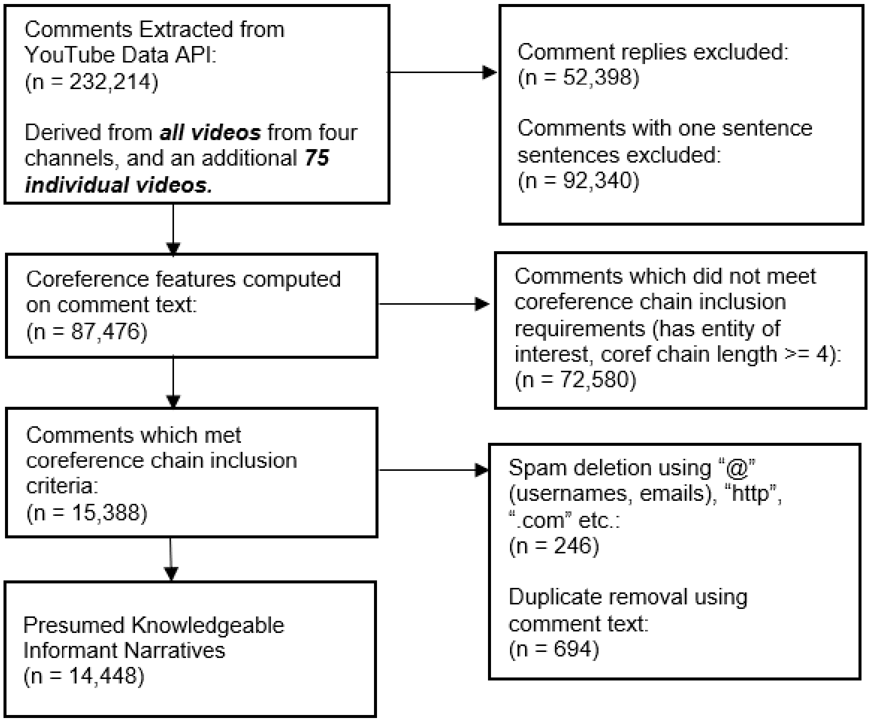
Comments Extracted from YouTube Data API: (n = 232,214)
Derived from ***all videos*** from four channels, and an additional ***75 individual videos.***
Comment replies excluded: (n = 52,398)
Comments with one sentence sentences excluded: (n = 92,340)
Coreference features computed on comment text: (n = 87,476)
Comments which did not meet coreference chain inclusion requirements (has entity of interest, coref chain length >= 4): (n = 72,580)
Comments which met coreference chain inclusion criteria: (n = 15,388)
Spam deletion using "@" (usernames, emails), "http", ".com" etc.: (n = 246)
Duplicate removal using comment text: (n = 694)
Presumed Knowledgeable Informant Narratives (n = 14,448)

Figure 2. Filtering process to compile "knowledgeable informant narratives"

Annotation and Coding Rubric Creation

A random sample of 215 KI narratives served as the *development set* for creating a qualitative coding rubric to classify "possible delusion" (appendix B). The rubric was developed iteratively, adding criteria with supporting examples as they were identified in the development set. This rubric was then applied to an additional *324 randomly sampled KI narratives*, which served as the reference set in assessing LLM model performance. Given that this was a pilot test, sample sizes were limited by resource constraints in annotation. After limited training, KI narratives could be classified at roughly 1.6 per minute.

An important consideration in the coding rubric was distinguishing "possible delusion" from actual delusions. Given a corpus of YouTube comments, it is typically left ambiguous whether the commenter attempted to correct a false belief, and if so, whether the individual with that false belief changed their mind. Essentially, it is not possible to know with certainty the duration of a false belief, or whether the belief was "fixed", i.e. persisted despite evidence to the contrary. However, a "fixed" false belief is a key criteria in diagnosing delusional disorders during routine patient care (2). With this constraint, we assumed that false beliefs were persistent, unless otherwise indicated by the commenter. Note, from the perspective of patient care, false beliefs may be stressful to the caregiver regardless if they are persistent. For example, if an IWD continuously suggests someone stole their purse, the situation still requires effort and finesse to defuse, even if a caregiver succeeds in convincing them it was never lost.

LLM Prompting

LLM prompts were developed iteratively, using the 'text-davinci-003' model via the OpenAI API (9). The 'text-davinci-003' model was chosen due to its natural language instruction following capabilities, which would be required for intuitive use by qualitative researchers. In addition, the 'text-davinci-003' model is designed for text insertion and replacement which were important functions for extracting excerpts. While it is not possible to guarantee reproducible results via the OpenAI API, parameters of the LLM model were chosen to *minimize stochasticity* in outputs. The final prompt is summarized in Figure 3 below. Although the OpenAI API does not always yield a properly formatted response, the results were consistent and parseable with a simple Python function.

```
You are qualitative researcher reviewing YouTube comments from channels related to dementia.
You are following a coding rubric which outlines specific cases to classify as possible delusion.


## Qualitative Coding Rubric

{qualitative coding rubric with criteria and examples}

…

### Task Instructions

Does the following text show evidence of a dementia patient which may have delusions? If yes,
extract an excerpt from the text to justify your response.


Structure your response as follows:

Possible Delusion: true / false

Excerpt: Excerpt


Text: {text}
```

Figure 3. The final LLM prompt included the qualitative coding rubric as a few-shot prompt, task instructions, and a comment to analyze.

Evaluation

We evaluated classification performance under two scenarios, first with all instances in the test set, then excluding instances which contained the string "delus" in the comment text. The second approach sets a higher bar for performance, simulating the challenging task of identifying possible delusion when a knowledgeable informant lacks the vocabulary, knowledge, or desire to identify delusion themselves. We report results for this more challenging use case only. From standard contingency tables, we calculated the following outcome values: positive predictive value (PPV), sensitivity, F-score, specificity, accuracy. A brief review of misclassified instances was conducted. Finally, the LLM's ability to extract excerpts according to prompt instructions was assessed on a set of 1337 instances where an LLM classification was available. This set included instances from both the annotated sets (development and test), as well as instances which were not annotated. We evaluated first whether the LLM correctly executed instructions to only extract excerpts in cases of positive classifications. Given the tendency of LLM's to "hallucinate" facts, we then assessed if the extracted excerpts were found in the original comment text.

Note, the specific Python libraries, including versions, used to conduct this work are documented in the public code repository (17).

Results

Classification Results

Classification results for the more challenging use case which excludes comments with explicit references to delusion are presented in Table 1 below. Per annotators, 15.3% of comments in a test set of 313 instances showed evidence of possible delusions in an individual with dementia. Note, one test set instance was discarded due to a poorly formatted response from the LLM which did not conform to specified format provided in the instructional prompt.

Current classification performance would not provide a satisfying experience for qualitative researchers. 73% sensitivity indicates that over one fourth of comments with evidence of possible delusion are incorrectly classified and would be filtered if the model were used as a pre-screen to human review. 56% PPV indicates that nearly one half of comments included for human review would not be relevant. 87% accuracy is misleading as a success metric and is primarily due to class imbalance. Review of false negative and positive instances did not yield any obvious trends underlying classification errors. The sample size in each class was insufficient to rigorously assess the impact of comment level features, for instance, number of tokens, on classification accuracy. A review of 10 randomly sampled true positive excerpt extractions showed promising results. In eight of ten cases, the LLM model extracted excerpt was similar or identical to the excerpt extracted by annotators to support their classification. Overall, these results suggest the need for fine-tuning via gradient adjustments to achieve satisfactory recall.

Table 1: Classification Report – Excluding comments with explicit references to delusion

Outcome Metric	**Value (Support)**
PPV	0.56 (35/62)
Sensitivity	0.73 (35/48)
Specificity	0.90 (238/265)
Accuracy	0.87 (273/313)
F-Score	0.64
Annotated Prevalence	0.15 (48/313)

Instruction Following Results

In addition to classification accuracy, the capability of the LLM to follow instructions in extracting excerpts was assessed. Note, the performance reported in this section does not consider the accuracy of classifications, nor the quality of extracted excerpts, i.e., whether the excerpt was supported a positive classification. When making a positive classification, the LLM appears to have *extracted* an excerpt 100% of the time, i.e. it correctly executed prompt instructions. When making a negative classification, the LLM extracted an excerpt 9% of the time, contrary to prompt instructions.

Whether extracted excerpts appeared in the original comment text was assessed programmatically. First, for 378 instances in which an excerpt was extracted, the original comment and LLM excerpts were stripped of all characters except digits and alphanumeric characters, then lowercased. Given these transformations, the LLM extracted excerpt was a substring of the original comment text 83% of the time. For the remaining 17% of extracted excerpts, a string similarity score was calculated for all substrings of the original comment with the same length of the LLM excerpt (18). The maximum similarity score at the comment level was then recorded as a measure of how faithfully the LLM extracted excerpts.

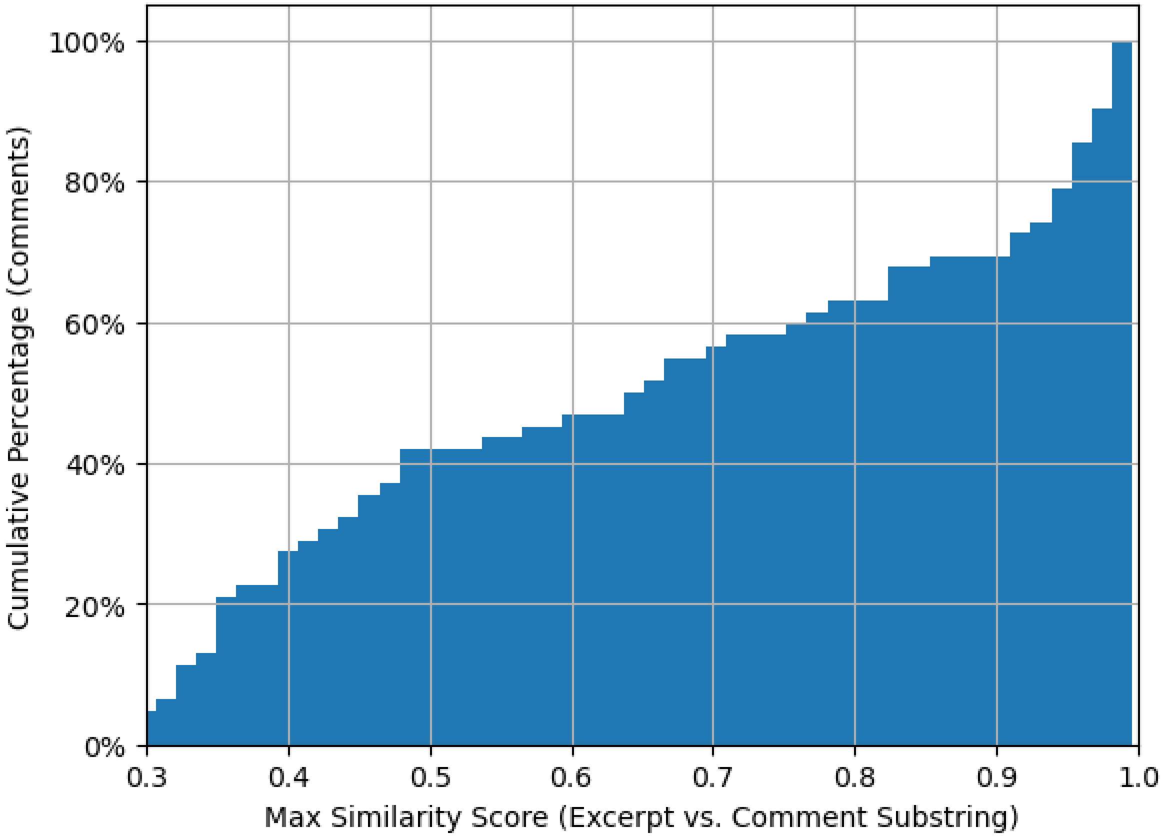

Figure 4: The maximum similarity score between the excerpt and comment substrings. Note, only cases where excerpts do not exactly match a substring in the comment are included.

A review of specific bands of similarity scores revealed trends in the LLM's behavior. For cases with a similarity score above 0.9, the differences between the extracted excerpt and the original text would likely not be relevant to a qualitative researcher. For example, in several cases, the LLM clarified the subject of a phrase with a pronoun, e.g. changing "but has 1 or 2 delusions a month" to "she but has 1 or 2 delusions a month". In another case, the LLM corrected a spelling error, e.g. changing "sit there and cpunt the people" to "sit there and count the people". Cases with similarity scores between 0.7 and 0.9 tended to be a combination of two non-contiguous excerpts from the original comment, at times joined with the string "[…]". This reflects the LLM's training for the purpose of summarization. Scores lower than 0.7 tended to include "hallucinated" text by the LLM or reiteration of irrelevant examples from the few-shot prompt.

These results illustrate both the remarkable ability of LLM's to perform a task given natural language instructions as well as their inherent unpredictability. While a successfully extracted excerpt resembles a "program", the LLM does not execute rules-based logic. Instead, *it appears to execute logic* by predicting the next token. Ultimately, *how a task is performed* will be determined probabilistically by the distribution of instances in the LLM's training corpus. The model may extract an excerpt verbatim, prefix with a pronoun, or hallucinate, but the user cannot know which in advance. One approach to address this constraint is to *ground* the LLM in existing factual sources using deterministic programs, then prompt the LLM to correct itself (19). For example, in this use case, an excerpt altered by the LLM could be grounded by searching for the closest matching substring in the original comment text, then returning the original substring.

Conclusion

In this study, we compiled a corpus of YouTube comments in which knowledgeable informants relate narratives regarding individuals with dementia. Using a large language model, we attempted to classify if a comment contained evidence of *possible delusion*, instructing the model to *extract excerpts as evidence* in positive cases. The results suggest that fine-tuning is required to achieve performance which would be satisfactory to a qualitative researcher, attempting to use an LLM *intuitively*. The results also highlight challenges in working with large language models which may produce *unexplainable* and *unpredictable* outputs. Possible next steps include fine-tuning an LLM for this use case as well as shifting to an open source alternative to ensure reproducibility of results.

Acknowledgements

We would like to acknowledge Lauren Fisher RN, PHD candidate for her feedback in developing the qualitative coding rubric. Daniel Mowery PHD assisted by annotating a portion of the test set to assess the quality of the coding rubric. Patryk Orzechowski PHD assisted with feedback regarding methodology.

References

1. 2022 Alzheimer's Disease Facts and Figures. Alzheimers Dementia: Alzhiemer's Association; 2022.
2. Ismail Z, Creese B, Aarsland D, Kales HC, Lyketsos CG, Sweet RA, et al. Psychosis in Alzheimer disease - mechanisms, genetics and therapeutic opportunities. Nat Rev Neurol. 2022;18(3):131-44.
3. Cummings J, Pinto LC, Cruz M, Fischer CE, Gerritsen DL, Grossberg GT, et al. Criteria for Psychosis in Major and Mild Neurocognitive Disorders: International Psychogeriatric Association (IPA) Consensus Clinical and Research Definition. Am J Geriatr Psychiatry. 2020;28(12):1256-69.
4. Edmonds N. Dementia and Delusions: Why do delusions happen and how to respond? YouTube: Dementia Careblazers; 2019 [Available from: https://www.youtube.com/watch?v=cVvi2X39Sio.
5. Hoffman AK. The Reverberating Risk of Long-Term Care. Yale Journal of Health Policy, Law and Ethics. 2015;15(1).
6. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med. 2022;5(1):46.
7. Topaz M, Adams V, Wilson P, Woo K, Ryvicker M. Free-Text Documentation of Dementia Symptoms in Home Healthcare: A Natural Language Processing Study. Gerontol Geriatr Med. 2020;6:2333721420959861.
8. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models Internet: arXiv [Preprint]; 2022 [Available from: https://arxiv.org/abs/2206.07682.
9. GPT-3.5 API Documentation Internet: OpenAI; [Available from: https://platform.openai.com/docs/models/gpt-3-5.
10. Kádár Á, McCann POL, Hudson R, Schmuhl E, Landeghem SV, Boyd A, et al. Website: Explosion.ai. 2022. Available from: https://explosion.ai/blog/coref.
11. Dobrovolskii V. Word-Level Coreference Resolution Internet: arXiv [Preprint]; 2021 [Available from: https://arxiv.org/abs/2109.04127.
12. OpenAI. GPT-4 Technical Report Internet: arXiv [Preprint]; 2023 [Available from: https://cdn.openai.com/papers/gpt-4.pdf.
13. Awesome Prompt Engineering github.com2023 [Available from: https://github.com/promptslab/Awesome-Prompt-Engineering.
14. Dementia Careblazers YouTube.com [Available from: https://www.youtube.com/c/DementiaCareblazers.
15. SpaCy: Industrial-Strength Natural Language Processing Internet [Available from: https://spacy.io/.
16. YouTube Data API [Available from: https://developers.google.com/youtube/v3.
17. Roberts D. BMIN521 Final Project Github.com [Available from: https://github.com/dr00b/id_possible_delusion_in_KI_narrative_text.
18. Helper Functions for Computing Deltas [Available from: https://docs.python.org/3/library/difflib.html.
19. Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, et al. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback Internet: arXiv [Preprint]; 2023 [Available from: https://arxiv.org/abs/2302.12813.

Appendix A: Entities of interest in identification of knowledgeable informant narratives.
my mom
my mother
my dad
my grandma
my grandmother
my husband
my father
my grandfather
my grandpa
my wife
my sister
my mum
my brother
my friend
my aunt
my parent
my granny
my uncle
my life
my daddy
my partner
my grandparent
my mama
my lo
my nana
my sibling
my grandad
my patient
my client
my nan
my lowd
my mil
my momma
my neighbor
my papa
my boyfriend
my hubby
my gran
my gma
my girlfriend
my pop
my spouse
my stepdad
my granddad
my man
my gram
my girl
my lady
my auntie
my mam
my love
my ma
my stepfather
my ex
my fil

my neighbour
my stepmom
my dear
my mommy
my grammy
my grandmom
my step
my abuela
my gramp
my mamaw
my sis
my inlaw
my elder
my gramma
my great-
my nanny
my boss
my fiance
my gpa
my grampa
my mamma
my papaw
my aunty
my grama
my twin
my guy
my step-
my bro
my colleague
my fiancee
my gf
my godmother
my granddaddy
my nanna
my pa
my abuelita
my bf
my fiancé
my grandpas
my granpa
my hub
my maw
my moma
my nonna
my stepmother
my gp
my grandmomma
my grannie
my lowvd
my opa
my pawpaw
my sib
my sweetheart
my buddy
my companion
my dude

Appendix B: Qualitative Coding Rubric

Annotation Task Overview

The goal of this annotation task is to identify Youtube comments in which a ***commenter*** relates a story regarding an ***elderly individual with dementia*** which contains evidence of ***possible delusion***. The keyword here is possible. It is *impossible* to adjudicate the truth of statements made in a Youtube comment without concrete evidence. However, it is possible to make a reasonable guess.

In this context, there are two subtasks:

1. Indicate if the comment has evidence of “possible delusion” in an *individual with dementia*. Choose a criteria 1.1-1.6 as justification for any positive examples.
2. Extract specific sentences from the text to justify your response.

Underlying Assumptions

1. Given a purposive sample of videos and channels, comments relate primarily to dementia, even if a dementia specific keyword is not used explicitly. *Unless explicitly stated otherwise, assume commenters describing a person with a condition are referring to an individual with some form of dementia.*
2. Comments are from real people acting as “knowledgeable informants”, i.e. a third-party who may report or disconfirm beliefs held by the individual with dementia. They are not bots or sarcastic posters.
3. *Commenters statements are truthful*. For example, many delusions experienced by dementia patients relate to infidelity or theft. If a commenter asserts an individual with dementia’s claims are false, assume the commenter is correct.

Assumption 1 – Motivating Examples:

In the following comment, the user responds to the content of the video, which outlined strategies for addressing delusions in dementia patients. The commenter never explicitly mentions dementia however.

```
I'm finding myself in more and more situations, where being honest with my loved one just makes matters worse.  There are times when he appears to have presence of mind and a pretty good understanding of reality.  It's during those episodes of clarity, when he tends to make statements or ask questions.  Where I find my replies of truth to be  like walking into a trap.  As if I just opened Pandora's box to his pent up hostility, anger and verbal abuse.
```

Similarly, in the following comment, there is no discussion of dementia as a condition, or an underlying pathology. Based on context and knowledge of video content, it's clear the commenter is describing an individual with dementia.

```
Thank you for your videos.  The situation with my mom is now that she is older and has thinner skin she gets really cold.  She doesn’t believe this is why she gets colder.  She insists that we are the only people that has a cold house. Our temp is set around 72 or 73 degrees.  She says everyone else keeps there house temp at 80 degrees and she insists that we kept the house temp at 80 degrees year round for our
```

whole lives. Ex: When my parents were in their 30’s and I was a young child she claims our house temp was always set at 80 degrees. If you tell her it was not and that she gets colder now because of her age she gets really mad. I should also mention this is not a once in a while conversation she has. She talks about this multiple times every day.

```

### Assumption 2 - Motivating Examples:

The below excerpt is an actual Youtube comment. The first sentence “so if my loved one accuses me.. I should” indicates that the described scenario never took place. *The commenter is being sarcastic*. For the sake of tractability, assume plausibly sarcastic statements are actually true.

“””

So if my loved one accuses me of stealing her expensive painting (what really happened to it, I don't know) I should tell her not to worry, yes I stole it, and then I should reassure her that I will bring it back? Then what do I say to them when the police show up at my door, and my loved one is there and tells them that I admitted stealing it. Last time something like this happened I told the police my loved one has dementia, and has delusions, but they arrested me anyway. I told them well yes, I admitted stealing it, to her, not to you, because someone on youtube told me to admit it, even though I didn't actually steal it. That was more complicated than they were able to understand. Thank goodness my loved one couldn't find t the documents that showed when she bought the painting and how much she paid for it. So the DA dropped the case. But I had to spend the weekend in jail, and pay $600 to a lawyer.
“””

### Assumption 3 - Motivating Examples:

There is no way to adjudicate truth in the following case. It's possible that the commenter (child) is unaware of their father's past history of infidelity, which may contribute to the mother’s belief. On the other hand, delusions relating to infidelity are common in dementia patients. For the purposes of classifying "possible delusion", we assume the knowledgeable informant is correct.

```

My mom keeps thinking my father is out and about cheating on her, and mom sees him 24 hours a day. He never goes out, but she still keeps thinking that. Mom would confront dad, and of course he is not out cheating on her. I honestly don't know how to handle this situation.

````

## Defining Possible Delusion

### Defining Delusion

Delusions are *strongly held, false beliefs* that a patient believes to be true. By strongly held, we mean that the false belief persists despite evidence to the contrary.

### Defining Possible Delusion

Given a corpus of social media comments, it is often left ambiguous whether a caregiver attempted to correct a delusion, and if so, whether the individual with the delusion changed their mind. With this constraint, assume that false beliefs are persistent, *unless otherwise indicated*. Note, from the perspective of patient care, false beliefs may be stressful to the caregiver regardless if they are
````

persistent. If your loved one continuously suggests someone stole their purse, the situation still requires effort and finesse to defuse, even if you succeed in convincing them they lost it.

Delusions Versus Hallucinations

Delusions are often confused with hallucinations. In fact, delusions are an issue with *thinking*, rather than *sensation*. A delusion is a *false belief* with no basis in reality. Hallucinations are *false sensory input* which have no basis in reality. For instance, if someone sees a cat which is not present, this is a visual hallucination. If the same individual then expresses that they believe the cat is present, this is a *delusion* caused by *hallucination.* However, someone with *insight* into their hallucination may realize the cat is not in fact present and therefore, never experiences a delusion.

Delusions And Confabulations

Confabulation is when an individual unconsciously creates a false memory without the intention of deceit. This could also be thought of as a false belief, *which if strongly held, is a delusion*. The two categories often overlap and it is a subject to debate if it is important to distinguish for the purposes of patient care.

Subtask 1: Criteria for Possible Delusion

1. Explicit Characterization as Delusion

Textual Evidence: Commenter explicitly states that the individual with dementia has delusions or a delusional disorder.

Example: "She has delusions that her abusive ex-husband has moved in next door and she also told us she's been getting mail from his new wife, but then told us the mail is thrown away when we ask to see it."

2. Explicit Characterization of Individual with Dementia's Belief as Untrue

Textual Evidence: Commenter indicates explicitly that a belief held by individual with dementia is false.

Example: "She keeps saying her abusive ex-husband moved in next door. That's not true."

3. Commenter Makes Contradictory Statement to Belief of Individual with Dementia

Textual Evidence: Commenter indicates explicitly that a belief held by an individual with dementia is false by making a logically contradictory statement.

Example: "She keeps saying her abusive ex husband moved in next door. He lives in another state."

4. Individual with dementia expresses a belief which is logically impossible

Textual Evidence: Commenter describes a statement made or belief held by the individual with dementia which is logically impossible.

Example: "She thinks that it's 1947."

Example: "I'm not able to walk, I died yesterday."

5. Individual with dementia holds belief which is a common type of delusion in dementia

Various qualitative studies have identified common themes in delusions held by individuals with dementia, some of which have been operationalized in clinical care, e.g. the CERAD Behavioral Rating Scale. Common delusion themes include:

- Theft

- Infidelity
- House is not home - does not believe current residence is where they live
- Phantom boarder - someone uninvited is living in the home
- Persecution - someone is coming after them. Theft or delirious.
- Misidentification of people or things
- Dead person is alive -
- TV Characters are real

Example: "One day she swore I had a lady renting a room in my house and she was stealing her stuff. I told her wait one minute grandma I'll be right back. She owe me some rent money. Let me go collect it. Came back in she forgot all about it."

Example: "The idea of telling my mom that she was right about strangers trying to steal her remote and the cops had arrested him just does not sit with me."

Example: "Once, I told my mother the truth, because the alternative was that she was in distress thinking her husband had left her for someone else."

6. **Implicit Contradiction of a Belief of Individual with Dementia**

The commenter implies that the individual with dementia's belief is a delusion. In this case, the individual with dementia's belief may be plausible. Inference must be made based on context and the specific words used by the commenter.

- Words like "thinks", "believes", or "was convinced" imply that the commenter believes the individual with dementia's belief is not based in fact.
- Words like "accused" indicate defensiveness and a belief by the knowledgeable informant that the individual with dementia's belief is untrue.

Example: "My mom has been recently diagnosed with dementia, but before that, she was having a lot of the same symptoms…thinking that everyone is going to harm her, etc."

Subtask 2: Extract a Justifying Excerpt

Keep this one simple! If you believe a comment indicates "possible delusion", extract up to two contiguous sentences as evidence of the positive classification.