GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

Asking Questions about Scientific Articles—Identifying Large N Studies with LLMs

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Paroiu, Razvan	National University of Science and Technology	
Ruseti, Stefan	National University of Science and Technology	
Dascalu, Mihai	National University of Science and Technology	
Trausan-Matu, Stefan	National University of Science and Technology	
McNamara, Danielle S.	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2023

Check type of content being submitted and complete one of the following in the box below:

☑ If article: Name of journal, volume, and issue number if available

□ If paper: Name of conference, date of conference, and place of conference

□ If book chapter: Title of book, page range, publisher name and location

□ If book: Publisher name and location

□ If dissertation: Name of institution, type of degree, and department granting degree

Electronics, 12(19)

DOI or URL to published work (if available) https://doi.org/10.3390/electronics12193996

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** National Center for Educat through **[Grant number]** R305N2100 to **Institution]** Arizona State U.The opinions expressed are those of the authors and do not represent views of the **[Office name]** National Center for Educa or the U.S. Department of Education.

electronics

MDPI

Article

Asking Questions about Scientific Articles—Identifying Large N Studies with LLMs

Razvan Paroiu [1], Stefan Ruseti [1], Mihai Dascalu [1,2,*], Stefan Trausan-Matu [1,2] and Danielle S. McNamara [3]

[1] Computer Science and Engineering Department, National University of Science and Technology Politehnica of Bucharest, 313 Splaiul Independentei, 060042 Bucharest, Romania; razvan.paroiu@upb.ro (R.P.); stefan.ruseti@upb.ro (S.R.); stefan.trausan@upb.ro (S.T.-M.)

[2] Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania

[3] Department of Psychology, Arizona State University, Tempe, AZ 85287, USA; dsmcnama@asu.edu

* Correspondence: mihai.dascalu@upb.ro

Abstract: The exponential growth of scientific publications increases the effort required to identify relevant articles. Moreover, the scale of studies is a frequent barrier to research as the majority of studies are low or medium-scaled and do not generalize well while lacking statistical power. As such, we introduce an automated method that supports the identification of large-scale studies in terms of population. First, we introduce a training corpus of 1229 manually annotated paragraphs extracted from 20 articles with different structures and considered populations. Our method considers prompting a FLAN-T5 language model with targeted questions and paragraphs from the previous corpus so that the model returns the number of participants from the study. We adopt a dialogic extensible approach in which the model is asked a sequence of questions that are gradual in terms of focus. Second, we use a validation corpus with 200 articles labeled for having *N* larger than 1000 to assess the performance of our language model. Our model, without any preliminary filtering with heuristics, achieves an F1 score of 0.52, surpassing previous analyses performed that obtained an F1 score of 0.51. Moreover, we achieved an F1 score of 0.69 when combined with previous extraction heuristics, thus arguing for the robustness and extensibility of our approach. Finally, we apply our model to a newly introduced dataset of ERIC publications to observe trends across the years in the Education domain. A spike was observed in 2019, followed by a decrease in 2020 and, afterward, a positive trend; nevertheless, the overall percentage is lower than 3%, suggesting a major problem in terms of scale and the need for a change in perspective.

Keywords: large language models; question answering; scientific article processing

check for updates

Citation: Paroiu, R.; Ruseti, S.; Dascalu, M.; Trausan-Matu, S.; McNamara, D.S. Asking Questions about Scientific Articles—Identifying Large N Studies with LLMs. *Electronics* **2023**, *12*, 3996. https://doi.org/10.3390/electronics12193996

Academic Editor: Arkaitz Zubiaga

Received: 3 August 2023
Revised: 17 September 2023
Accepted: 18 September 2023
Published: 22 September 2023

1. Introduction

The exponential growth of scientific publications [1] increases the effort for researchers to identify relevant articles. Even if multiple alternatives exist, such as Google Scholar (https://scholar.google.com/; accessed on 6 June 2023) or Semantic Scholar (https://www.semanticscholar.org/; accessed on 6 June 2023) with articles from a broad range of fields, the process remains cumbersome. Thus, additional automated tools are well-suited to refine the queries—for example, Elicit (https://elicit.org/; accessed on 6 June 2023) is designed as an AI research assistant focused on automating workflows, including the literature review process.

From the researchers' perspective, determining the appropriate number of participants for a new study involves identifying the minimal effect size that holds significance within their research and then finding the minimum number of participants needed to attain the specific effect size [2,3]. Additionally, a commonly employed heuristic involves an in-depth examination of the existing literature for similar research studies and extracting the participant count from those studies, which subsequently serves as a benchmark for the participant number in the researcher's own study [4].

From a complementary perspective, the scale of studies is a frequent barrier to research as it hampers the generalizability of findings. Employing a smaller subset of a population for study and subsequently extrapolating the findings to the broader population is a commonly used method among researchers. However, it is imperative that the characteristics of the smaller sample closely align with those of the larger population to which the researcher aims to generalize. The applicability of the findings to settings or individuals beyond the examined sample, referred to as "generalizability", can be influenced by factors such as the size of the sample and the approach used for sampling.

Although definitive guidelines regarding the required number of participants in a study are not universally established, large-scale studies are frequently associated with thousands of participants [5]; as such, we consider $N = 1000$ as our threshold. These studies yield a greater statistical significance, although the sampling technique may dramatically impact the results and induce bias [6].

When referring to the Education domain, scale plays a critical role as low-scaled studies are frequently encountered, and a special focus should be placed on large-scale studies. As such, we focus on developing automated methods that support identifying large-scale studies to address the previous challenges.

Our research objective is to find and evaluate articles in terms of scale by asking questions in natural language about the targeted papers presented as full-text papers, not only abstracts. More specifically, we prompt an FLAN-T5 XXL language model [7] with targeted questions and paragraphs from a testing corpus such that the model eventually returns the number of participants from the study. We build on the work of Corlatescu et al. [8] and introduce a generalizable approach in which we use large language models to answer specific questions about research papers.

Our main contributions are as follows:

- We introduce three new datasets. The first dataset consists of manually annotated paragraphs extracted from 20 articles with different considered populations. The second dataset is a 200-article validation corpus labeled binary whether the study has a large scale (i.e., contains more than 1000 participants). The third dataset consists of 460,164 unlabeled articles crawled from eric.ed.gov (accessed on 6 June 2023) database and then parsed into JSON files. We publicly release our datasets at https://largenineducation.org/datasets-and-publications (accessed on 6 June 2023).
- We define a dialogic approach for extracting the number of participants from any given study by prompting a pre-trained open-source Large Language Model (LLM); as such, we establish a sequence of questions that are gradual in terms of focus. We publicly release our codebase at https://github.com/readerbench/article-analyzer (accessed on 6 June 2023).
- We combine the LLM approach with previous extraction heuristics to obtain state-of-the-art results in identifying the scale of the study.
- We apply our method to the ERIC dataset, emphasizing the problem of scale in Education.

We opted to select an open-source LLM to ensure reproducibility while making everything available to the general public for future usage.

2. Related Work

In this section, we begin by presenting existing solutions for extracting relevant information from scientific articles. Following that, we provide a summary of the state-of-the-art LLMs and their applications in the field of NLP.

2.1. Extracting Relevant Information from Scientific Articles

While focusing on the task of extracting large-scale studies, the authors of ref. [8] managed to achieve an F1 score of 0.633 on a manually annotated corpus of 1000 articles by using multiple heuristics based on advanced text searches, regular expressions, and Natural Language Processing (NLP) techniques such as speech tagging and syntactic dependency parsing. They also applied the same strategy to a corpus of around 33,531 full-text articles extracted by crawling several journals and conference websites such as Frontiers, CHB, CE, L@S, AIED, and EC-TEL. The authors discovered 3347 studies with more than 1000 participants for their dataset. Their method had limitations in handling false positives, such as references to the number of participants from other studies or coefficients that were also named *N*, and one solution they proposed for future research was to use a Machine Learning model to find the number of subjects from any study.

Experiments of prompting language models were made in several previous works [9–13] for extracting relevant data—for example, by prompting a neural network with a simple question such as "What is the location?" [14]. However, prompt-based methods assigned to LLMs have been shown to be limited in Named Entity Recognition (NER) tasks. Problems with zero-shot experiments and the lack of robust prompts to obtain the necessary information have been observed [15–21]. One solution was to develop strategies for automatically converting the problems to Question Answering (QA) models, which improved the zero-shot capability for NER [14].

An important strategy also considered in our work asks the model to generate the question to which it has to respond. This method, known as Ask Me Anything (AMA) [22], produces multiple prompts from which a human can select which one best suits the task to solve. The motivation of this method is to reduce the effort dedicated to discovering the prompt that returns the best results. The model is asked to generate questions that it has to answer only after the best formats of questions have been studied. The study argued that open-ended questions are the best alternative.

Other studies also suggest that LLMs work effectively at few-shot information extraction from clinical data [23,24]. A problem in extracting information from clinical notes is the lack of manually annotated datasets by specialists. Despite this, LLMs have proven reliable at extracting information through zero-shot or few-shot, even if they have not been trained on medical datasets, and, in some situations, prompting GPT-3 produced better results than prompting PubMedBERT [25], a model trained only on clinical data [23].

Similarly, LLMs such as GPT-3 [26] and PaLM [27] were proven to be excellent few-shot learners without needing expensive human annotations. In some situations, extracting multiple concepts from a given text is required, and more modern large language models, such as T0 and InstructGPT, have been fine-tuned with different training objectives that encourage the model to perform well in these scenarios [28,29]. However, we introduce a method that can be gradually extended to extract relevant meta-data from various studies and can be easily deployed with limited resources.

The Elicit research assistant (https://elicit.org/; accessed on 6 June 2023) is another application where pre-trained language models are used for information retrieval. Elicit is an online platform based on GPT-3 where users enter a question about their research interest, and the neural network answers the query using information from about 175 million articles. The service returns the top relevant papers and also generates additional content, such as abstract summaries, based on the data gathered from those articles.

2.2. Language Models

In NLP, transfer learning is achieved by pre-training a model on large amounts of unlabeled data and then fine-tuning it on smaller labeled datasets. A unified text-to-text model is a pre-trained neural network where both input and output are text strings. One such example is the Text-To-Text Transfer Transformer (T5), which employs several transfer learning strategies proven to produce the best results [30]. Some of these strategies have already been used by popular models such as BERT [31] and XLNet [32].

The FLAN-T5 language model is a fine-tuned version of the T5 language model [7]. Its main advantage is that it can be used in zero-shot or few-shot scenarios, as it was fine-tuned on a large number of tasks, while T5 was only pre-trained, and it needs further fine-tuning. FLAN-T5 even outperforms a considerably larger language model—PaLM 62B [33] on the BigBench Hard benchmark. This advantage is especially useful for our purpose when an article describes a study in which n male subjects and m female subjects participated, and the model must return the sum $n + m$.

Wei et al. [34] also argued that fine-tuning models on a wide variety of NLP datasets and tasks enhances the few-shot effectiveness of LLMs. The researchers trained a language model of 137B parameters on over 60 different tasks in the field of NLP. The results obtained by this model (FLAN) were later compared with GPT-3, a pre-trained neural network with 175B parameters. The final conclusion was that this process improved FLAN's performance compared to the GPT-3, both in zero-shot and few-shot tasks [34].

Large Language Models use few-shot learning instead of fine-tuning the neural network. In contrast to the fine-tuning strategy, the model receives a smaller set of data which helps the neural network learn the current task, and there is also no weight change. This was motivated by the observation that people do not need a lot of supervised information to learn new language tasks if they have already mastered the basics of language. Although GPT-3 demonstrates that, in some tasks, the zero-shot and few-shot learning almost match the performance of fine-tuned systems, fine-tuning still gives better results than few-shot learning in the majority of situations [26].

Given their continually increasing learning capabilities [35], current LLMs are the future generation of pre-trained models, eliminating the need to train new neural networks from scratch each time new tasks arise. Researchers have shown that neural models learn new tasks through fine-tuning, and at the same time, they can retain their previously acquired knowledge through pre-training. This phenomenon is called Continual Learning, and these models have also proved that they can combine newly acquired knowledge with old knowledge to a certain extent [35].

Besides GPT-3 and GPT-4 [36], another important model worth mentioning is BLOOM [37]. This model was developed through the collaboration of hundreds of researchers trained on the ROOTS corpus that contains data from hundreds of sources. Most of the data is in natural language, but it also contains code from several programming languages. The model has a Transformer decoder architecture similar to GPT3, having 176 billion parameters; however, smaller versions also exist. Bloom's main advantage is that it is open-source and can be used as long as there is enough computing power and memory.

3. Method

We envisioned two consecutive filtering steps involved in the process of finding large-scale studies. Each of these steps corresponds to a prompt that is given to a neural network to answer whether or not a paragraph contains the necessary information. Our criterion for large scale is that it involves more than 1000 participants (i.e., $N \geq 1000$); nevertheless, our model provides the actual number of participants, and this threshold can be adjusted accordingly for modeling the evolution of large-scale studies. We adopt a dialogic approach in which the model is asked a sequence of questions that are gradual in terms of focus.

This section provides details on the three targeted datasets and experiments with prompting LLMs, as well as additional experiments wherein we provide insights into alternatives that we considered, but which achieved lower performance.

3.1. Datasets

Our first corpus is used to find the most suitable prompts. This corpus comprises 20 articles selected from a group of 50 randomly extracted articles from the EdPub corpus introduced by [8]. The 20 articles were selected based on their differences. Ten articles were not studies but contained results from previous studies. Others had a lot of numerical information. The other ten articles were studies, but some did not contain the number of participants in the study or a specific number of schools from which the participants were selected. Other articles included the number N of individuals involved in the study only in one paragraph from the entire article. We included the title, abstract, number of subjects who participated in the study, and the list of paragraphs from each article of our corpus.

For each paragraph, we included its primary section, text, and the number of participants involved in the study.

Second, a newly annotated corpus of 200 articles was used to validate the results. The articles were selected randomly from a larger corpus extracted from crawling multiple Internet sources [8]. The information available for each article consisted of the publication venue, the year when the paper was published, the title, whether it was a study that employed more than 1000 participants, and the total number of individuals involved in the studies. From the corpus of 200, only 15 papers were large-scale studies. Even though this corpus contains considerably more articles than the previous corpus of 20 annotated papers, the articles were labeled globally and not at the paragraph level.

Third, we also compiled a large corpus by crawling the eric.ed.gov (accessed on 6 June 2023) database using the BeautifulSoup library https://pypi.org/project/beautifulsoup4/ (accessed on 6 June 2023) from which we extracted all articles ranging from 1965 to 2022. For each article, we downloaded the following metadata: the title of the article, the authors, the date when the article was published, the abstract, the institution where the authors were working, the ISBN and/or ISSN, the language in which the article was written, the publisher, and its URL. Our analysis goes beyond the abstracts as we downloaded all available PDFs.

The content of the articles, structured into sections and paragraphs, was extracted using heuristics based on the distribution of used fonts and sizes. As such, our custom parser discriminated between headings and normal text while ignoring captions, footnotes, or references. The extraction method worked best when applied to papers with at least 4 pages; as such, we ignored anything below that. In addition, we ignored articles with more than 20 pages since they were most often books. Thus, we successfully extracted the full content for 170,524 articles.

3.2. Experiments with Prompting LLMs

All our experiments consider the FLAN-T5 XXL language model [7], a fine-tuned version of the T5 language model that introduced a unified text-to-text strategy. Before prompting the model in a dialogic approach, we filtered the paragraphs based on the primary section to which they belonged. By doing this, we eliminated the paragraphs belonging to the Introduction section because these most likely contain numerical data from other articles frequently presented in the state-of-the-art section. We made this assumption based on both our hands-on experience of manually annotating the training and testing datasets and also on the findings of a separate study where Barry et al. [38] reviewed 940 articles along with their respective state-of-the-art. We also eliminated the paragraphs from the Conclusions and References sections because it is unlikely to find there the number of participants from the current study; this decision is also based on our hands-on experience, whereas information about the research from the article is absent in the References section. We also concatenated multiple paragraphs into one sequence of paragraphs until it reached 2000 characters. We made this decision because multiple paragraphs are rather short, and our model performs better when more context is included in the prompt.

The Flan-T5 model was loaded and run using the HuggingFace Transformers library. A collection of prompts was created based on the specifically sought information (whether discerning if a given paragraph describes a new study or finding the count of participants within the study). Each prompt was then systematically paired with paragraphs drawn from the training dataset. For every such prompt–paragraph pair, the model was prompted with a structured sequence: the designated prompt, followed by the descriptor “Text:”, the paragraph content, and ultimately, the identifier “Answer:” (to signify the model’s anticipated response). The prompts were created either manually or generated with the AMA (Ask Me Anything) [22] method.

In the first prompt, we decided to ask the model whether a certain paragraph describes a new study or not. To do this, the model must do a binary classification of paragraphs. As such, instead of generating an answer such as “Yes”, “No”, “True”, or other similar alternatives, we looked at the logits returned by the model for the outputs “Yes” and “No”, and then we computed the probability of the two answers. The decision considers the cross-entropy loss for the “Yes” and “No” labels. Following this, our approach involved interpreting the label associated with the lower loss as indicative of the model’s higher confidence in that particular answer.

We decided to accept only the answers for which the neural network was confident because there are numerous cases when the model considers that multiple paragraphs are describing a new study, whereas they only describe studies that have been published in other articles. We did this by computing the ratio between the probability of answering “Yes” and the probability of answering “No”. Situations when the losses of both labels exhibited similar values indicated that the model lacked sufficient confidence in selecting the correct answer.

Finally, we established a threshold for this ratio using the 20-article corpus. We then computed the F1 scores of the results for all the ratios between 0 and 10 using prompts based on seven different questions, and the best 3 results, together with the threshold for which the outcome was possible, can be seen in Table 1.

Table 1. Results from 7 different prompts for which the model answers if a sequence of paragraphs describes a new study or not (bold marks the best F1 score).

Prompt	TP	TN	FP	FN	P	R	F1	RT
Is this paragraph describing a new study?	4	223	13	11	0.23	0.26	0.25	0.9
Is this paragraph introducing a study with participants?	9	227	9	6	0.50	0.60	0.54	5.9
Is this paragraph introducing a study?	7	222	14	8	0.33	0.46	0.38	2.6
Is this sentence introducing research that includes participants?	10	210	26	5	0.27	0.66	0.39	9.5
Does this paragraph introduce a study with participants?	10	203	33	5	0.23	0.66	0.34	7.8
Is this text introducing research that includes participants?	10	203	33	5	0.23	0.66	0.34	9.9
Does this text introduce a new study that includes several participants? The text must include the number of participants.	10	230	6	5	0.62	0.66	**0.64**	2.7

TP = true positive; TN = true negative; FP = false positive; FN = false negative; P (Precision) = TP / (TP + FP); R (Recall) = TP / (TP + FN); F1 = 2*P*R / (P + R); RT = ratio threshold;).

For the second prompt, we instructed the model to return the number of participants from the current study. In order to force the model only to generate numbers, we used a custom logits processor, which overrides the probability of any token that is not a number or a special token with 0. We also decided to accept only the answers for which the neural network is confident by getting the probability of the generated answer and accepting it only if it was above a chosen threshold. We then prompted the model with 25 different questions for every sequence of paragraphs from our testing corpus. Seven prompts were manually chosen, and 18 were generated by our model using the AMA method [22]; the corresponding results are presented in Table 2.

The text “Write the question that gives the following answer given the context: *Answer*:—the number of participants from the study—*Context*:—the sequence of paragraphs to which the model has to answer—” was prompted to the neural network for each sequence of paragraphs from our testing corpus. The model produced 1047 different questions, but most were customized for the sequence of provided paragraphs and could not be used generally. As a result, we only chose 18 questions that we considered generic.

Since the ratio threshold for the first prompt and the probability threshold for the second prompt are tightly coupled with each other, we chose the two thresholds by looking at the F1 score after applying the two prompts in a sequence. This means that the second prompt is applied only if the ratio from the first prompt exceeds the threshold. For the first prompt, we used “Does this text introduce a new study that includes several participants? The text must include the number of participants.” because it was the most effective at answering if a sequence of paragraphs describes a new study or not (as seen in Table 1). For the second prompt, we used the last four prompts that produced the best results (as seen in Table 2) individually in extracting the number of participants from the study.

Table 2. Results from 25 different prompts for which the model is instructed to extract from each paragraph the number of participants to the study. The first 7 prompts were manually chosen, whereas the last 18 were generated through AMA (bold marks the best 4 F1 scores).

Prompt	TP	TN	FP	FN	P	R	F1	PT
How many subjects took part in the following study?	9	187	51	4	0.15	0.69	0.24	0.006
How many subjects were involved in the next study (answer in digits)?	10	129	110	2	0.08	0.83	0.15	0.004
How many subjects were involved in total in the next study (provide a number)?	9	148	92	2	0.08	0.81	0.16	0.005
How many subjects were involved in the next study?	7	150	88	6	0.07	0.53	0.12	0.005
The following study had how many subjects?	9	130	109	3	0.07	0.75	0.13	0.005
How many subjects participated in the following study?	10	175	64	2	0.13	0.83	0.23	0.006
Extract the number of entries in the study:	11	180	58	2	0.15	0.84	0.26	0.006
How many subjects were in the study?	10	186	53	2	0.15	0.83	0.26	0.006
How many people were in the study?	7	178	60	6	0.10	0.53	0.17	0.007
How many subjects were in the experiment?	9	194	45	3	0.16	0.75	**0.27**	0.006
How many people were in the experiment?	8	200	38	5	0.17	0.61	**0.27**	0.007
How many subjects participated in the experiment?	8	182	58	3	0.12	0.72	0.20	0.006
How many people participated in the experiment?	8	178	62	3	0.11	0.72	0.19	0.006
How many subjects participated in the study?	10	162	77	2	0.11	0.83	0.20	0.006
How many people participated in the study?	8	163	76	4	0.09	0.66	0.16	0.007
How many subjects are in the study?	9	206	32	4	0.21	0.69	**0.33**	0.006
How many people are in the study?	6	204	35	6	0.14	0.50	0.22	0.008
How many subjects were involved in the study?	9	177	63	2	0.12	0.81	0.21	0.006
How many people were involved in the study?	9	173	66	3	0.12	0.75	0.20	0.006
How many subjects were selected?	4	202	35	10	0.10	0.28	0.15	0.005
How many people were selected?	7	145	94	5	0.06	0.58	0.12	0.004
How many subjects were involved in the investigation?	4	204	34	9	0.10	0.30	0.15	0.007
How many people were involved in the investigation?	8	204	35	4	0.18	0.66	**0.29**	0.006
How many subjects participated in the study?	10	162	77	2	0.11	0.83	0.20	0.006
How many people participated in the study?	8	163	76	4	0.09	0.66	0.16	0.007

TP = true positive; TN = true negative; FP = false positive; FN = false negative; P (Precision) = TP / (TP + FP); R (Recall) = TP / (TP + FN); F1 = 2*P*R / (P + R); PT = Probability threshold.

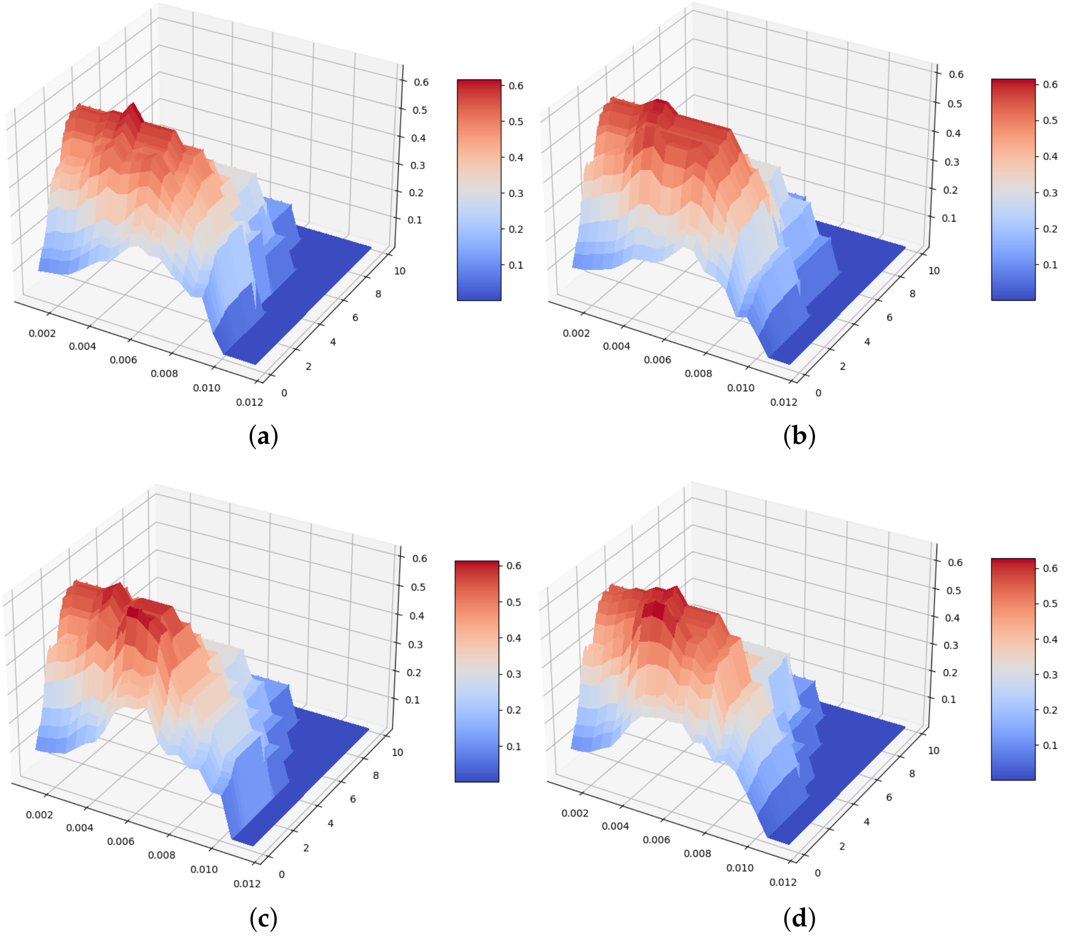
Table 3 depicts the results of these experiments together with the ratio threshold and probability threshold for which these results were possible. Figure 1 displays the F1 scores results for the complete sequence of ratio thresholds between 0 and 10 and the entire sequence of probability thresholds between 0 and 0.012.

Table 3. Results using the second prompts selected from Table 2 together with the first prompt selected from Table 1 (bold denotes the best F1 score for the specified ratio threshold and probability threshold).

Prompt	TP	TN	FP	FN	P	R	F1	RT	PT
How many subjects are in the study?	9	231	5	6	0.64	0.60	0.62	1.2	0.005
How many people were involved in the investigation?	10	230	6	5	0.63	0.67	**0.645**	1.3	0.0045
How many subjects were in the experiment?	8	234	2	7	0.80	0.53	0.64	2.7	0.004
How many people were in the experiment?	8	233	3	7	0.73	0.53	0.615	2.7	0.0035

TP = true positive; TN = true negative; FP = false positive; FN = false negative; P (Precision) = TP / (TP + FP); R (Recall) = TP / (TP + FN); F1 = 2*P*R / (P + R); RT = Ratio threshold; PT = Probability threshold.

Figure 1 also illustrates a significant decline in the F1 score as the probability threshold surpasses 0.006, regardless of any ratio threshold value. This large decrease indicates that the model's confidence in its responses concerning the number of participants from the study is limited in many instances. A similar trend is also observed when considering the ratio threshold, where ratios exceeding 4 are infrequent, and the threshold excludes a substantial number of results, thus contributing to the reduction of the F1 score.



(**a**) (**b**) (**c**) (**d**)

Figure 1. F1 scores of the results of the combinations between the initial prompt and the best 4 prompts presented in Table 1. The ratio threshold spans from 0 to 10, while the probability threshold ranges from 0 to 0.012. (**a**) How many subjects were in the experiment? (**b**) How many people were in the experiment? (**c**) How many subjects are in the study? (**d**) How many people were involved in the investigation?

3.3. Combining Prompted LLMs with Extraction Heuristic

In order to further enhance our approach, we introduce a combined method that first filters potential documents of interest using the heuristics defined by Corlatescu et al. [8], followed by the usage of the prompted LLM. First, we applied the heuristic-based method and continued applying the method based on neural networks to any article for which the heuristics predicted the existence of a number of participants greater than 1000. We first divided the article into paragraphs and then prompted the FLAN-T5 model for each paragraph with the predefined prompts using the computed thresholds.

As such, we combine the strengths of both approaches: the initial filtering is more rigorous and focused on extracting potential numbers of interest, whereas the LLM better extracts relevant information given its contextualized representations and inferencing capabilities. The final processing flow is presented in Figure 2.

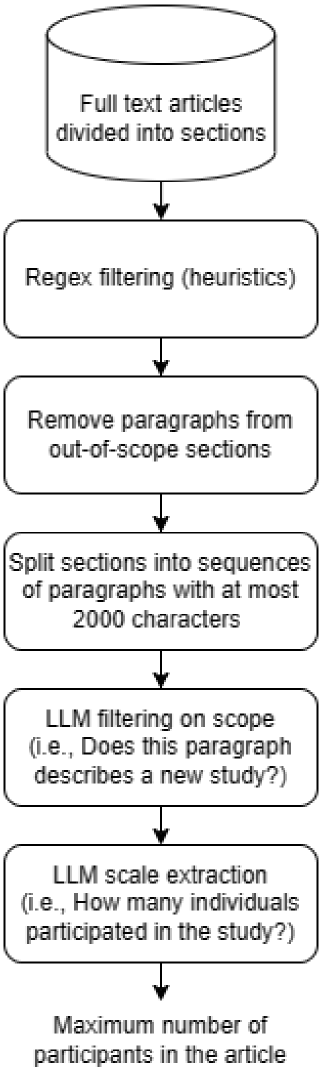

Full text articles divided into sections
Regex filtering (heuristics)
Remove paragraphs from out-of-scope sections
Split sections into sequences of paragraphs with at most 2000 characters
LLM filtering on scope (i.e., Does this paragraph describes a new study?)
LLM scale extraction (i.e., How many individuals participated in the study?)
Maximum number of participants in the article

Figure 2. Application flow chart.

3.4. Additional Experiments Yielding Inferior Results

In the following, we present several experiments which did not yield satisfactory results. Nevertheless, we see fit to introduce these details to support our method's argumentation further.

Besides FLAN-T5 XXL, the same method was implemented using BLOOM 7B1 [37]; however, as a result of BLOOM's many hallucinations, the results were inferior. Since the version of BLOOM used in our experiment had 7 billion parameters, and FLAN-T5 XXL has 11 billion parameters, we assume that the entire version of BLOOM would achieve better results; however, the 176B version requires 8 NVidia A100 GPUs for inference, which defies our goal of having a manageable model.

Another attempt was to use three prompts instead of two prompts. The third prompt instructed the model to return the number of subjects in digits using the answer from the second prompt as input. We experimented with 3 different prompts, and the results were promising in some situations. We noticed that, in numerous situations, when the model was prompted with the second question and returned a false positive value, it returned the answer as a text, not in digits. Afterward, we forced the model to return for the second prompt only digits; as such, the third prompt became irrelevant.

We also experimented with different parsing algorithms. The PDF dataset extracted from ERIC was parsed into text and then into paragraphs using the algorithm developed in the previous research of Corlatescu et al. [8]. Moreover, we also experimented with the

Grobid library https://grobid.readthedocs.io/en/latest/ (accessed on 6 June 2023). Grobid (GeneRation Of Bibliographic Data) is a library based on ML algorithms to parse PDF documents into data structured in XML/TEI. This output contains the text from the PDF file structured in paragraphs and some metadata. The library can be used on any type of PDF file, but it is specialized in scientific articles. The Grobid library was successfully used on the validation corpus of 200 articles because the dataset contained only scientific articles published in reputable journals. In the case of all articles from ERIC, the PDF files had, in many situations, an irregular or custom format that could not be parsed with Grobid. As such, we decided to continue the experiments using the custom parser developed by Corlatescu et al. [8].

4. Results

We used the 200-article corpus to evaluate our method. The best two prompts from the 20-article corpus and their corresponding thresholds (see Table 3) were applied to the new dataset. This corpus does not contain annotated paragraphs; thus, we applied our method to each individual paragraph and returned the maximum discovered value.

The first filtering eliminated the paragraphs that belonged to the primary sections: "introduction", "conclusions", and "references". In order to eliminate situations where the section's name starts with an uppercase letter or the entire name contains only uppercase letters, we lowered the case of the letters and eliminated punctuation. The second filtering removed paragraphs not describing a study, in accordance with the results obtained from prompting the model with the question: "Does this text introduce a new study that includes several participants? The text must include the number of participants." with a confidence ratio of 1.3. In the second phase, we prompted the model with the second question: "How many people were involved in the investigation?" with a confidence probability of 0.045. Table 4 displays the final results of our method together with a side-by-side analysis using the heuristics by Corlatescu et al. [8].

Table 4. Confusion matrix of our method (left; F1 = 0.52) and heuristics (right; F1 = 0.51) using the validation corpus of 200 articles.

	Predicted: 0	**Predicted: 1**		**Predicted: 0**	**Predicted: 1**
Actual: 0	166	22	Actual: 0	174	14
Actual: 1	0	12	Actual: 1	3	9

The results obtained when combining these two methods on the corpus of 200 articles are presented in Table 5. The high increase in the F1 score argues that the techniques complement one another.

Table 5. Confusion matrix of heuristics rules combined with our method using the validation corpus of 200 articles (F1 = 0.69).

	Predicted: 0	**Predicted: 1**
Actual: 0	183	5
Actual: 1	3	9

5. Discussion

Following the results obtained by applying the LLM method to the corpus of 200 articles, we observed that the model recognized all situations when an article had over 1000 participants (12 articles in total); as such, the recall score was 1.0, while the heuristic method obtained a recall score of 0.75.

In contrast, the false positive results obtained by the LLM are more numerous than those obtained by the heuristic-based method, this being mainly because the neural model hallucinates about numbers not present in the text. Because of this, the precision was 0.35, while the heuristics-based method had a precision of 0.39.

The combined method inherits a lower recall score due to the initial heuristics (R = 0.75) but achieves a considerably higher precision (P = 0.64), given the LLM's better contextualization.

5.1. Evolution of Large-Scale Studies in ERIC

Next, we modeled the evolution of large-scale studies from the ERIC dataset. The same strategy and corresponding prompts evaluated on the annotated dataset were applied to the large ERIC dataset. Out of the 170,524 articles with extracted text, we focused on recent articles with a timeframe of 10 years, namely, between 2013 and 2022.

The combined method highlights an even gloomier view than the initial findings of Corlatescu et al. [8], namely, that only roughly 3% of articles have a large scale (see Figure 3 and Table 6). We observed an increase in large-scale studies in 2019; this was followed by a decrease in 2020 and afterward a positive trend across subsequent years. Although we do not know with certainty, we consider that this phenomenon is due to the beginning of the COVID-19 pandemic, when numerous large-scale studies were conducted to examine the effects of pandemics on education.

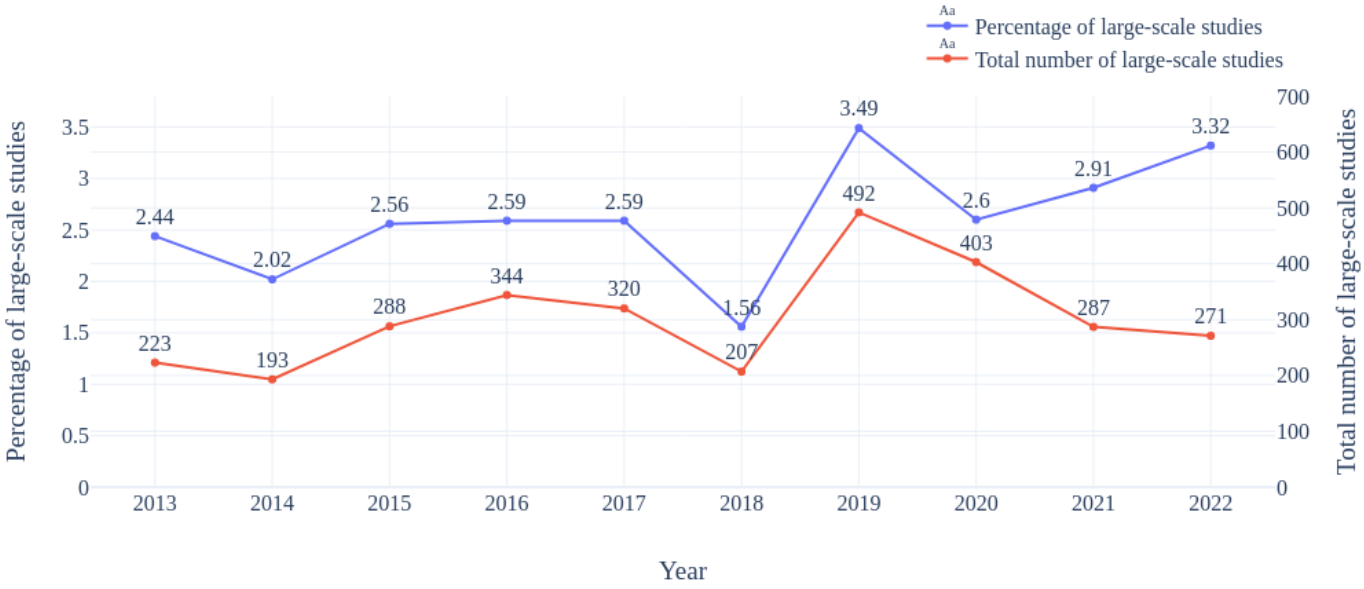


Figure 3. Percentage and number of large-scale studies in the ERIC dataset.

Table 6. Total number of large-scale studies discovered in the ERIC dataset using the heuristic-based method followed by our current zero-shot strategy.

Year	Total Number of Articles	Total Number of Large-Scale Articles (Discovered by the Heuristics Method)	Total Number of Large-Scale Studies (Discovered by Both Methods Combined)	Percentage of Large-Scale Studies
2022	8143	674	271	3.32%
2021	9857	851	287	2.91%
2020	15,491	1134	403	2.6%
2019	14,061	995	492	3.49%
2018	13,225	622	207	1.56%
2017	12,355	853	320	2.59%
2016	13,267	938	344	2.59%
2015	11,216	798	288	2.56%
2014	9545	655	193	2.02%
2013	9109	554	223	2.44%
Total count	116,269	8074	3028	2.60%

Our findings are particularly significant to the field of educational research because they have an impact on the validity, generalizability, and reliability of research findings. Larger sample sizes improve reliability by reducing the impact of random fluctuations. As a result, the findings are more likely to be statistically significant than accidental results. Because educational research aims to provide insights with broad relevance across many contexts, people, and settings, generalizability is also important. Large-scale studies, which include thousands of individuals from various backgrounds, produce results that are more generalizable to a larger population.

Furthermore, variability is important given the complex nature of educational environments, where a variety of factors influence learning outcomes. Larger studies are also better at capturing this complexity, which leads to stronger conclusions about the correlations between variables. The subtle effects of certain interventions or circumstances on educational learning outcomes are another crucial issue to take into account. These impacts, which can go undetected in smaller-scale studies, are easily picked up by large-scale studies.

5.2. Disparities between the Current Method and the Previous Heuristics-Based Approach

The misclassifications produced by our method are examples of its limitations, which require further expansion of the training dataset with negative samples. Three specific situations are targeted. First, generalizing from individuals and documents to any type of entity (e.g., one study considered match observations). Second, variables from statistic analyses were combined into a single number (e.g., "F(1,206)" was not interpreted as between-groups degrees of freedom followed by within-groups degrees of freedom, but as a single number). Third, combinations of two populations were not summed up adequately (e.g., "teachers in Germany (n = 868; male = 23.1, SD = 3.3; female: 75.2%) and Costa Rica (n = 284, male = 25.8, SD = 6.6; female: 55.3%").

The numerous false positives produced by our method are a clear example of its limitations. In most situations, the second prompt hallucinates data not included in the sequence of paragraphs provided as context. There are also some situations when the second prompt is unable to recognize whether the data belong to a referenced state-of-the-art study or the current paper; nevertheless, the first prompt excludes these instances since they do not describe a new study. Table 7 details situations when the first or the second prompt would provide false results but have returned correct ones when combined.

It should be noted that the false positives and false negatives returned by our method are primarily due to the hallucinations of the neural model. This problem can be reduced by using more advanced neural models with more parameters in the future. We will now enumerate the types of problems that occurred in the previous method based on heuristics:

- Number *N* is a reference to another study that is part of SOTA.
- The number *N* does not represent the number of participants in the current study but only another number of entities from the text.
- Errors arising from the time of parsing the PDF into text. These errors are mostly eliminated using the method presented in this article because the model can determine if a number is concatenated to the text (without space) and can extract it separately.
- The number of participants represents the summation of several numbers in the text (e.g., if 100 students and 50 teachers participated in the study, then the number of participants is 150). These errors are eliminated in some cases using the current method, but not completely.

Our approach has an additional advantage over the heuristics method since we can get the exact number of participants from the study for any provided text. In contrast, the heuristics method simply returned if an article has a large *N* or not. If we modify the open-source code for the heuristics to return the precise number of research participants from the study, the algorithm occasionally generates a text instead of a number. The text returned by the algorithm corresponds to the section of the article where the large *N* was discovered (in some cases, several numbers with at least four digits can be noticed in that section). Future modifications will be required to extract only a number rather than a text.

Table 7. Situations when the first or second prompt provided false results, but their combined version worked well.

Case	Example (The Spans Contain Only the Important Parts of the Entire Sequence of Paragraphs of 2000 Tokens Given at Input)	Result
The second prompt extracts the number of participants from the SOTA, but the first prompt specifies that the text does not describe a new study.	"To illustrate, Borrero et al. (2010) conducted descriptive assessments with 25 children with feeding problems and their caregivers."	The second prompt makes the model state that the article describes a study in which 25 participants took part, but the first prompt states that this text does not describe a new study.
The second prompt hallucinates data that do not exist in the context (with a rather high probability of being true), but the first prompt blocks it as not being part of a new study.	"The purpose of this exploratory study was to examine online teachers' selfreported frequency and confidence in performing online learning tasks. Two questions guided this exploratory study: 1. Is there a difference between returning teachers' and new teachers' self-reported frequency of performed online teaching tasks (supporting student learning, supporting content learning, making learning adaptations, assessing learning, and managing learning processes)? 2. Is there a difference between returning teachers' and new teachers' self-reported confidence in performing online teaching tasks (supporting student learning, supporting content learning, making learning adaptations, assessing learning, and managing learning processes)?"	In a text with no specified number, the second prompt causes the neural network to return the number 161, with a probability of 0.08 (the threshold used is 0.0045, and the number 161 does not even represent the number of words in the text).
The first prompt assumes that the context describes a new study, but the second prompt, even if it extracts some data (such as the number of schools where a previous study was conducted), it is not confident enough that its answer is correct.	"math and science teachers from two middle schools first experienced inquiry"	The second prompt returns the number 2 with a probability of 0.002 (lower than 0.0045), but the first prompt specifies that the paragraph describes a study in which the number of participants is included in the text (the first prompt is: "Does this text introduce a new study that includes several participants? The text must include the number of participants.").
The first prompt assumes that the context describes a new study, but the second prompt, even if it extracts some data (such as the number of schools where a previous study was conducted), it is not confident enough that its answer is correct.	"A one-tailed, independent samples *t*-test (*N* = 102, df = 100) was conducted to identify statistically significant areas"	The first prompt makes the model return that the text describes a new study in which the number of participants is specified. The ratio of confidence is 3.06 (higher than 1.3), but the second prompt even if it returns the number 102, it still does so with a probability of 0.001 (lower than 0.0045).

6. Conclusions and Future Work

This paper introduces an extensible method in which an FLAN-T5 language model is asked a sequence of gradual questions in terms of focus to mimic a dialogue with the model. After performing experiments with different prompts and identifying the optimal configuration, our model surpassed previous heuristics in terms of F1 scores on a newly introduced validation corpus. We also argued that combining our method with the previously researched heuristics ensures a more robust model.

The main advantage of our method over the Elicit tool is that it is based on an open-source pre-trained neural network and open-source technologies. Another important aspect that gives our method a significant edge over Elicit and similar tools is that it generates answers based on the complete article content and not only on its abstract; this plays a crucial role since population information is frequently not present in the abstract. Given our validation corpus, only 124 articles mentioned the scale of the study in the abstract. Also, we provide a rigorous evaluation of our model compared to exporting the predictions from GPT-3. Nevertheless, our validation pinpointed some limitations of our prompts that need to be tackled in future analyses.

Three new datasets are also introduced in this paper. The first dataset consists of manually annotated paragraphs collected from 20 different articles. The second dataset, which was used as a validation corpus, consists of 200 articles labeled binary, indicating whether the study is a large-scale study (i.e., includes more than 1000 participants). Finally, the third dataset consists of 460,164 unlabeled articles that were parsed into JSON files after being crawled from the database of eric.ed.gov (accessed on 6 June 2023). Our ERIC dataset is available at https://largenineducation.org/datasets-and-publications (accessed on 6 June 2023). Since our method is based on a zero-shot strategy, future research will focus on experimenting with few-shot learning and fine-tuning an open-source model.

Author Contributions: Conceptualization, S.R. and M.D.; methodology, S.R., R.P., M.D. and D.S.M.; software, R.P. and S.R.; validation, R.P., S.R. and M.D.; formal analysis, S.R. and R.P.; investigation, R.P. and S.R.; resources, R.P., S.R. and M.D.; data curation, R.P. and S.R.; writing—original draft preparation, R.P.; writing—review and editing, M.D., S.R., S.T.-M. and D.S.M. visualization, R.P.; supervision, M.D., S.T.-M. and D.S.M.; project administration, M.D.; funding acquisition, M.D. and D.S.M. All authors have read and agreed to the published version of the manuscript.

Funding: The results presented in this article have been funded by a grant from Schmidt Futures and by the Ministry of Investments and European Projects through the Human Capital Sectoral Operational Program 2014–2020, Contract no. 62461/03.06.2022, SMIS code 153735. The research reported here was also supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305N210041 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We made freely available the training and validation datasets on which we performed our experiments, together with the extracted ERIC https://largenineducation.org/datasets-and-publications (accessed on 6 June 2023). Additionally, we release our code as open-source on GitHub (https://github.com/readerbench/article-analyzer accessed on 6 June 2023) in order to support future research on the dataset.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bornmann, L.; Haunschild, R.; Mutz, R. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 1–15. [CrossRef]
2. Martínez-Mesa, J.; González-Chica, D.A.; Bastos, J.L.; Bonamigo, R.R.; Duquia, R.P. Sample size: How many participants do We need in my research? *An. Bras. Dermatol.* **2014**, *89*, 609–615. [CrossRef]
3. Marc, B. How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *J. Cogn.* **2019**, *2*, 16. [CrossRef]
4. Capili, B. Selection of the Study Participants. *AJN Am. J. Nurs.* **2021**, *121*, 64–67. . [CrossRef] [PubMed]
5. Wagner, R.K.; Ridgewell, C. A large-scale study of specific reading comprehension disability. *Perspect. Lang. Lit.* **2009**, *35*, 27. [PubMed]
6. Kaplan, R.M.; Chambers, D.A.; Glasgow, R.E. Big data and large sample size: A cautionary note on the potential for bias. *Clin. Transl. Sci.* **2014**, *7*, 342–346. [CrossRef] [PubMed]
7. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *arXiv* **2022**, arXiv:2210.11416.
8. Corlatescu, D.; Ruseti, S.; Toma, I.; Dascalu, M. Where are the Large N Studies in Education? Introducing a Dataset of Scientific Articles and NLP Techniques. In Proceedings of the Ninth ACM Conference on Learning @ Scale, New York, NY, USA, 1–3 June 2022; pp. 461–465.

9. Chen, X.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; Chen, H.; Zhang, N. LightNER: A Lightweight Tuning Paradigm for Low-resource NER via Pluggable Prompting. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022, pp. 2374–2387.
10. Perez, E.; Kiela, D.; Cho, K. True few-shot learning with language models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 11054–11070.
11. Gao, T.; Fisch, A.; Chen, D. Making Pre-trained Language Models Better Few-shot Learners. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Online, 1–6 August 2021; Association for Computational Linguistics: Cambridge, MA, USA, 2021.
12. Li, D.; Hu, B.; Chen, Q. Prompt-based Text Entailment for Low-Resource Named Entity Recognition. In Proceedings of the 29th International Conference on Computational Linguistics; International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 1896–1903.
13. Le, V.H.; Zhang, H. Log Parsing with Prompt-based Few-shot Learning. *arXiv* **2023**, arXiv:2302.07435.
14. Liu, A.T.; Xiao, W.; Zhu, H.; Zhang, D.; Li, S.W.; Arnold, A. QaNER: Prompting question answering models for few-shot named entity recognition. *arXiv* **2022**, arXiv:2203.01543.
15. Utama, P.; Moosavi, N.S.; Sanh, V.; Gurevych, I. Avoiding Inference Heuristics in Few-shot Prompt-based Finetuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 10–11 November 2021; pp. 9063–9074. [CrossRef]
16. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Online and Punta Cana, Dominican Republic, 2021; pp. 3045–3059. [CrossRef]
17. Ma, R.; Zhou, X.; Gui, T.; Tan, Y.; Li, L.; Zhang, Q.; Huang, X. Template-free Prompt Tuning for Few-shot NER. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 June 2022; pp. 5721–5732. [CrossRef]
18. Ding, N.; Chen, Y.; Han, X.; Xu, G.; Wang, X.; Xie, P.; Zheng, H.; Liu, Z.; Li, J.; Kim, H.G. Prompt-learning for Fine-grained Entity Typing. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6888–6901.
19. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to Prompt for Vision-Language Models. *Int. J. Comput. Vis.* **2021**, *130*, 2337–2348. [CrossRef]
20. Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; Stenetorp, P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Volume 1: Long Papers, pp. 8086–8098. [CrossRef]
21. Ishibashi, Y.; Bollegala, D.; Sudoh, K.; Nakamura, S. Evaluating the Robustness of Discrete Prompts. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 2373–2384.
22. Arora, S.; Narayan, A.; Chen, M.F.; Orr, L.J.; Guha, N.; Bhatia, K.; Chami, I.; Sala, F.; Ré, C. Ask Me Anything: A simple strategy for prompting language models. In Proceedings of the Eleventh International Conference on Learning Representations, Vienna, Austria, 7–11 May 2022.
23. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. Large language models are few-shot clinical information extractors. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 1998–2022.
24. Oniani, D.; Chandrasekar, P.; Sivarajkumar, S.; Wang, Y. Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks: Algorithm Development and Validation Study. *JMIR AI* **2023**, *2*, e44293. [CrossRef]
25. Danilov, G.; Ishankulov, T.; Kotik, K.; Orlov, Y.; Shifrin, M.; Potapov, A. The Classification of Short Scientific Texts Using Pretrained BERT Model. *Stud. Health Technol. Inform.* **2021**, *281*, 83–87. [PubMed]
26. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
27. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling Language Modeling with Pathways. *arXiv* **2022**, arXiv:2204.02311.
28. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegle, A.; Le Scao, T.; Raja, A.; et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv* **2021**, arXiv:2110.08207.
29. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155.
30. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
31. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. . [CrossRef]

32. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, Canada, 8–14 December 2019.
33. Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H.W.; Chowdhery, A.; Le, Q.V.; Chi, E.H.; Zhou, D.; et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv* **2022**, arXiv:2210.09261.
34. Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned Language Models Are Zero-Shot Learners. In Proceedings of the International Conference on Learning Representations, Online, 25–29 April 2022, pp. 1–46.
35. Scialom, T.; Chakrabarty, T.; Muresan, S. Fine-tuned Language Models are Continual Learners. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6107–6122.
36. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
37. Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv* **2022**, arXiv:2211.05100.
38. Barry, E.S.; Merkebu, J.; Varpio, L. State-of-the-art literature review methodology: A six-step approach for knowledge synthesis. *Perspect. Med. Educ.* **2022**, *11*, 281–288. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.