Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success)

Chantal Shaib[1] **Millicent L. Li**[1] **Sebastian Joseph**[2]
Iain J. Marshall[3] **Junyi Jessy Li**[2] **Byron C. Wallace**[1]
[1]Northeastern University, [2]The University of Texas at Austin,
[3]King's College London
{shaib.c, li.mil, b.wallace}@northeastern.edu
iain.marshall@kcl.ac.uk
{sebaj, jessy}@utexas.edu

Abstract

Large language models, particularly GPT-3, are able to produce high quality summaries of general domain news articles in few- and zero-shot settings. However, it is unclear if such models are similarly capable in more specialized, high-stakes domains such as biomedicine. In this paper, we enlist domain experts (individuals with medical training) to evaluate summaries of biomedical articles generated by GPT-3, given zero supervision. We consider both single- and multi-document settings. In the former, GPT-3 is tasked with generating regular and plain-language summaries of articles describing randomized controlled trials; in the latter, we assess the degree to which GPT-3 is able to *synthesize* evidence reported across a collection of articles. We design an annotation scheme for evaluating model outputs, with an emphasis on assessing the factual accuracy of generated summaries. We find that while GPT-3 is able to summarize and simplify single biomedical articles faithfully, it struggles to provide accurate aggregations of findings over multiple documents. We release all data and annotations used in this work.[1]

1 Introduction

Large language models have been shown to be capable of producing high-quality and reasonably accurate summaries in *zero-shot* settings (Goyal et al., 2022; Liang et al., 2022), with GPT-3 besting fully supervised models in generic news summarization, according to human judgments (Goyal et al., 2022). In this work we evaluate if such models are similarly able to summarize medical literature, a high-stakes domain that demands factual accuracy.

Specifically, we use the newest iteration of GPT-3 (`text-davinci-003`; `GPT3-D3` from here) to generate summaries of (a) individual articles describing individual randomized controlled trials (RCTs) evaluating the efficacy of interventions, and, (b) collections of such articles that describe several trials addressing the same underlying clinical question (e.g., evaluating the same medication). These constitute single- and multi-document summarization tasks, respectively. In the single-document case, we also evaluate the ability of `GPT3-D3` to summarize in *plain language*. We enlist domain experts (with medical training) to annotate model outputs, and seek to address the following questions.

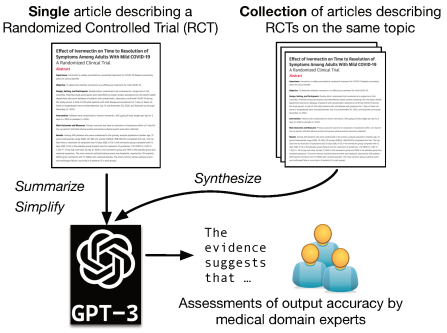

Figure 1: We enlist domain experts to evaluate the factual accuracy of summaries and simplifications of medical articles describing clinical trials. We consider both single- and multi-document settings.

RQ1 Does `GPT3-D3` produce *faithful* summaries of medical articles?

RQ2 Can `GPT3-D3` accurately *simplify* while also summarizing such texts?

RQ3 Can `GPT3-D3` *synthesize*—aggregate the findings presented in—multiple input articles in a way that accurately reflects the totality of the evidence?

RQ4 What sort of factual mistakes does `GPT3-D3` make when performing these tasks (if any), and what are the risks implied by such errors?

Overall, we find that `GPT3-D3` performs single-document summarization and simplification with reasonably good accuracy. However, it is less able to accurately synthesize evidence reported in *collections* of trials (in the multi-document case). We

[1]https://github.com/cshaib/summarizing-medical-evidence

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1387–1407
July 9-14, 2023

release all model outputs and accompanying annotations to facilitate additional work on this topic.

2 Single Document Summarization

Data We sample 100 articles describing randomized control trials (RCTs) indexed in the Trialstreamer database (Marshall et al., 2020), which also provides automatically extracted "key results"[2] alongside titles and abstracts. We search for trials published after November 28 2022, following the release date of GPT3-D3, to ensure the model has not seen any of the studies during pre-training.

Experimental Setup Using the RCT data described above, we evaluate the ability of GPT3-D3 to faithfully summarize and simplify biomedical texts in a zero-shot setting. We also compare GPT3-D3 summaries to summaries generated using Flan-T5 (Wei et al., 2021), but qualitatively find that GPT3-D3 summaries are much higher quality. We provide results of this comparison in Appendix F.3. Specifically, we prompt GPT3-D3 to separately produce: (i) a technical summary, and, (ii) a plain language summary (August et al., 2022). See Appendix C for all prompts.

Study Design We designed an evaluation scheme that captures the sensitivity of medical information. To assess factuality, we collect annotations about omissions and errors with respect to main results, and key components of the trials including populations, interventions, and outcomes ("PICO" elements; Richardson et al. 1995). Where appropriate, we ask annotators to highlight spans of generated text that are inconsistent with the input—these might be "new" concepts introduced or spans that directly contradict the input. To gauge overall linguistic quality, we solicit assessments regarding the fluency and usefulness of a summary on a Likert scale (1932). We include additional questions about the simplification of technical terms for the plain language summaries. We provide a complete taxonomy of the survey in Appendix H.

Annotations We recruited 3 domain experts with medical training on the Upwork platform,[3] and task them each with annotating 100 samples. In total, we collect 300 annotations (3 annotations per sample). We use Label Studio[4] as our interface.

[2]Extracted sentence communicating the main findings.

[3]https://www.upwork.com

[4]https://labelstud.io/

3 Multiple Document Summarization and Evidence Synthesis

Data For multi-document summarization, we download meta-analyses from the Cochrane Library (these are reviews of medical evidence, usually RCTs).[5] Our final sample contains 50 multi-document studies comprising meta-review titles, reference abstracts (inputs), and target conclusions (target summaries) written by domain experts, 10 of which were published post-GPT3-D3 release. [6]

Experimental Setup Because inputs comprise multiple abstracts, these (together with generated tokens) often exceed the token capacity of GPT3-D3. In our dataset, about 41% of the samples exceeded this upper-bound. We report information about our data, including average length, in Appendix B. To address the upper-bound problem, we adopt a simple two-phase strategy for multi-document summarization. First, we generate independent summaries for each abstract, using the single-document summarization prompt described in Section 2. Then, we include all the generated single-document summaries in our multi-document synthesis prompt[7] (examples in Appendix C).

Study Design Our evaluation rubric asks for assessments of generated outputs as compared to: (a) inputs, and, (b) target summaries. Specifically, we ask if generated summaries are supported by the *summaries* provided as inputs in the multi-document case, and to what extent they agree with target (reference) summaries. We also ask annotators to highlight spans of text in generated outputs that disagree with paired target summaries. We reproduce the full rubric in Appendix H.

With respect to annotators, we use the same procedure described in Section 2; we recruited 3 new medical experts and tasked them each with annotating 50 samples, for a total of 150 annotations.

[5]https://www.cochranelibrary.com/

[6]At the time of retrieval we were only able to extract 18 samples post-GPT3-D3 release. We excluded any updates (meta-analyses with ≤ 1 reference abstract). There was no discernible difference in the performance, however, more data is needed to evaluate this effect

[7]Note that we have yet to see prior work systematically investigate a strategy for zero-shot multi-document summarization; due to the prompt-sensitive nature of LLMs (Liang et al., 2022), we do not guarantee that we obtained the best prompt despite fairly extensive trials.

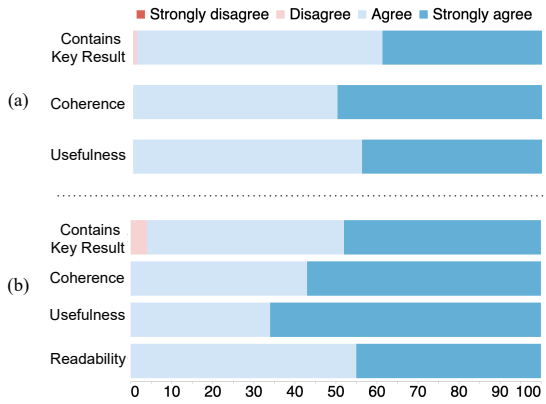

Figure 2: Average scores for assessing overall faithfulness, coherence, and usefulness of generated (a) regular summaries and (b) simplified summaries. GPT3-D3 produces high-quality regular and simplified summaries.

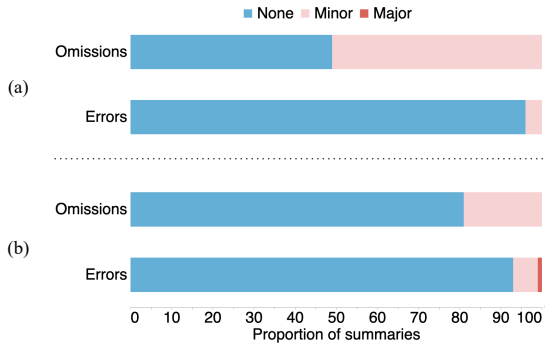

Figure 3: Average number of errors and omissions made in the generated (a) regular and (b) simplified summaries. Most mistakes made in both cases are minor, and omissions are more frequent than errors.

4 Results

RQ1: Does GPT3-D3 produce faithful summaries of medical articles? In the single document setting, we find that GPT3-D3 generates summaries of biomedical abstracts that are fairly high-quality. Figure 2 (a) shows that annotators rated a majority of the summaries as being coherent, useful, and capturing "key results".

When GPT3-D3 does err, it tends to make minor mistakes or omit details. The latter is more common than the former, as shown in Figure 3 (a).

RQ2: Can GPT3-D3 accurately simplify while summarizing medical texts? Shown in Figure 2 (b), GPT3-D3 produces simplified summaries that are similarly deemed to be coherent and useful, and which appear to contain key results. Simplified outputs are scored highly in terms of readability, indicating that these summaries would be understood by someone without medical training.

In comparison to the technical summaries, Figure 3 (b) shows that there are fewer omissions but a slightly higher amount of errors. These may be problematic, but — importantly — some omissions are expected in a simplified summary, as certain details that are important for an accurate summary for a technical audience may not be necessary to convey key information to a more general audience.

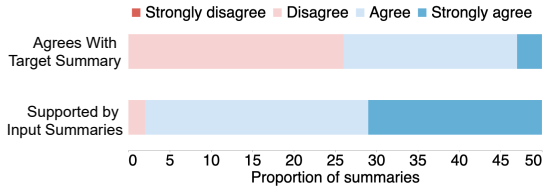

Figure 4: Proportion of summaries that reflect the target summary and are supported by the input summaries in the multi-document setting. While most summaries follow from the input, less than half are rated as agreeing with the target summary.

RQ3: Can GPT3-D3 *synthesize* findings presented in multiple input articles in a way that accurately reflects the totality of the evidence? We now evaluate GPT3-D3's performance on multi-document summarization, i.e., its ability to synthesize evidence (Wang et al., 2022). Figure 4 shows that most summaries generated by GPT3-D3 in this setting are supported by the inputs. This is consistent with our findings in **RQ1**: GPT3-D3 is able to summarize faithfully with respect to given input. However, we find that generated summaries do not consistently agree with the target summaries. Indeed, Figure 4 shows that generated summaries disagree with the targets in over half of cases. This discrepancy suggests that human-written summaries in the biomedical domain require a level of synthesis that is not captured by GPT3-D3 .

RQ4: What sort of factual mistakes does GPT3-D3 make and what are the risks? In RQ1, we reported that GPT3-D3 sometimes omits key information. Figure 5 characterizes the types of omissions and errors made, with respect to PICO elements. GPT3-D3 tends to underspecify elements in the summary more often than generating inaccuracies. Appendix F provides further details regarding underspecification. In the simplification task, GPT3-D3 capably simplifies most technical terms in the generated output (Figure 6).

Regarding RQ3, we showed that there are often discrepancies between generated and target summaries, despite the former being supported by the inputs. Human-written summaries of trials may be

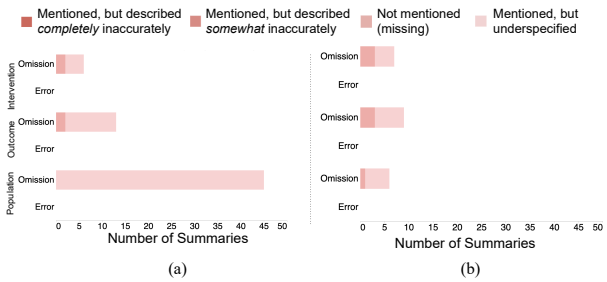

Figure 5: Granular omissions and errors annotated in (a) technical and (b) simplified summaries. Most omissions come from underspecifying key components.

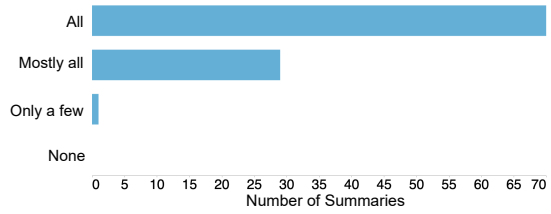

Figure 6: In the simplification case, the model usually replaces complex terms with simpler ones.

more cautious in their conclusions. We measure the evidence strength and direction of both the target and generated summaries, and find that GPT3-D3 tends to recommend marginal or substantive beneficial effects regarding interventions in the majority of the summaries (Figure 7).

Overall, we find that GPT3-D3 copies frequently from inputs. This results in summaries that are often faithful to the input. It may also be one reason that summaries tend to have more omissions (rather than errors) in the single document case, and it may also explain how summaries in the multi-document case often disagree with the reference synopsis while also being supported by (some subset of) the inputs. We calculate the degree of overlap and similarity between inputs and generated summaries from GPT3-D3 for both single-document and multi-document summarization at the sentence level (Figure 8). GPT3-D3 often copies sentences verbatim. In other cases, it changes phrasings but only very slightly (see Appendix F for examples).

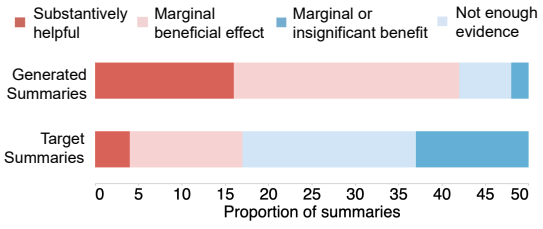

Figure 7: Proportion of summaries that are reported as beneficial in the generated summaries and the target summaries. The generated summaries tend to report beneficial effects in most of the summaries.

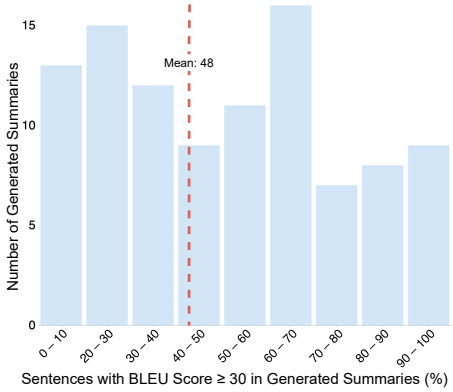

Figure 8: Percentage of sentences in the generated summaries with a BLEU score of 30 or higher, which indicates high similarity.

Further, Figure 8 shows how many sentences in each summary have a BLEU score of ≥ 30; which indicates the sentences are highly aligned. Over 70% of the summaries have at least a quarter of the sentences copied from the input. Appendix F shows some examples of highly similar summaries and sentence pairs.

5 Related Work

More broadly in summarization, several efforts have called for increased emphasis on human (rather than automated) evaluation of generated texts, increased deployment of human-centered systems for text generation evaluation (Khashabi et al., 2021), and greater focus on building benchmarks that incorporate human preferences (Liang et al., 2022; Fabbri et al., 2021). And indeed, Goyal et al. (2022) find that summaries produced by GPT3-D3 are often preferred by humans over alternative model outputs even when automated metrics disagree. Such findings have motivated the manual analysis we conduct for this work. As far as we know, there has not been any work that assess the degree to which GPT-3 is proficient at summarizing biomedical and clinical data in both single-document and multi-document cases.

Our analysis of summarization in the biomedical space complements recent work analyzing the question answering capabilities of such models in this domain (Singhal et al., 2022; Liévin et al., 2022) and the degree to which they encode medical knowledge implicitly (Sung et al., 2021). Other work has considered using summarization

of biomedical texts as assistive tools for reading (August et al., 2022).

6 Conclusions

We evaluate the ability of GPT3-D3 to faithfully summarize and simplify medical literature. The expert annotations we collect indicate that GPT3-D3 performs single-document tasks quite well, but struggles with multi-document summarization. This highlights the ability to aggregate across documents as a direction for future work. We release all data and annotations to facilitate such work in the medical space going forward.

Limitations

This evaluation focussed on expert manual assessments of model outputs and their factual accuracy. Domain expertise (in medicine) was invaluable for this task, but is also expensive and therefore limited the scale of our evaluation. Consequently, all findings are derived over a modest sample (100s) of triple-annotated instances.

Another limitation here is that we have considered only articles describing *randomized control trials (RCTs)*. We focused on such articles because RCTs are the most reliable means of assessing medical interventions, and therefore inform the practice of evidence-based medicine; summarizing such articles is therefore critical to help physicians stay on top of the evidence. Moreover, RCTs provide a natural grounding with respect to factuality, given that all such trials will investigate the relative efficacy of an intervention for a particular condition (i.e., on a specific population of patients) and with respect to an outcome of interest. That said, this is restrictive by design, and our analysis has therefore excluded large swaths of other types of medical texts.

Ethical Considerations

In Appendix D, we note the costs of hiring domain experts for annotation.

Large language models (such as GPT3-D3) have been shown capable of generating concise and fluent summaries. But these often contain factual inaccuracies. This poses unique risks in the domain of medicine, where inaccurate summaries of published evidence have the potential to (mis-)inform patient care. This work has attempted to empirically assess the tendency of models to introduce inaccuracies into summaries of medical literature by enlisting domain experts to identify and characterize omissions and errors in model generated summaries. Understanding such issues is a first step toward designing methods to mitigate them.

While we found that GPT3-D3 appears to produce summaries of single biomedical article abstracts that are reasonably factual, relying on such outputs still poses risks, and even in this setting we would caution against trusting model outputs without further verification at present. Moreover, we found that in the multi-document case—i.e., on the task of synthesizing evidence reported across multiple clinical trials—GPT3-D3 struggles to provide synopses that agree with reference (expert written) summaries. In sum, despite their ability to produce consistently plausible outputs, our view is that summaries of medical literature produced by LLMs should not yet be used to directly inform care given the risks of factual inaccuracies. More research is needed to better characterize the kinds of mistakes such models make, and ultimately to mitigate them.

Acknowledgements

This research was partially supported by National Science Foundation (NSF) grants IIS-2145479 and RI-2211954, and by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant 2R01LM012086.

References

Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2022. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Transactions on Computer-Human Interaction*.

Alexander Richard Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *ArXiv*, abs/2209.12356.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. Genie: Toward reproducible and standardized human evaluation for text generation. In *Conference on Empirical Methods in Natural Language Processing*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–A13.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Appendix

A Model details

We use the following parameters to prompt `GPT3-D3`: temperature = 0.7, top-p = 1.0, frequency penalty = 0.0, presence penalty = 0.0. We set our maximum token length to 1000 to avoid artificially introducing any omission errors.

B Dataset statistics

We provide some basic information about the dataset in Table 2. Because we used `GPT3-D3`, we do not have a clear idea about how the tokenization is done. To be as transparent as possible, however, we still provide the number of tokens when tokenized with SpaCy[8]. Since we use `GPT3-D3`, we opt to use a tokenization scheme that focuses mainly on general English (so we did not use a specialized tokenizer for biomedical texts to replicate as similar a tokenization as possible).

C Prompts

For single-document summarization, we follow prior work to select our prompts. From (Goyal et al., 2022; August et al., 2022), we use the following prompts for the technical summary and the plain language summary:

- Summarize the above.
- My fifth grader asked me what this passage means: “““ [TEXT TO SIMPLIFY] ””” I rephrased it for him, in plain language a fifth grader can understand.

To our knowledge, there is no prior work investigating prompt constructions for multi-document summarization generally (or evidence synthesis specifically). Table 1reproduces prompts we considered for this, but we ultimately used:

- “““ [GENERATED INPUT SUMMARIES] ””” What does the above evidence conclude about “““ [TITLE] ”””?

Figure 9 shows an example of the input structure and prompts we provide to `GPT3-D3` in the multi-document setting. For the few-shot setting, we evaluate using up to 5 examples in context. Figure 10 shows the input structure for this setting in the second phase.

[8] https://spacy.io/

Prompts:
Write a meta-analysis based on the above evidence.
Summarize the above evidence.
Synthesize the above.

Table 1: Examples of prompts tried for multi-document summarization.

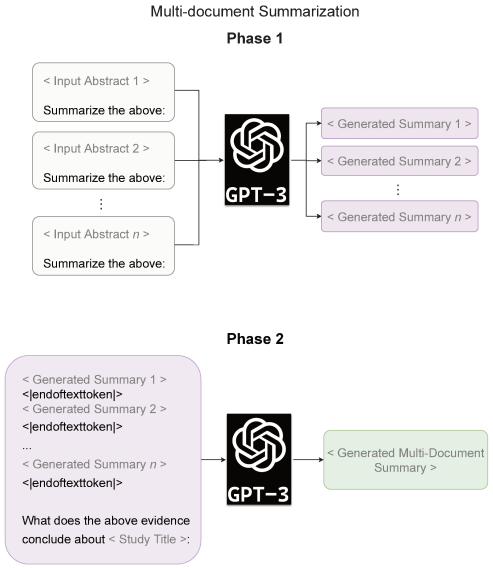

Figure 9: Input structure and prompts for the multi-document setting.

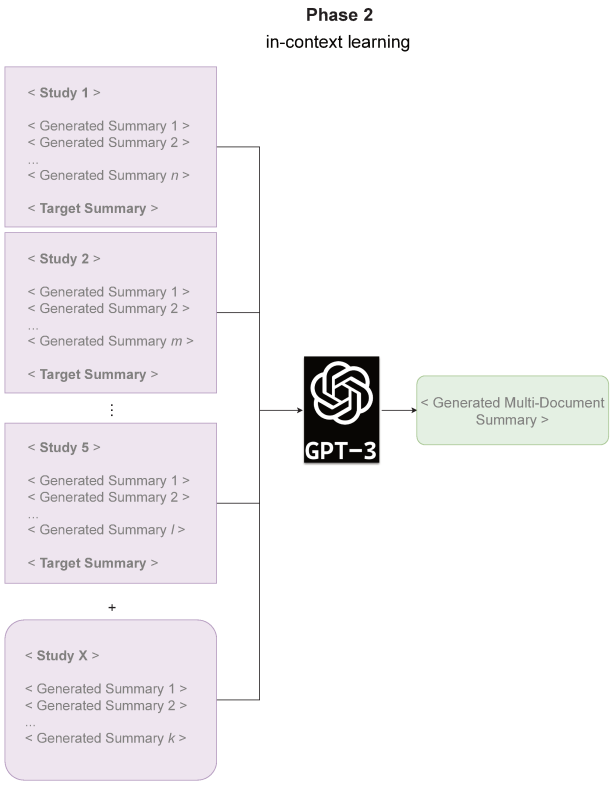

Figure 10: Adding in-context learning examples in the second step in multi-document summarization in the few-shot settings.

D Annotation details

We calculate the inter-annotator agreement score (Cohen's kappa), which averaged 0.59 amongst all annotators.

We also transparently reveal the cost of annotating on Upwork. The total cost of hiring 3 workers on Upwork was a little more than $3,700 USD. Because annotations on a more specialized platform cost significantly more, we hired fewer annotators than one would hire on generic crowdworking websites.

Since each Upworker requested different payment amounts (which is the nature of the platform), we provide the averages per hour for the work. For the single-document case, each annotation took on average 15-20 minutes per sample, and with 100 samples, the upper-bound was 33.3 hours for the entire task per annotator. For the multi-document case, each annotation took on average 10-15 minutes per sample, and with 50 samples, the upper-bound was 12.5 hours for the entire task per annotator. Both tasks had three annotators annotating each.

E Survey details

For each data point (and for each question in the interface), the annotator first evaluates the standard summary and then evaluates the plain language summary, before completing the survey in its entirety. We reproduce our survey questions and the corresponding answer options. These include the evaluation categories that we care about: For **standard (technical) summaries**, we focus on factuality, linguistic quality, and holistic evaluation; For **plain language summaries**, we include an additional section on readability because the purpose of these is to simplify technical language such that a layperson might understand the summary. We provide details regarding the structures of the surveys we used and our rationales behind their construction below.

E.1 Single-document summarization

In the single-document summarization case, the inputs comprise **study abstracts**, **titles**, and we also show to the user **key results**, which were automatically extracted (Marshall et al., 2020). (We do not have reference summaries for these examples.) The goal of expert evaluation was to quantify the extent to which `GPT3-D3` accurately summarizes these article inputs. We reiterate that we consider **two** different types of summarization strategies: standard (technical) summarization and plain-language summarization. We reproduce the questions asked for these summary types below, which vary only slightly in their focus.

Factuality Many of our questions chosen in our taxonomy revolve around factuality since factual accuracy is extremely in domain-specific work.

1. The model summary accurately conveys the key results in the input. Given the model summary, we seek to evaluate whether the key results that are automatically extracted are reflected in the output. This is a matter of degree, so we solicit assessments rated on a Likert scale.

2. Highlight sentences in the model summary (if any) that directly contradict the input (highlight model summary on the right). We collect additional annotations on which portions of the model summary contradict the input. We did not further analyze these highlights here, but do release them as part of the data collected.

3. Highlight any concepts that are new in the model summary that don't appear in the input

Type of statistic	Single-document	Multi-document
Average number of tokens per input (all)	293.06	1451.68
Average number of tokens per input (abstract(s) only)	293.06	1353.04
Average number of tokens per input (study title only)	N/A	10.28
Average number of tokens per input (abstract titles only)	N/A	88.36

Table 2: General dataset statistics for reference. Note that in the single-document case, we only use abstracts in our zero-shot generation, so the remaining rows for anything other than abstracts only are labeled "N/A".

(highlight model summary on the right). Here the idea is to allow the annotator to mark "hallucinated" content in outputs (not supported by the input).

4. How are details about the population described in the summary, relative to the input text? The patient population is a critical component of clinical trials in medicine, and so it is important that summaries accurately describe this element. In particular we ask both whether the population is described (at all), and also the degree to which it is described *accurately*.

5. How are details about the intervention described in the summary, relative to the input text? Another key element of trials is the intervention (e.g., medicine or treatment) being evaluated. Therefore, as for study populations, we collect annotations regarding whether this is captured (and if it is captured accurately).

6. How are details about the outcome (what was measured) described in the summary, relative to the input text? The outcome measured (e.g., mortality) is the final foundational component of trials. As in the preceding two cases, we ask annotators to assess whether this is reported upon faithfully.

7. Are there any omission(s) unrelated to the population, intervention, or outcome? We evaluate whether the model omits any information regarding the key trial elements—population, intervention, and outcome—just described. For more details about types of omissions, refer to section F.2.

8. Are there any errors? We also ask whether there are any errors (in general) in the model summary.

Linguistic quality

9. The model summary is coherent, fluent, and without grammatical errors. This is intended to capture the readability or fluency of the generated output, independent of its veracity.

Holistic evaluation Finally, we ask for a holistic evaluation of the output.

10. The output is a concise, accurate, and potentially useful summary of the input. Continuing with more holistic questions, this is intended to capture the perceived (potential) utility of generated summaries, according to the domain experts we hired as annotators.

In the case of plain summarization, we ask the annotator to rate whether **10. The simplified text is accurate and would be understandable by a (lay) patient.** This effectively conveys the potential utility of automatically produced lay summaries, because the purpose of these outputs would be make medical evidence more accessible to (inexpert) patients.

11. If there was anything not elaborated or covered, feel free to leave a comment in the box. We conclude with an open-ended text box to collect notes or thoughts not otherwise captured.

Readability For **plain language summaries**, we include a section on readability, given the focus on making evidence more digestible in this case.

12. The simplified model text is less technical and more approachable, thus making it easier to understand. This question measures the degree to which the annotator judges the model to have successfully simplified the text.

13. Technical terms in the input are being substituted with simpler language in the simplified model text. This is a more focussed question regarding simplification to quantify whether the model consistently swaps jargon terms for more accessible language.

E.2 Multi-document summarization

The inputs in the multi-document case comprises collections of articles describing trials, and the targets are syntheses of these (which put together

the findings they report). We sampled these meta-reviews from previously conducted evidence syntheses, and so in this case we have target summaries, which we provide to the annotator. We not consider simplification in the multi-document setting.

Factuality We again focus on factuality of model outputs.

1. Highlight any spans in the generated summary that disagree with the target summary. We ask for annotators to mark any explicit contradictions featured in the generated output.

2. The generated summary is supported by putting together the given summaries of the individual articles. The core of multi-document summarization is the piecing together of multiple documents into a coherent summary that accurately reflects the inputs in aggregate. This question is intended to measure the degree to which the model does so.

3. The generated summary agrees with the target summary. Because we have reference (target) summaries in this case, we directly ask whether and to what degree the model generated synopsis seems to agree with this.

4. Rate the degree to which the *generated* summary shows the extent that there is evidence supporting the effectiveness of the intervention(s) of interest (as indicated in the studies). The *generated* summary suggests... Here we aim to assess whether the model output implies that the intervention studied in the constituent trials is supported by the findings reported within them.

5. Rate the degree to which the *target* summary shows the extent that there is evidence supporting the effectiveness of the intervention(s) of interest (as indicated in the studies). The *target* summary suggests... Similarly, we ask whether the reference summary implies that the intervention in question is effective.

Holistic evaluation As above we seek to elicit an overall impression of summary accuracy and quality.

6. If there was anything not elaborated or covered, feel free to leave a comment in the box. Much like for single-document summarization, the survey provides an additional box for annotators to give information about the specific data point that was asked.

F Additional evaluation

F.1 Few-shot

Few-shot We experimented briefly with *few*-shot prompting (Appendix G), but qualitatively this did not seem to outperform zero-shot summarization, hence our focus on evaluating the latter.

For few-shot generation, we insert in-context training examples after the first summarization phase by concatenating the summaries and the target conclusions of inputs (see Appendix C). We evaluate using up to 5 shots.

F.2 Underspecified elements

Table 3 and Table 4 show the additional options selected when an element (e.g., population) was marked as "underspecified" in the survey for the technical and simplified cases, respectively.

There can be many reasons why an element could be marked underspecified. Because we try to remove as much ambiguity as possible, we opt to identify the reasons under each category (*Population*, *Intervention*, *Outcome*) the specific reasoning. The questions we ask in both the regular and plain summarization case are both different because of the audience we address in either case. In the regular summarization case, the reader is intended to be a domain expert; in the plain summarization case, the reader is intended to be laymen, and so we alter the types of questions we ask as a result.

We find that plain summaries (Table 4) have fewer errors than that of regular summaries (Table 3), whereas regular summaries have a higher number of specific omissions. However, plain summaries seem to have more omissions in areas outside of the scope of what we identify as salient omissions. We can hypothesize that given more complex language, it could be that annotators can more easily identify salient information in the text. On the other hand, there are nuances in regular summaries that cannot be extrapolated via plain summarization prompts, and instead we must use regular summaries to gather more critical information (in addition to the fact that the questions asked in the plain summarization case tends to be simpler). Although, with regular summaries, summarizing on a deeper level may result in using more convoluted language. Nonetheless, each type of prompt (regular and plain) seem to be well-suited for the task at hand; what matters is the context in

Type of Error	Number of Articles
Population	
Omits demographic information	0
Omits sample size	41
Other	1
Intervention	
Does not describe comparator intervention	2
Omits dosage or other important detail about administration	1
Other	0
Outcome	
Omits description of specific measurements of high-level outcomes	4
Omits one or more of multiple outcomes	8
Other	0

Table 3: Types of errors and the number of articles with the corresponding error, for regular summarized articles.

Type of Error	Number of Articles
Population	
Missing completely	1
Missing key details (patients vs patients with depression)	2
Inaccurate	0
Other	1
Intervention	
Missing completely	1
Missing comparator	2
Inaccurate	0
Other	2
Outcome	
Missing completely	0
Missing part outcomes	3
Missing key details that would be important for a lay person to know	1
Inaccurate	0
Other	0

Table 4: Types of errors and the number of articles with the corresponding error, for plain summarized articles.

which the prompt is used, and what information is needed for the user.

F.3 Flan-T5

We compared GPT-3 zero-shot results to Flan-T5 (Wei et al., 2021). We find that Flan-T5 produces substantially shorter summaries (2-3 sentences on average). We provide examples of generated summaries in Figure 11. Qualitatively, these seemed far worse than GPT-3 generated outputs, so we did not evaluate these further in this work.

F.4 ROUGE scores

ROUGE-1	ROUGE-2	ROUGE-L
0.27	0.06	0.16

Table 5: ROUGE scores on **multi-document** biomedical summaries using GPT3-D3

We provide the standard automatic metric of ROUGE (Lin, 2004) to analyze multi-document summarization. We do not have ROUGE scores for single-document summarization since we lack ground truth data. However, the focus of this work is on the capability of GPT3-D3 to faithfully summarize biomedical literature (i.e., to generate accurate summaries); human experts remain the best judges of factuality. Noting this and prior work by Goyal et al. (2022) make ROUGE scores (and other automatic metrics) rather unreliable to judge the capabilities of these large language models on summarization.

F.5 Similarity

We provide additional examples of sentences and summaries with high similarity to the input abstract.

G Examples of generated summaries

We include examples of generated summaries we annotated, both standard summaries and plain language in the single and multi-document case (Table 14, 13).

We also provide examples of few-shot generations along with the zero-shot and target summaries for comparison (Figure 15). Note that the few-shot examples reflect the same evidence strength and recommendation as the zero-shot examples, thus

Sentence from Abstracts	Sentence from Generated Summary	BLEU
These findings suggest that access to care and differences in treatment may be responsible for racial disparities in colorectal cancer.	These findings suggest that access to care and differences in treatment may be responsible for racial disparities in colorectal cancer.	100
After corrections for multiple comparisons, only PFC effects on praise and emotion strategies at post-treatment, and praise and withdrawn/depressed behavior at follow-up, maintained.	After corrections for multiple comparisons, only PFC effects on praise and emotion strategies at post-treatment, and praise and withdrawn/depressed behavior at follow-up, were maintained.	91.93
AIM To assess the safety and efficacy of hybrid closed-loop (HCL) insulin delivery 24/7 versus only evening and night (E/N), and on extended 24/7 use, in free-living children with type 1 diabetes.	This study aimed to assess the safety and efficacy of hybrid closed-loop (HCL) insulin delivery 24/7 versus only evening and night (E/N), and on extended 24/7 use, in free-living children with type 1 diabetes.	91.20
We find that protocol compliance, as measured by correlations between e-cigarette use measures and cotinine levels, was only achieved in the first week of the study and declined thereafter.	The findings showed that protocol compliance, as measured by correlations between e-cigarette use measures and cotinine levels, was only achieved in the first week of the study and declined thereafter.	90.46
CONCLUSIONS Our findings suggest that the SERT-enriched functional network is dynamically different in ASD during processing of socially relevant stimuli.	The findings suggest that the SERT-enriched functional network is dynamically different in ASD during processing of socially relevant stimuli.	89.96

Table 6: Examples of highly extractive sentence pairs found from generated summaries for single-document summarization.

Sentence from Abstracts	Sentence from Generated Summary	BLEU
CONCLUSIONS: Drug-induced remission of JIA-U did not persist when adalimumab was withdrawn after 1-2 years of treatment.	However, remission of JIA-U did not persist when adalimumab was withdrawn after 1-2 years of treatment.	84.80
CONCLUSION: This study suggests that increasing the dose of inhaled steroids at the onset of an exacerbation of asthma is ineffective and should not be included in asthma self management plans.	The evidence suggests that increasing the dose of inhaled corticosteroids at the onset of an exacerbation of asthma is ineffective and should not be included in asthma self management plans.	79.19
RESULTS: Following maternal betamethasone administration (day 2), fetal heart rate variation was reduced by 19% and fetal body and breathing movements by 49% and 85%, respectively.	Dexamethasone had a greater beneficial effect, reducing fetal heart rate variation by 19% and fetal body and breathing movements by 49% and 85%, respectively.	56.71
OBJECTIVE: This study aimed to investigate the effect of endometrial injury using Pipelle catheter in the follicular phase (cycle day 5, 6, or 7) of the stimulation cycle on pregnancy rates in patients undergoing intrauterine insemination.	The evidence suggests that endometrial injury using a Pipelle catheter in the follicular phase (cycle day 5, 6, or 7) of the stimulation cycle may improve pregnancy rates in women undergoing intrauterine insemination (IUI).	56.22
CONCLUSION: Based on these results, it is suggested that VAC has advantages when compared to the Bogota bag as a temporary closure method in the management of abdominal compartment syndrome.	Furthermore, the VAC system has advantages compared to the Bogota bag as a temporary closure method in the management of abdominal compartment syndrome.	54.32

Table 7: Examples of highly extractive sentence pairs found from generated summaries for multi-document summarization.

we do not evaluate them at this point.

H Additional figures

Evaluation Category	Question or Statement	Answer Choices
Factuality	The model summary accurately conveys the key results in the input	Strongly disagree; disagree; agree; strongly agree
Factuality	Highlight sentences in the model summary (if any) that directly contradict the input (highlight model summary on the right)	Multiple tokens highlighted
Factuality	Highlight any concepts that are new in the model summary that don't appear in the input (highlight model summary on the right)	Multiple tokens highlighted
Factuality	How are details about the population described in the summary, relative to the input text?	The population is not mentioned (missing) in the model summary; The population is mentioned, but described completely inaccurately; The population is mentioned, but described somewhat inaccurately; The population is mentioned, and described accurately; The population is underspecified; Not applicable (N/A)
Factuality	How are details about the intervention described in the summary, relative to the input text?	The intervention is not mentioned (missing) in the model summary; The intervention is mentioned, but described completely inaccurately; The intervention is mentioned, but described somewhat inaccurately; The intervention is mentioned, and described accurately; The intervention is underspecified; Not applicable (N/A)
Factuality	How are details about the outcome (what was measured) described in the summary, relative to the input text?	The outcome is not mentioned (missing) in the model summary; The outcome is mentioned, but described completely inaccurately; The outcome is mentioned, but described somewhat inaccurately; The outcome is mentioned, and described accurately; The outcome is underspecified; Not applicable (N/A)
Factuality	Are there any omission(s) unrelated to the population, intervention, or outcome?	No omission; Minor omission(s); Major omission(s)
Factuality	Are there any errors?	No errors; Minor error; Major error
Linguistic Quality	The model summary is coherent, fluent, and without grammatical errors	Strongly disagree; disagree; agree; strongly agree
Holistic evaluation	The output is a concise, accurate, and potentially useful summary of the input	Strongly disagree; disagree; agree; strongly agree
Holistic evaluation	If there was anything not elaborated or covered, feel free to leave a comment in the box	Free text

Table 8: Questions used in our survey for annotators to evaluate standard summaries

Evaluation Category	Question or Statement	Answer Choices
Factuality	The simplified model text accurately conveys the key results in the input	Strongly disagree; disagree; agree; strongly agree
Factuality	Highlight sentences in the input (if any) that directly contradict the simplified model text (highlight input on the right)	Multiple tokens highlighted
Factuality	Highlight any concepts that are new in the simplified model text that don't appear in the input (highlight model summary on the right)	Multiple tokens highlighted
Factuality	How are details about the population described in the simplified model text, relative to the input text?	The population is not mentioned (missing) in the simplified model text; The population is mentioned, but described completely inaccurately; The population is mentioned, but described somewhat inaccurately; The population is mentioned, and described accurately; The population is underspecified; Not applicable (N/A)
Factuality	How are details about the intervention described in the simplified model text, relative to the input text?	The intervention is not mentioned (missing) in the simplified model text; The intervention is mentioned, but described completely inaccurately; The intervention is mentioned, but described somewhat inaccurately; The intervention is mentioned, and described accurately; The intervention is underspecified; Not applicable (N/A)
Factuality	How are details about the outcome (what was measured) described in the simplified model text, relative to the input text?	The outcome is not mentioned (missing) in the simplified model text; The outcome is mentioned, but described completely inaccurately; The outcome is mentioned, but described somewhat inaccurately; The outcome is mentioned, and described accurately; The outcome is underspecified; Not applicable (N/A)
Factuality	Are there any omission(s) unrelated to the population, intervention, or outcome?	No omission; Minor omission(s); Major omission(s)
Factuality	Are there any errors?	No errors; Minor error; Major error
Linguistic Quality	The simplified text is coherent, fluent, and without grammatical errors	Strongly disagree; disagree; agree; strongly agree
Holistic evaluation	The simplified text is accurate and would be understandable by a (lay) patient	Strongly disagree; disagree; agree; strongly agree
Holistic evaluation	If there was anything not elaborated or covered, feel free to leave a comment in the box	Free text

Table 9: Questions used in our survey for annotators to evaluate simplified model summaries

Evaluation Category	Question or Statement	Answer Choices
Readability	The simplified model text is less technical and more approachable, thus making it easier to understand.	Strongly disagree; disagree; agree; strongly agree
Readability	Technical terms in the input are being substituted with simpler language in the simplified model text.	None at all; Only a few; Mostly all; All

Table 10: Additional questions used in our survey for annotators to evaluate simplified model summaries

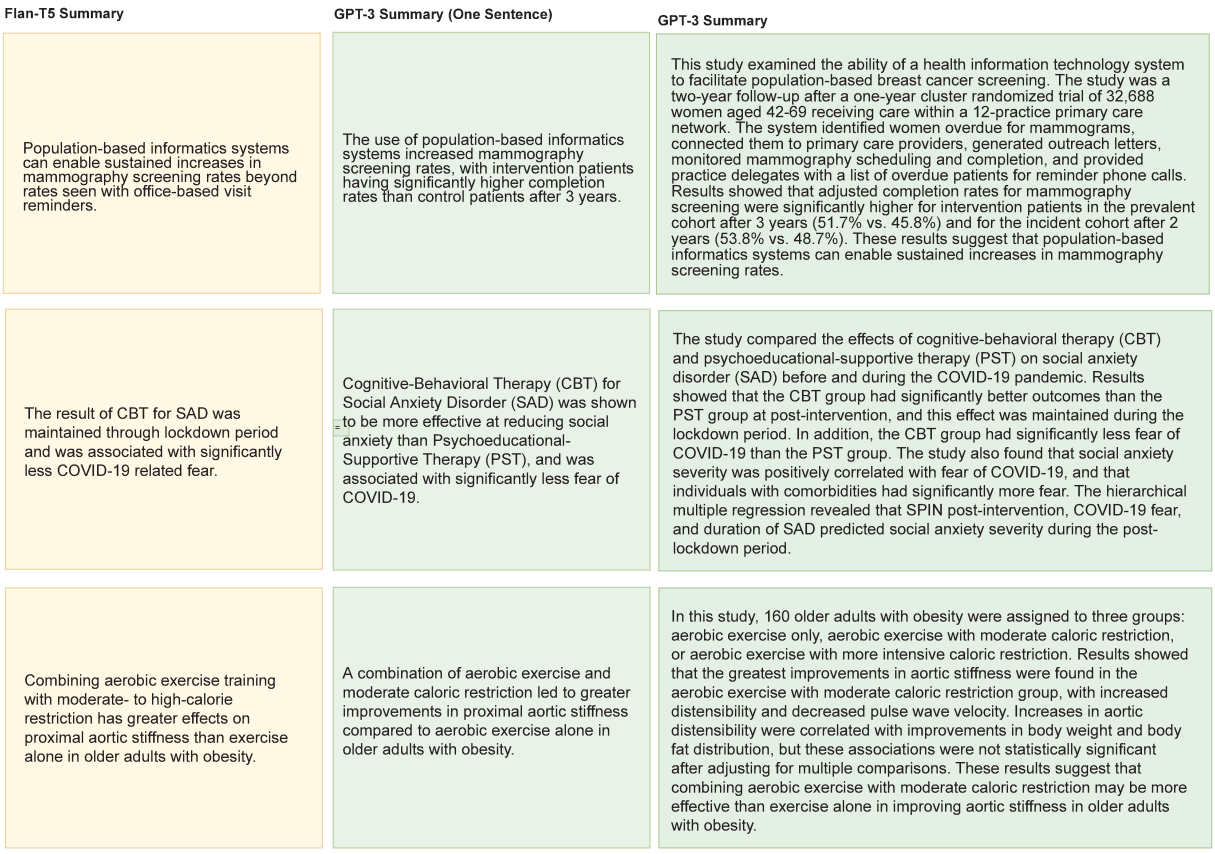

Figure 11: Sample generated summaries from Flan-T5 and GPT-3

Evaluation Category	**Question or Statement**	**Answer Choices**
Factuality	Highlight any spans in the generated summary that disagree with the target summary	Multiple tokens highlighted
Factuality	The generated summary is supported by putting together the given summaries of the individual articles	Strongly disagree; disagree; agree; strongly agree
Factuality	The generated summary agrees with the target summary	Strongly disagree; disagree; agree; strongly agree
Factuality	Rate the degree to which the *generated* summary shows the extent that there is evidence supporting the effectiveness of the intervention(s) of interest (as indicated in the studies). The *generated* summary suggests...	There is not enough evidence to draw any meaningful conclusions; The intervention has a marginal or insignificant comparative benefits; The intervention may have a marginal beneficial effect; The intervention is substantively helpful
Factuality	Rate the degree to which the *target* summary shows the extent that there is evidence supporting the effectiveness of the intervention(s) of interest (as indicated in the studies). The *target* summary suggests...	There is not enough evidence to draw any meaningful conclusions; The intervention has a marginal or insignificant comparative benefits; The intervention may have a marginal beneficial effect; The intervention is substantively helpful
Holistic Evaluation	If there was anything not elaborated or covered, feel free to leave a comment in the box	Free text

Table 11: Questions used in our survey for annotators to evaluate multi-document model summaries

Input Abstract	Generated Summary
The Kinect video game (KVG) has received attention as an intervention method for cerebral palsy (CP). However, evidence remains limited. PURPOSE To investigate the effects of training using Xbox Kinect on lower extremity motor function, balance, and gait in adolescents with spastic diplegia CP. METHODS This study was a pilot randomized controlled trial. Ten participants were randomly allocated to either the KVG training group (n = 5) or the conventional training (CT) group (n = 5). The Selective Control Assessment of the Lower Extremity (SCALE) tool, Pediatric Balance Scale (PBS), and GAITRite were used for the outcome measurements. RESULTS In the comparison between the groups, the KVG group showed significant improvements in all the items in the SCALE (except for right hip abduction) and PBS score as compared with the CT group. CONCLUSIONS KVG training might be an effective intervention for the rehabilitation of adolescents with spastic diplegia CP.	This study investigated the effects of training using Xbox Kinect on lower extremity motor function, balance, and gait in adolescents with spastic diplegia CP. Results showed that the KVG group had significant improvements in all the items in the SCALE (except for right hip abduction) and PBS score compared to the conventional training group. The findings suggest that KVG training may be an effective intervention for the rehabilitation of adolescents with spastic diplegia CP.
OBJECTIVE To evaluate the effect of antrectomy in which resection was started from 2 cm or closer to the pylorus on % excess weight loss (EWL), nausea, vomiting, and complication rates. STUDY DESIGN Comparative study. PLACE AND DURATION OF STUDY Antalya Training and Research Hospital, from April 2018 to December 2018. METHODOLOGY Patients in whom laparoscopic sleeve gastrectomy (LSG)were done starting at a level of 2 cm or closer to pylorus were included in the study. Patients were divided into one of the two groups based on the distance between the pylorus and the resection margin: group 1 having resection ≤10 mm and group 2 at 11-20 mm. Above mentioned parameters were compared in both groups. RESULTS Ninety-two patients were included. Postoperative nausea and vomiting rates were similar in both groups. At the end of the first year, % EWL was 82.9% and 73.5% in groups 1 and 2 (p=0.003). CONCLUSION Starting antrectomy at a distance of 2 cm or less from the pylorus is safe and effective. Starting antrectomy at a distance of 1 cm or less from the pylorus in LSG provides effective weight loss without increasing complications. KEY WORDS Bariatric surgery, Antrectomy, Laparoscopic sleeve gastrectomy, Complications.	This study evaluated the impact of antrectomy in which resection was started from 2 cm or closer to the pylorus on % excess weight loss (EWL), nausea, vomiting, and complication rates. Ninety-two patients were included and divided into two groups based on the distance between the pylorus and the resection margin. Postoperative nausea and vomiting rates were similar in both groups. At the end of the first year, % EWL was significantly higher in group 1 which had resection ≤10 mm compared to group 2 at 11-20 mm (p=0.003). Overall, starting antrectomy at a distance of 2 cm or less from the pylorus in laparoscopic sleeve gastrectomy is safe and effective and provides effective weight loss without increasing complications.
OBJECTIVE To comparatively examine the effectiveness and safety of the combination therapy of traditional Chinese medicine formula Bushen-Shugan granule and psychotherapy (BSSG-P) and Chinese herbal medicine Bushen-Shugan granule (BSSG) alone in the treatment of moderate to serious mood disorder in menopausal women. METHODS In our previous clinical studies, BSSG-P had been proved to be superior to BSSG, psychological treatment, and placebo in improving mild mood disorder in menopausal women. In this study, we analyzed the efficacy of BSSG-P and BSSG in the treatment of moderate to serious mood disorder. Eighty-five eligible participants, who were diagnosed as menopausal women with moderate to serious mood disorder and categorized as kidney deficiency and liver-qi stagnation pattern, were randomly assigned into two groups and treated with BSSG-P or BSSG. They were subjected to an 8-week treatment period and a 4-week follow-up study. The primary outcome instrument was the Greene Climacteric Scale, Self-Rating Depression Scale (SDS), and Self-Rating Anxiety Scale (SAS), respectively. RESULTS When comparing all time points with baseline, both BSSG-P and BSSG markedly decreased the total score of Greene, SDS, and SAS and the score of each dimension, in which BSSG-P exerted superior effect after 8-week treatment and 4-week follow-up (P <0.05). Furthermore, BSSG-P also showed great advantage in reducing the score of Greene, SDS, and SAS for menopausal women with moderate mood disorder at the end of the 8th and 12th week when compared with BSSG (P < 0.05), whereas there was no significant difference between groups at any time point for patients with serious mood disorder (P >0.05). No serious event occurred in both groups, and no significant difference was found between groups in adverse event proportion. CONCLUSIONS BSSG-P was superior to BSSG in improving the physical and psychological symptoms of menopausal women with mood disorder. For patients with moderate mood disorder, BSSG-P showed obvious advantages; however, no superiority was observed for serious mood disorder.	This study examined the effectiveness and safety of a combination therapy of traditional Chinese medicine formula Bushen-Shugan granule and psychotherapy (BSSG-P) and Chinese herbal medicine Bushen-Shugan granule (BSSG) alone in the treatment of moderate to serious mood disorder in menopausal women. The study found that BSSG-P was superior to BSSG in improving the physical and psychological symptoms of menopausal women with mood disorder, especially for those with moderate mood disorder. No serious event occurred in both groups and no significant difference was found between groups in adverse event proportion.

Figure 12: Examples of generated summaries where all sentences have a high similarity scores.

Input:

To investigate the optimal blood pressure (BP) in patients with coronary artery disease (CAD), we conducted subgroup analysis using SPRINT data. The study sample included 1206 participants with CAD (of whom 692 underwent coronary revascularization) and 8127 participants without CAD. Participants were randomized into two groups (systolic BP target of 140 mm Hg vs. 120 mm Hg). The primary outcome was a composite of cardiovascular events. After a median follow-up of 3.9 years, the hazard ratios (HRs) for the primary outcome were 0.65 (95% confidence interval (CI) 0.53-0.79) and 1.05 (95% CI 0.76-1.46) among those in the non-CAD and CAD subgroups, respectively (P value for interaction 0.02). Intensive BP treatment was a protective factor for all-cause death (HR 0.60, 95% CI 0.37-0.96) in the CAD subgroup, compared with standard BP treatment. The HRs (95% CI) for stroke were 3.57 (1.17-10.85) and 1.03 (0.29-3.62) among those in the coronary revascularization and non-revascularization subgroups, respectively (P value for interaction 0.13). For safety events, intensive BP treatment increased the risk of hypotension (HR 2.00, 95% CI 1.06-3.79) and electrolyte abnormalities (HR 2.38, 95% CI 1.25-4.56) in the CAD subgroup, while the risk of serious adverse events did not increase (HR 1.03, 95% CI 0.88-1.20). These results suggest that positive benefits from intensive BP treatment might be attenuated in patients with CAD who are under better secondary prevention. The risk of stroke might increase at the systolic BP target of 120 mm Hg in case of coronary revascularization, although the confidence interval was wide.

Generated Technical Summary:

The study investigated the optimal blood pressure for patients with coronary artery disease (CAD) and found that after a median follow-up of 3.9 years, intensive BP treatment was a protective factor for all-cause death in the CAD subgroup. However, the risk of stroke might increase in patients who underwent coronary revascularization. Intensive BP treatment was also associated with an increased risk of hypotension and electrolyte abnormalities in the CAD subgroup, while the risk of serious adverse events did not increase.

Generated Simplified Summary:

This study looked at people with and without coronary artery disease (CAD), a type of heart disease. They wanted to find out if lowering the blood pressure to a certain level (120 mmHg) was better for their health than having it at a slightly higher level (140 mmHg). They found that for people without CAD, lowering their blood pressure was helpful. But for those with CAD, lowering their blood pressure did not help as much. They also found that people who had had a special surgery to help their CAD (called coronary revascularization) had a higher chance of having a stroke if their blood pressure was lowered too much. In conclusion, it is important to keep a healthy blood pressure, but for people with CAD, it is important to not lower it too much.

Figure 13: An example input and output (technical and simplified summaries) for the single-document summarization task.

Input:

Study title:
Felbamate add-on therapy for drug-resistant focal epilepsy

1. **Abstract title:** Felbamate: a double-blind controlled trial in patients undergoing presurgical evaluation of partial seizures.

This double-blind, placebo-controlled trial studied the efficacy and safety of felbamate as an investigational antiepileptic drug. **[...]** This trial demonstrated that felbamate is an effective and safe treatment for refractory partial-onset seizures.

2. **Abstract title:** Felbamate for partial seizures: results of a controlled clinical trial.

In a double-blind, placebo-controlled trial, 56 patients with partial seizures and a requirement for four or more seizures per month despite concomitant therapeutic blood levels of phenytoin and carbamazepine were given Felbamate. **[...]** Overall, this trial suggests that Felbamate is a safe and effective treatment for severely refractory epilepsy.

3. **Abstract title:** Felbamate: a clinical trial for complex partial seizures.

This randomized, double-blind, three-period crossover study found that felbamate (FBM) was well-tolerated by 28 patients with complex partial **[...]** Although there was no significant reduction in seizure frequency between placebo and FBM periods, the data suggested a strong antiseizure effect of FBM when the lower CBZ levels were taken into account.

Generated Technical Summary:

The evidence suggests that Felbamate is an effective and safe add-on therapy for patients with drug-resistant focal epilepsy. The trial showed that Felbamate was superior to placebo in reducing seizures, with no major adverse events reported. Additionally, the data suggests that Felbamate may have an antiseizure effect when combined with lower levels of carbamazepine.

Target:

In view of the methodological deficiencies, the limited number of included studies and the differences in outcome measures, we have found no reliable evidence to support the use of felbamate as an add-on therapy in people with drug-resistant focal-onset epilepsy. A large-scale, randomised controlled trial conducted over a longer period of time is required to inform clinical practice.

Figure 14: An example input, output, and target for the multi-document summarization task.

Generated Technical Summary (0-shot)	Generated Technical Summary (5-shot)	Target
The evidence suggests that Felbamate is an effective and safe add-on therapy for patients with drug-resistant focal epilepsy. The trial showed that Felbamate was superior to placebo in reducing seizures, with no major adverse events reported. Additionally, the data suggests that Felbamate may have an antiseizure effect when combined with lower levels of carbamazepine.	The evidence from these three studies suggests that felbamate is an effective and safe antiepileptic drug for refractory partial-onset seizures. It is well-tolerated by patients taking concomitant carbamazepine, and may reduce the need for higher doses of this medication. It is also associated with a decrease in seizure frequency, although further studies are needed to confirm this result.	In view of the methodological deficiencies, the limited number of included studies and the differences in outcome measures, we have found no reliable evidence to support the use of felbamate as an add-on therapy in people with drug-resistant focal-onset epilepsy. A large-scale, randomised controlled trial conducted over a longer period of time is required to inform clinical practice.
The evidence suggests that peri-implantation glucocorticoid administration does not provide a significant benefit to assisted reproductive technology cycles when it comes to improving pregnancy and implantation rates. While it may improve the pregnancy rate in patients with positive autoantibodies, it has not been shown to be effective in improving the overall outcome of IVF-ET cycles. Additionally, it has not been shown to be effective in reducing the risk of OHSS.	The evidence suggests that the use of low-dose glucocorticoids as an adjunct to IVF-ET does not have a significant effect on pregnancy or implantation rates, and does not reduce the incidence of ovarian hyperstimulation syndrome. Further research is needed to confirm these findings.	Overall, there was insufficient evidence that administration of peri-implantation glucocorticoids in IVF/ICSI cycles influenced clinical outcomes. These findings were limited to the routine use of glucocorticoids in subfertile women undergoing IVF or ICSI.
The evidence suggests that bioengineered nerve conduits and wraps are a safe and effective option for repair of transected peripheral nerves in the upper limb. They are comparable to conventional microsurgical techniques for repairing nerve lacerations, and for shorter nerve gaps, may provide superior results in terms of sensory recovery. Additionally, the use of a conduit eliminates the donor-site morbidity associated with nerve-graft harvesting.	The evidence suggests that nerve repair using a biodegradable poly(DL-lactide-epsilon-caprolactone) Neurolac nerve guide, a collagen nerve conduit, or a polyglycolic acid conduit can be as effective as conventional microsurgical techniques in repairing transected peripheral nerves in the hand. Furthermore, these methods can potentially reduce the occurrence of cold intolerance, and may be beneficial for repairing nerves with gaps of 4 mm or less. Further research is needed to confirm these findings.	Based on the available evidence, this review does not support use of currently available nerve repair devices over standard repair. There is significant heterogeneity in participants, injury pattern, repair timing, and outcome measures and their timing across studies of nerve repair using bioengineered devices, which make comparisons unreliable. [...] provide high-certainty evidence and facilitate more detailed analysis of effectiveness of emerging, increasingly sophisticated, bioengineered repair devices.

Figure 15: Examples of generated summaries in the few-shot setting and their associated target summaries

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7 (after conclusion)

☑ A2. Did you discuss any potential risks of your work?
RQ4, section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2, 3

☑ B1. Did you cite the creators of artifacts you used?
Section 2, 3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2, 3

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix, and will be released with the data

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix and Section 2, 3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 2, 3

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No - annotation work like this does not require IRB, and i have discussed this with our folks here before

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We only hired annotators based on their expertise, demographic/geographic characteristics were not part of this.