TYPE Original Research
PUBLISHED 20 August 2025
DOI 10.3389/fninf.2025.1609077

Large language models can extract metadata for annotation of human neuroimaging publications

Matthew D. Turner[1]*, Abhishek Appaji[2], Nibras Ar Rakib[3], Pedram Golnari[4], Arcot K. Rajasekar[5], Anitha Rathnam K V[6], Satya S. Sahoo[4], Yue Wang[5,7], Lei Wang[1] and Jessica A. Turner[1]

[1]Department of Psychiatry, The Ohio State University, Columbus, OH, United States, [2]Department of Medical Electronics Engineering, B.M.S. College of Engineering, Bengaluru, India, [3]Faculty of Information, University of Toronto, Toronto, ON, Canada, [4]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, United States, [5]School of Information and Library Science, University of North Carolina, Chapel Hill, NC, United States, [6]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India, [7]Carolina Health Informatics Program, University of North Carolina, Chapel Hill, NC, United States

OPEN ACCESS

EDITED BY
Pritpal Singh,
Central University of Rajasthan, India

REVIEWED BY
James Joseph Driscoll,
University Hospitals of Cleveland,
United States
Maheshi Dissanayake,
University of Peradeniya, Sri Lanka
Bayan Altalla,
King Hussein Cancer Center, Jordan

*CORRESPONDENCE
Matthew D. Turner
✉ matthew.turner2@osumc.edu

RECEIVED 09 April 2025
ACCEPTED 28 July 2025
PUBLISHED 20 August 2025

CITATION
Turner MD, Appaji A, Ar Rakib N, Golnari P, Rajasekar AK, K V AR, Sahoo SS, Wang Y, Wang L and Turner JA (2025) Large language models can extract metadata for annotation of human neuroimaging publications.
Front. Neuroinform. 19:1609077.
doi: 10.3389/fninf.2025.1609077

We show that recent (mid-to-late 2024) commercial large language models (LLMs) are capable of good quality metadata extraction and annotation with very little work on the part of investigators for several exemplar real-world annotation tasks in the neuroimaging literature. We investigated the GPT-4o LLM from OpenAI which performed comparably with several groups of specially trained and supervised human annotators. The LLM achieves similar performance to humans, between 0.91 and 0.97 on zero-shot prompts without feedback to the LLM. Reviewing the disagreements between LLM and gold standard human annotations we note that actual LLM errors are comparable to human errors in most cases, and in many cases these disagreements are not errors. Based on the specific types of annotations we tested, with exceptionally reviewed gold-standard correct values, the LLM performance is usable for metadata annotation at scale. We encourage other research groups to develop and make available more specialized "micro-benchmarks," like the ones we provide here, for testing both LLMs, and more complex agent systems annotation performance in real-world metadata annotation tasks.

KEYWORDS

large language models, metadata annotation, information extraction, human neuroimaging, ontologies, document annotation, text mining

1 Introduction

Scientific publications are highly stylized writings with strong rules and norms for formatting and arranging the information that they convey, unlike the freedom afforded to writings like web pages, business letters, and other working literature. This regularity allows scientifically literate readers to deftly read and process these publications. Despite these restrictions, scientific publications are generally free-text documents, and the freedom afforded to their writing allows an incredibly high degree of variability which limits the discovery of relevant publications within the broader research workflow. Basic text search (e.g., Google web search, PubMed, etc.) has often been ineffective as it has worked, historically, at the word and phrase (syntactic) level, which detaches the search terms from much of the surrounding (semantic) context. For example, a web search or searching a collection of documents for a term, such as "schizophrenia," will turn up many papers that mention schizophrenia in passing or even in the context of "without schizophrenia," and those papers limit the findability of papers which present relevant research results directly relevant to the topic.

Historically, human curation of the scientific literature has been required to overcome these limitations, but this approach fails as the scale of scientific publication increases (Tan et al., 2024; Yadav et al., 2024). Curation refers to the extraction or development of specific, context-relevant, information from publications and the organization of this information into "annotations" which we treat as metadata that are attached to the original publications (Tan et al., 2024). This metadata specifically allows programmatic methods for accessing and further processing this published literature. This "further processing" may include finding links to data attached to the publications, as well as processing the literature for systematic or quantitative reviews.

This literature curation and metadata annotation comes with a number of challenges. Perhaps the foremost of these is economics: outside of a few projects such as PubMed or corporate databases, there are very few resources available for detailed human curation of the scientific literature. But even if such resources existed, there are still problems. First, the task requires expertise in the relevant scientific area or specialized training of the annotators. Second, annotation requires substantial attention to detail, as well as high levels of focus, both of which are difficult for human annotators to achieve. Finally, the task is intrinsically monotonous which makes it difficult for people to maintain accuracy for more than a short period.

Large language models (LLMs) have disrupted artificial intelligence and machine learning research in various ways across multiple domains of use, including science (Fernandez et al., 2023; Maik Jablonka et al., 2023; Thirunavukarasu et al., 2023; Zhao et al., 2023; Nejjar et al., 2024; Sahoo et al., 2024). Within scientific research, LLMs are being used for a variety of tasks, but here we focus on their use in processing the scientific literature itself. Prior work has shown that LLMs can be used in several relevant natural language tasks such as summarization of both single and multiple documents or gathering materials for reviews (Lála et al., 2023; Agarwal et al., 2024), and annotation (Ding et al., 2023). Here we explore the capabilities of one of the most sophisticated of these models (GPT-4o from OpenAI) to do an annotation task traditionally done by human experts.

Given the impressive performance of LLMs in many tasks it seems obvious to try them on the annotation task. However, results have been mixed: Wadhwa et al. (2024) find that GPT-3 achieves state of the art performance in relation extraction from texts while Kristensen-McLachlan et al. (2023) find that LLMs are unreliable as annotators for problems in tweet classification. Aldeen et al. (2023) find that LLM performance in annotations vary widely by the specific annotation task tested. These results suggest that LLM performance on annotation will depend critically on two things: the specifics of each annotation task itself and the choice of LLM used to do the task.

When considering the capability of LLMs to do metadata annotation we must acknowledge that the human-made gold standard data, which we compare to machine annotation, is far from perfect. In a previous study (Sahoo et al., 2023), we used trained undergraduate students to do annotation of neuroimaging papers with metadata relevant to understanding the type of experimental data reported in these papers. Even with well-trained students and multiple quality control measures (using multiple annotators per annotation, more experienced students reviewing the work of junior students, multiple passes over the annotations, reviews by subject matter experts, etc.) there were many errors in the final set of annotations. Additionally, the co-developed Neurobridge ontology, like all non-probabilistic ontologies, has crisp boundaries which require arbitrary decisions in terminology. This leads to the inconsistent application of the terms when annotating.

We take the position that annotating publications with machine readable metadata extracted or generated from the unstructured text of the publications is sufficiently valuable to warrant its creation. One potential criticism of this approach could be that LLMs can interact with unstructured text directly, therefore what purpose does such metadata serve? While it is possible that LLMs could, for each new problem, process many publications directly, there are issues with this approach. First are the direct costs involved. Commercial LLMs do these sorts of tasks well, but open weight LLMs (such as the Llama series of models from Meta) do not currently perform as well, as can be seen from the various LLM leaderboards. Second, there are serious concerns about the environmental impacts, both energy and water usage, for the large commercial models (Strubell et al., 2019; Hisaharo et al., 2024; Jiang et al., 2024). If eventually open-weight LLMs could take on these tasks with less damage to the environment, there are still the compute costs and the direct work of deploying and maintaining these models. (Running an LLM, even at a moderate scale, is not as simple as downloading and running an executable). Therefore, minimizing the number of times a publication must be processed has intrinsic value. Finally, while LLMs are capable of sophisticated processing of publications, they are not instantaneous, so passing many publications through an LLM workflow is, and will likely remain for the near-term future, unacceptably slow for on-demand use.

Preprocessing to extract such metadata and curating the results as a searchable repository for the future enables reuse and repurpose, thus saving valuable resources. Further, the benefits of having such a repository are additive as it enables aggregation with additional metadata based on other future LLM extractions. We believe that a growing prompt-annotated metadata repository would be a very useful tool for researchers who can search for past annotated data and add to the repository with focused annotation prompts that they find relevant for their current research. Community annotated metadata using AI-prompts will be a valuable addition for future research and improve the metadata signatures for datasets with minimal effort.

Many neuroimaging papers describe research studies where the authors make their data available, but this is often done without putting the data directly into public repositories or making the data easily findable via internet search. The goal of the Neurobridge project (Sahoo et al., 2023; Wang L. et al., 2023; Wang X. et al., 2023) is to facilitate the ability to find neuroimaging data described in the published literature. With this goal in mind, the characteristics of a magnetic resonance imaging (MRI) dataset that we have annotated here address the question: "If I received this dataset from the authors, what would I receive?" Specifically, we focus on aspects of the data that allow either data reuse or which could be used to address further questions, such as: What types of MRI scans were collected? Who were the participants who were recruited? What were the parameters used in the scanner? Therefore, we are interested in descriptions of subject or participant populations, including diagnostic status; the structure of the experimental groups; anatomical MRI imaging parameters; and classification of functional MRI methods (e.g., resting state MRI versus task-based MRI); and detailed descriptions of task-based MRI, with classification into specific categories. The Neurobridge Ontology serves as standardized terminology that can be used as annotations. This list of questions will expand as Neurobridge continues.

We present examples of three specific targeted LLM annotation tasks in this paper that relate to data findability. We note that this same information is of value in the curation of papers for systematic reviews and meta-analyses. Specifically, we develop several "micro-benchmarks" that are directly relevant to our curation problem, and we evaluate the leading large language model on this annotation task. By "micro"-benchmark we mean an exceptionally well curated, therefore usually small, collection of test cases with gold-standard human extracted metadata. (We might call such extremely vetted data a "platinum" standard, but there is no term in current use for such data). We expect this collection of micro-benchmarks to expand as part of the ongoing Neurobridge research and encourage the development of additional benchmarks by other research groups using LLMs for these types of annotation tasks.

We provide the benchmarks we have developed, along with all testing materials and code for comparing model performance. All materials are available at this paper's GitHub repository, please see the Data Availability Statement. We expect this repository (or a linked one) to eventually contain additional test sets, papers, and labels, for other annotation targets and we welcome contributions from other research groups working on similar annotation problems.

2 Materials and methods

2.1 Experimental data

The text used in these experiments are the published texts of scientific papers (hereafter "publications") and several sets of human annotations treated as standards for comparison with the LLM results or used as performance comparisons. We used a subset of 186 open access full-text publications from the PubMedCentral (PMC) as our main collection of scientific papers in this study (see below). Several sets of highly curated labels, available on subsets of these publications, were used to jump start annotations as described below.

2.1.1 Annotations: targets

Human-assigned annotations were either available, partially available, or newly produced for the following aspects of the publications. The previously available annotations for these tasks are all taken from (2023).

Each of these sets of annotations defines a specific annotation task:

- Task 1: General Imaging Type - (25 publications) The 3 labels available for this annotation are broad categories of types of human neuroimaging: (1) T1 weighted anatomical images, (2) resting state functional MRI, and (3) task based functional MRI. This is the simplest category of labels we used and only requires classifying the general neuroimaging modalities reported in the publications.
- Task 2: Structural Imaging Parameters - (44 publications) These are the parameters used to collect T1 weighted images, such as magnetic field strength, TR, TE, etc. (see Table 1 for a full listing.) There are 12 of these parameters we identified as being the most used to describe anatomical imaging, but it is common to just give a subset of these 12, as some of the parameters can be derived from combinations of the others. Which parameters are explicitly stated in a publication is a choice by the authors and, while guidelines exist, individual presentations can be idiosyncratic. While more complex than general imaging type, this task requires finding the specific parameters that are listed in the publication and recording only those explicitly present. These annotations were prepared expressly for this research and full details of how they were made are given in the next Section "2.1.2 Annotations: process."
- Task 3: Experimental Group Information - (30 publications) From previous research we have trained annotator labels for the participant (subject) groups used in the research presented in each of the 30 publications. Specifically, there is a set of 41 diagnostic labels as potential annotations and the annotator identified which label is appropriate for each experimental group presented in the publication. Here we require the LLM annotator to determine the number of participants in each group as well, which the human annotators were not asked to do previously. These participant count annotations are new to this work.

These three annotation tasks present a variety of challenges for the annotation of neuroscientific publications.

Task 1, the general imaging type is the simplest task as it only requires recognizing that something is present in the text, usually presented in clear or highly regularized language, and only requires assigning a generic label to indicate that the thing's category is present. As such it constitutes a classical multilabel classification problem (Tsoumakas et al., 2006).

Task 2, identifying the structural imaging parameters is also a simple annotation case, only requiring copying values present in the text. Although the parameters for T1 structural images need to be stated explicitly in the methods of each paper, there has been a lack of standardization in how these parameters are reported, and some of the parameters can be derived from others. This creates a complex system of reporting where there are many different combinations of reported parameters that may imply the same underlying set of scanning parameters (see below).

Additionally, this requires a sensible grouping of the parameters by scan, as some parameters are present but have different values in different scans (T1, T2, etc.). For the annotations to be correct, the parameters reported must belong to the correct scan. While this is generally easy for humans, machines may confuse or combine parameters from different scans. Our pilot studies of this process on earlier OpenAI models such as GPT-3.5 and GPT-4 showed many errors of this type. We also note that our initial human annotations contained significant errors as well (discussed below), so again we emphasize that human annotations are never perfect without intense review, revision, and curation.

We selected 12 structural imaging parameters for extraction by the LLM (see Table 1). Every paper should report a subset of these parameters. We expect the LLM to simply copy the values if present in the paper and report the value as is, with one exception. That exception was "scan acquisition time" which, when present, we requested to be reported in units of seconds no matter how it was listed in the source publications. (Acquisition time was reported in the original publications in a wide variety of different formats and units). There are some minor differences in what was annotated by the student annotators and the LLMs, see the discussion of the annotation process (below) for more details.

Task 3, the experimental group information is the most complicated of our annotation tasks. Here we chose a set of publications where each publication had exactly one or two participant groups. The process requires that an annotator do the following:

1. Read the language used in the publication to describe the experimental groups and determine the psychiatric diagnoses that are present in natural language. Note that this language is not fully standardized across either area of research or publications. As one example of a standardization issue, we note that this language has changed over time: with the publication of the Diagnostic Standards Manual 5 (DSM-5) the separate diagnostic labels "alcohol dependence" and "alcohol abuse" were combined into "alcohol use disorder." So, as an example, the distribution of terminology is non-stationary.
2. The description in the publication must be mapped into the set of terms in our controlled vocabulary. Even assuming perfect reading of the publications and full understanding of the controlled vocabulary, it is possible for annotators to choose different terms given intrinsic vagueness of language.
3. The label for the group must be paired with the correct final count of subjects or participants. Any cases that might have the correct label, but which lack the correct count, are considered errors.

These operations are usually easy for people but have only recently been consistently achieved by LLMs. In our pilot studies, the GPT-3.5 and the first version of the GPT-4 models could not consistently do this task, often inventing new terms that were not in the controlled vocabulary ("hallucinating" in LLM terminology). Note also that in Sahoo et al. (2023), the students were not required to annotate the group sizes so we do not have a human comparison for these.

2.1.2 Annotations: process

2.1.2.1 Previous work: annotation tasks 1 and 3

Tasks 1 and 3, General imaging type and the experimental group information, were annotations from previous work. Briefly, both sets of annotations were made by trained (undergraduate) student annotators in several passes over the publications. During the first pass, students entered their choices into a spreadsheet and, when relevant, made recommendations for new ontology terms to the ontologists; the Neurobridge ontology was co-developed with the annotations. After consultations between the students and the ontologists, the draft ontology labels were entered into the spreadsheets. In a later phase, the raw text of the publications was marked up directly with the annotations in specially formatted files by a new group of trained students. During this latter phase the earlier work was checked by the new students and, finally, reviewed by "senior annotators" (students who were promoted from the annotation task or their supervisors) for correctness. See Sahoo et al. (2023) for full details.

As noted above, this process did not produce perfectly correct annotations. In fact, despite the extensive effort put into the process, many errors were present in the final collection of 186 publications annotated.

In this study, several senior authors reviewed, evaluated, and corrected the student annotations that were used in this study, producing a new much more extensively reviewed and corrected set of annotation labels for these two tasks. These were vetted repeatedly, and the final annotations were only accepted when both authors agreed. Additionally, after the LLM annotation, each of the disagreements between the humans and the LLM were reviewed again. This process was repeated until every disagreement between humans and LLM were accounted for; this yielded the final gold standard we now make available in this work. We note that this final set of gold standard annotations has been subjected to more extensive review than most "gold standard" evaluation data.

TABLE 1 The structural MRI parameters targeted for annotation from each publication (task 2).

MRI parameters targeted for extraction (12)	
T, tesla, magnetic field strength of the scanner	Voxel size, (1–3 numbers; mm; sometimes with exponents)
TR, repetition time (ms)	Matrix size, (usually 2–3 numbers, more rarely 1; unitless)
TE, echo time (ms)	*Slice thickness (mm)*
TI, inversion time (ms)	Acquisition time, (s; the only unit conversion in prompt)
Flip angle (degrees)	Number of slices (unitless)
FOV, field of view (1–3 numbers; mm)	*Image orientation (values: axial, sagittal, coronal)*

Graduate student authors labeled the annotations in Roman font, and a senior author labeled the items in *italics*. See the Section "4 Discussion" in the text for details.

2.1.2.2 New work: annotation task 2—structural imaging parameters

The structural imaging parameter annotations were made by the graduate student authors and reviewed by senior authors. The graduate students were required to label the set of 44 publications for 9 possible values (see Table 1). So, there were 9 × 44 = 396 annotation positions (potential parameter values) to fill in. Note that three of these labels were for FOV, matrix, and voxel size. Each of these parameters may be represented by 1, 2, or 3 numbers. The students reported the values as strings, not as individual numbers, so each of these count as a single annotation position. The LLM treated these 3 items as 9 distinct annotations, one for each number reported. This will affect the counts of potential annotation positions and the accuracy calculations in the discussion below. The graduate students each labeled about half of the publications annotated, then they switched and reviewed each other's work. Below we call this set of annotations (before review) the "graduate student annotations."

These annotations were reviewed and normalized by the first author. Normalization corrected for different encodings: some of the elements annotated by the students were cut-and-pasted while others typed, so different symbols, sometimes visually indistinguishable, were all corrected to specific standard symbols. Additionally, units were normalized (e.g., "10 degrees" and "10°" were both made the same) and generally removed from the student annotations and embedded in the format used to store the data (that is, either paired with a column name in spreadsheets or stored as a different data element in JSON formats, see task 2 below for the JSON). OpenRefine (version 3.8.2) was used as the primary tool for this cleaning process (Ham, 2013; Petrova-Antonova and Tancheva, 2020).

We note that during this review and normalization process the three remaining annotations were added: slice thickness, image orientation, and Tesla (magnetic field strength), which had not been part of the original annotations. The annotations are referred to as the "additional annotations" below. This led to an additional set of 132 (44 × 3) annotation positions.

Subsequently, all the annotations were reviewed by two senior authors. Note that by the end of this, some annotations have been through 5 passes: original annotator, second annotator, the first author (during normalization), a second senior author, then a repeated vetting by the first author, while the additional annotations have been through 3 passes (annotation by the first author, followed by review from another senior author and the first author a final time). We call all of these final, thoroughly reviewed annotations the "full human system annotations." Throughout this process, data was collected on human errors made in annotation at each stage; these will be discussed below. Note that we have three relevant stages to assess the accuracy of the annotations: the correctness of the graduate student annotations (before review) together with the correctness of the additional annotations (by the first author) before review, the annotations after being reviewed and updated by experts, and the correctness of this reviewed data during the post LLM review of differences. (Each of these can be considered a different stage of the development of the annotations.) The graduate student and additional annotations are comparable to what most papers call a "gold standard" data set, despite still containing many errors. The full human system annotations are much more deeply reviewed than is usual in this kind of work. Our new "gold standard" annotations are the final set that has been through all these processes.

This particularly intense review process yielded a set of annotations that are more correct than we have come to expect from most human annotated gold standard data sets. However, after LLM annotation, the entire set of human/LLM disagreements was reviewed again, which yielded 2 additional human errors that had made it through the entire process described above. We discuss the correctness of these sets of annotations in the appropriate section of the results below and we consider the costs involved there as well.

All of the final gold standard annotations are available in the GitHub repository for this project, see the Section "Data availability statement."

2.1.3 Publication texts details

LLM usage. The raw text of the publications given to the LLM in two of our experiments (tasks 2 and 3) were the BioC texts of the publications (Comeau et al., 2019). These BioC formatted papers were further processed to provide the plain text of the papers with footnotes, citations, and other technical apparatus removed. Due to context window size limitations of earlier LLMs, only the content of the papers to the end of the Sections "2 Materials and methods" were provided to the LLMs. This included: titles, abstracts, introductions, and methods sections; removing the results, discussion, and other latter parts of the papers. The sections were determined from the metadata provided in the BioC format. For task 1, general imaging type, the LLM was given the PMC provided PDFs rather than the BioC text. This task's workflow was dependent on the LLM provider's internal process for converting the PDFs into usable text for the LLM and, as such, we do not have access to the details of how this was done. We note that there were two runs which generated failures in parsing the PDFs. This appeared to be an intermittent fault and in both cases repeating the analysis in a new chat session resolved the problem with no changes. See below for more details (Section "2.2.1 LLM: GPT-4o").

Human usage. For annotation tasks 1 and 3, the human annotators had access to the full (all text sections) BioC text for their final markup, but they were not restricted to using only this text and were free to review PDF versions of the papers. For the structural imaging parameters task, the human annotators used the PMC provided PDFs as their texts.

2.2 Large language model and prompting strategies

2.2.1 LLM: GPT-4o

The present study focuses exclusively on the GPT-4o (omni) LLM, the 2024 flagship product from OpenAI, although we developed many of the experimental tasks on earlier models (GPT-3.5 and GPT-4 at various checkpoints). It is used as the reference model as it is an extremely capable model with a reasonable price point for use in research applications. As the goal of this paper is to establish whether the most powerful models available can reasonably solve these tasks, we did not attempt to systematically explore all currently available models but instead established a baseline against one of the most prominent current commercial models. Therefore, the results here may not fully generalize to other large commercial models. As there are new models released continuously, overall LLM model performance is a rapidly moving target. We provide code in the paper's GitHub repository that can be modified to enable testing other LLMs with our annotations.

We used two methods of submitting our prompts to the LLM. Task 1 was run manually via ChatGPT (OpenAI's chat interface), selecting the GPT-4o model and uploading the PMC-provided PDFs of the publications for analysis. Each paper was run with the prompt in a fresh chat session, using the default parameters for ChatGPT. Accessing GPT-4o via the chat interface precludes knowing the specific model version or "checkpoint" used; however all of these were run on the same day in August 2024, so they should all have used the same version of the model. The other tasks (2 and 3) were run through OpenAI's API, which allows direct programmatic access to the models via functions in a corresponding Python library released by the company.[1] The model used in those tasks was the GPT-4o LLM (version checkpoint: gpt-4o-2024-08-06). In these experiments, the raw BioC text of each paper in turn was appended onto the end of the prompt for the task and submitted for processing. As LLMs are stateless, this approach is the equivalent of using a fresh chat session as in the manually run experiment.

For these tasks, all LLM parameters were set to their defaults except for the "temperature" parameter. Temperature appears to be related to LLM response variability, although this claim is contested (Renze and Guven, 2024). Others commonly claim that it is a response "creativity" parameter, but this is also unlikely (Peeperkorn et al., 2024). For task 1 the temperature was left at the default for the ChatGPT interface. For task 3, the temperature was set to 0.2 (low), based on informal advice we received, but we did not explore other settings. For task 2 we had to set the temperature to zero to fix a specific problem. This task used the most complex JSON prototype of all the tasks. When the temperature was greater than zero, the LLM would often return simplified JSON, that is, it would pick out the annotations present in the publication and return only the JSON fields corresponding to the parameters explicitly reported. The JSON fields that corresponded to parameters not present would often be elided in unusual ways that were problematic for post-processing. Setting the temperature to zero made the LLM always return the exact same JSON each time, just with the JSON nulls replaced with values for any parameters present in the publication. All the other parameters remained present in the JSON and set to null. This is the behavior that we wanted. For the other tasks, no temperature changes were needed to maintain the expected JSON format. We take the position that JSON should always remain fully present (i.e., don't delete fields for information not available) to simplify the postprocessing of the JSON returned from the LLMs.

2.2.2 Prompting strategy

Given that we were using a capable leading edge LLM, we explored the quality of annotations under only the simplest possible usage (Schulhoff et al., 2025). Specifically, we used "zero-shot" prompts for these tasks (Reynolds and McDonell, 2021; Li, 2023). Prompts are the text that specify the task that the LLM is supposed to perform (Bhandari, 2024). A zero-shot prompt is one in which instructions for the task are given, but no specific examples are included. Performance with a zero-shot prompt can be thought of as a type of baseline performance or naive prompting strategy. If there are annotations that do not perform well under zero-shot prompts, so-called "few-shot" prompts can be used to improve performance. These prompts include both the task instructions along with one or more examples. If there are known cases that are challenging for the LLM, these problems can often be solved by adding examples that specifically cover the most challenging cases. For more complex annotation situations there are sophisticated systems, such as DSPy (Khattab et al., 2023a,b), that exist to automate few-shot prompt development with examples.

We selected the zero-shot approach for several reasons, the main one being that this approach is what most people would try first when faced with a new annotation problem. Current commercial LLMs are quite capable when used with zero-shot prompts, and any annotation problem which requires going beyond simple prompting strategies will likely require addressing specifics that may be unique to the class of annotations involved. Perhaps obviously these procedures are not perfect. However, human annotation has its own problems as noted above. No method, either human or automated, will ever annotate perfectly except in situations of extensive review and curation, but we find that the types of errors GPT-4o commits under this strategy are comparable to human annotation (particularly when that human annotation is given reasonable review and verification).

Additionally, our prompts contained JSON prototypes of varying complexity that guided the LLM responses in each annotation problem. See Figure 1 for the most complex example. We recommend using more detailed JSON, like that for task 2 here, as such prototypes contain information that improve LLM performance in annotating publications. The handling of the JSON format by LLMs has dramatically improved over the course of this study. The most recent OpenAI models offer a new feature, called "structured outputs," that allow for more fine-grained control of the JSON returned from the LLM.[2] We did not use these new features as they became available too late in the course of this study. All the outputs from the models presented here used only the more basic JSON-mode provided by OpenAI for their models along with in-prompt instructions to return JSON as the response.[3] Despite using this older and more limited control over the LLM output, we still achieved excellent results.

A system prompt is text appended to the start of the LLM session which sets a context for what follows. We used such prompts for tasks 2 and 3, and the full text of these prompts can be found in the GitHub repository. In each of these system prompts a persona for the LLM is defined which prompts the LLM to act as a "helpful assistant" and contains a directive that the LLM is "expert" in the annotation task at hand. Additionally, the LLM is given the directive to be "careful, thorough, and brief" in responses. Finally, the LLM is instructed to respond using JSON only, following the prototype provided. All of this follows industry practice for LLMs.

[1] github.com/openai/openai-python

[2] platform.openai.com/docs/guides/structured-outputs

[3] platform.openai.com/docs/guides/structured-outputs/json-mode#json-mode

```
{
  "sequence_type": "T1-weighted",
  "T": {
    "description": "Scanner Field
Strength",
    "unit": "Tesla",
    "value": null
  },
  "TR": {
    "description": "Repetition Time",
    "unit": "ms",
    "value": null
  },
  "TE": {
    "description": "Echo Time",
    "unit": "ms",
    "value": null
  },
  "TI": {
    "description": "Inversion Time (TI)",
    "unit": "ms",
    "value": null
  },
  "flip_angle": {
    "description": "Flip Angle",
    "unit": "degrees",
    "value": null
  },
  "FOV": {
    "description": "Field of View",
    "unit": "mm",
    "value": {
      "x": null,
      "y": null,
      "z": null
    }
  },
  "voxel_size": {
    "description": "Voxel Size or In
Plane Resolution",
    "unit": "mm",
    "value": {
      "x": null,
      "y": null,
      "z": null
    }
  },
  "matrix_size": {
    "description": "Matrix Size",
    "value": {
      "x": null,
      "y": null,
      "z": null
    }
  },
  "slice_thickness": {
    "description": "Slice Thickness",
    "unit": "mm",
    "value": null
  },
  "acquisition_time": {
    "description": "Acquisition Time",
    "unit": "seconds",
    "value": null
  },
  "number_of_slices": {
    "description": "Number of Slices",
    "value": null
  },
  "image_orientation": {
    "description": "Orientation of the
Scan",
    "value": null,
    "allowed_values": [
      "axial",
      "sagittal",
      "coronal"
    ]
  }
}
```

FIGURE 1
JSON prototype for the structural MRI parameters task (task 2). Each entry corresponds to one T1-weighted structural MRI parameter, and includes the standard units used for that parameter to guide the LLM response. The LLM returns this JSON in its entirety with any parameters present in the text replacing the relevant "null" values under the "value" keys. Parameters not explicitly listed in the publication text are left as JSON nulls. Note the use of allowed values for "image orientation."

For task 1 we did not set a system prompt, using instead the default provided by ChatGPT.

2.2.3 Evaluation

The tasks described here are evaluated in two important ways.

First, the tasks are analyzed using simple accuracy (the ratio of correct annotations, including "not present" or "not applicable" as required, to the count of all possible annotation positions). However, given the high accuracies reported here (all greater than 90%), we focus on comparing these numbers with human performance for calibration. As our claim is mere comparability, and LLM performance equals or exceeds human performance in all cases presented, we do not discuss any statistical analysis in the paper. We have provided some additional statistical analyses in the Supplementary material.

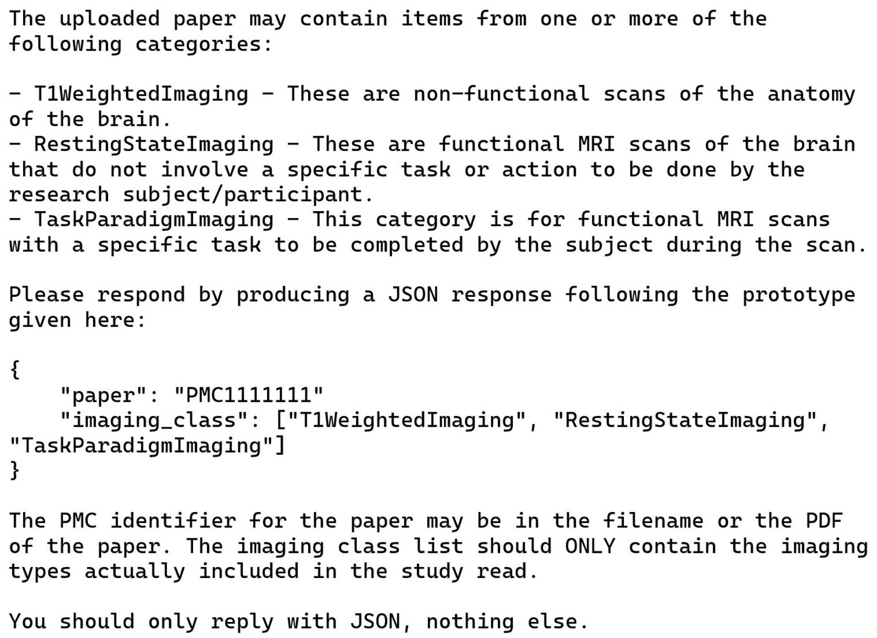
Second, for each task, there is a qualitative analysis which reviews each of the disagreements between the LLM and the human gold-standard data. It is known that LLMs occasionally do better at annotating than humans do (Nahum et al., 2024). We also find this in our results. This includes both finding additional errors made by humans that slipped through review and occasionally finding that the LLM annotations are better organized than the human work. Note that with this review we were compelled to change our conception of our gold standard, producing a higher quality final product than previous human-centered annotation processes have produced.

3 Results

3.1 Annotation task 1: general imaging type

This task used the prompt given in Figure 2. An issue arose in this case: for nine of the papers the given prompt failed to generate correct answers. For these papers, a new run was started and a prompt that said, in its entirety: "Please summarize this paper" was given. Then, after ChatGPT generated the summary of the paper, the original prompt was repeated verbatim; and this action substantially improved the performance of ChatGPT. Note that this is a chain; these two prompts were run in sequence in one session. This action improved the results even though the summaries were not relevant to the errors present in the initial response. Adding this summary prompt before the papers that worked initially did not change their annotations, so a fully automated process for this task via the API is trivial to implement.

The scores for this manual process are presented in Table 2. We scored the LLM based on both the initial response to the manual prompt and again with this prompt being repeated after generating a summary. The initial LLM responses for the publications analyzed are presented in the column "LLM without summary." Simple accuracy here is only 84.0% reflecting 12 errors: 8 omissions (all "T1 Weighted Imaging") and 4 false positives (all "Resting State Imaging") each of these divided by 75 possible annotation positions (3 annotation positions per document × 25 documents). After summarization, this was reduced to two errors, both false positives for "Resting State Imaging." Both the LLM with summaries and human student annotators scored an accuracy of 97.3% on 2 errors



```
The uploaded paper may contain items from one or more of the
following categories:

- T1WeightedImaging - These are non-functional scans of the anatomy
of the brain.
- RestingStateImaging - These are functional MRI scans of the brain
that do not involve a specific task or action to be done by the
research subject/participant.
- TaskParadigmImaging - This category is for functional MRI scans
with a specific task to be completed by the subject during the scan.

Please respond by producing a JSON response following the prototype
given here:

{
    "paper": "PMC1111111"
    "imaging_class": ["T1WeightedImaging", "RestingStateImaging",
"TaskParadigmImaging"]
}

The PMC identifier for the paper may be in the filename or the PDF
of the paper. The imaging class list should ONLY contain the imaging
types actually included in the study read.

You should only reply with JSON, nothing else.
```

FIGURE 2
Manual zero-shot prompt for general imaging type task (task 1). This text was pasted into the chat window after uploading the publication PDF. See text for details.

TABLE 2 Results for the general imaging type annotation task (task 1).

Parameter	*LLM without summary*	*LLM with summary*	*Student annotators*
Accuracy	84.0%	97.3%	97.3%
Fraction of perfect labeling	14/25	23/25	23/25

The LLM was tested under two conditions: once with just the basic prompt and a second time with this prompt following a request for the LLM to summarize the publication. See text for details.

for each (see below for details of the errors). It is reasonable to expect these labels to be among the most easily assigned, so this result with LLM and humans tied is not surprising. It also makes the point that human annotations are never perfect. The second row of the table is the fraction of "perfect" cases out of the total number of cases.

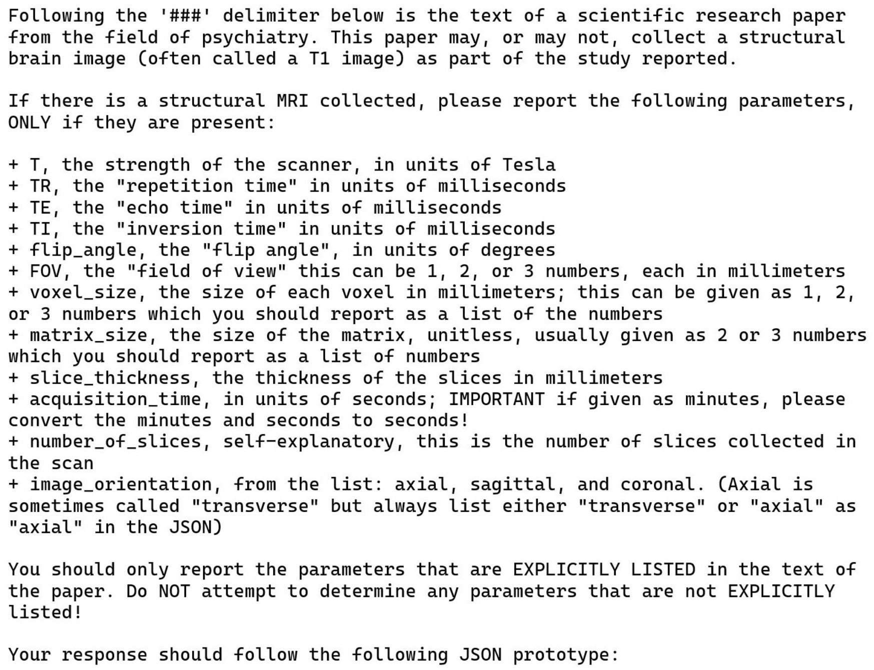
3.1.1 Qualitative analysis

The student annotators are scored as committing two errors, one false positive and one false negative for T1 Weighted Imaging. One of these is correct but based on information not included in the actual publication analyzed (the information was included in a supplement to the publication). As such it is unfair to score the students as wrong here (however in the ground truth this was set to "no T1 Weighted Imaging" to accurately reflect the contents of the publication text). The other error that the human annotators made was missing an oblique reference to T1 imaging in one paper. However, it is worth noting that this paper did not follow standard reporting practices, so the reference was easy to miss, although the LLM did both find this obscure reference and correctly interpret it.

The LLM (with summarization) committed two errors as well, both of which were the inclusion of the "Resting State Imaging" indicator when this type of imaging was not present (false positives). Without summarization beforehand, the LLM made this error four times, summaries resolved two of these. In the non-summary condition, the LLM also made eight errors (false negatives) for the T1WeightedImaging label. Given the prior probability of structural imaging being included in almost any given fMRI study this seems strange. Also, as noted, none of the generated summaries included any mention of structural imaging, so it is not clear why these summaries improved the performance in so many cases.

3.2 Annotation task 2: structural MRI parameters

The prompt and the JSON prototype for this task are presented in Figures 1, 3. Note that these are shown separately for the presentation here due to their size but were combined with each other and with the publication text to be analyzed into a single block of text when sent to the LLM via the API. The instructions in Figure 3 simply list the parameters to be collected, along with a one-line definition of what the parameters mean, and an admonition to only list parameters explicitly present in the publication. This example uses a more complex JSON format than our other tasks in that it also includes embedded information to guide the LLM toward a correct solution and alignment with the instructions. Each of the items shown in Figure 1 includes a brief informative description and the most common units of measurement used for



```
Following the '###' delimiter below is the text of a scientific research paper
from the field of psychiatry. This paper may, or may not, collect a structural
brain image (often called a T1 image) as part of the study reported.

If there is a structural MRI collected, please report the following parameters,
ONLY if they are present:

+ T, the strength of the scanner, in units of Tesla
+ TR, the "repetition time" in units of milliseconds
+ TE, the "echo time" in units of milliseconds
+ TI, the "inversion time" in units of milliseconds
+ flip_angle, the "flip angle", in units of degrees
+ FOV, the "field of view" this can be 1, 2, or 3 numbers, each in millimeters
+ voxel_size, the size of each voxel in millimeters; this can be given as 1, 2,
or 3 numbers which you should report as a list of the numbers
+ matrix_size, the size of the matrix, unitless, usually given as 2 or 3 numbers
which you should report as a list of numbers
+ slice_thickness, the thickness of the slices in millimeters
+ acquisition_time, in units of seconds; IMPORTANT if given as minutes, please
convert the minutes and seconds to seconds!
+ number_of_slices, self-explanatory, this is the number of slices collected in
the scan
+ image_orientation, from the list: axial, sagittal, and coronal. (Axial is
sometimes called "transverse" but always list either "transverse" or "axial" as
"axial" in the JSON)

You should only report the parameters that are EXPLICITLY LISTED in the text of
the paper. Do NOT attempt to determine any parameters that are not EXPLICITLY
listed!

Your response should follow the following JSON prototype:
```

FIGURE 3
Zero-shot prompt for the structural MRI parameters task (task 2). This prompt describes the potential MRI parameters to be returned. It provides a limited vocabulary for "image orientation" and the expected units for all other parameters. Also, it tells the LLM to convert scan acquisition times to units of seconds, no matter how this was originally reported in the publication.

each parameter. Although occasionally the publication text strays from these units, the LLM successfully mapped the parameters in the free text into the JSON with units used correctly. We did not ask the LLM to do any unit conversions, except for the "acquisition time" of the scans, which was reported in a wide variety of ways in the original publications. We prompted the LLM to convert any different time representations into seconds no matter how they were reported (Figure 3). This is the only explicit demand for unit conversion. Amazingly, given the wide variety of ways that acquisition time was written in the publications, the LLM was successful in converting and recording these for all but one case (see below).

Out of 792 annotation positions in the JSON (44 publications × 18 possible values) the LLM annotations did not match the gold standard annotations 35 times, for an overall (preliminary) accuracy of 95.6% under the assumption that the gold standard is fully correct. However, not all mismatches will turn out to be wrong, see the next Section "3.2.1 Qualitative analysis."

We reiterate that most of the values to be filled in are nulls, but we consider missing a value that is present to be as problematic as filling in a null that is not present. We also note that both the LLM and the humans made both types of errors.

3.2.1 Qualitative analysis

As the number of mismatches was small, we reviewed all 35 of them to determine the differences between human and LLM annotation. We include the PMC ID numbers and citations for any specific papers we review in detail.

The first group of mismatches are differences from the gold standard that are, in fact, correct. In one paper, PMC5037039 (Janes et al., 2016), we discovered that the LLM reported the Tesla rating of the scanner as 2.89T, not 3T as reported by our human annotators. In fact, both values were listed in the paper, with the 2.89 being more precise, so the response from the LLM is correct. In the JSON prototype we require 1, 2, or 3 values for voxel size. We note that the human annotators just copied these numbers from the publications verbatim (as strings) without much consideration (this is what they were told to do). The LLM recognized that papers reporting "1 mm^3 isotropic" for voxel size corresponded to 3D voxels with values of (1 mm, 1 mm, 1 mm), and returned the unpacked x, y, z values appropriately. This may be considered an error or not, depending on the interpretation or the goals of the annotation process, but this is a correct interpretation of the notation. This specific difference occurs in 8 papers and accounts for a total of 16 of the differences with the gold standard. Combining these with the previous difference for Tesla rating, changes the mismatches from 35 to 18 yielding a revised accuracy of 97.7%, if we prefer these changes to our original annotation scheme. The annotation supervisors were impressed with the LLM's recognition and correct use of the "isotropic" notation and noted that it was a poor choice in the design of the human annotation process not to require a similar unpacking from the student annotators.

The second group of mismatches are related to errors or ambiguities in the original papers. In one paper, PMC6031869 (Hua et al., 2018), the authors list the "in plane resolution" or voxel size as "231 × 232" which our expert considered most likely to be the values for the matrix. This is what the LLM assigned those values to; the student annotators did not. In PMC6551253 (Chumin et al., 2019), the text contains the odd expression. "Field of view = 192 × 168 matrix. . ." which the LLM labeled as FOV, but which our expert believes should be the matrix value. However, without expertise this is a hard case to resolve, although here our student annotators did choose the expert's preferred solution. Paper PMC6677917 (Lottman et al., 2019) stated "base resolution = 256" which our expert determined implies a matrix value of 256 × 256. The LLM reported this as a matrix with one value (x) of 256; the student annotators ignored this information entirely. One mismatch was due to the PDF to text conversion of one of the publications. In PMC6491039 (Sawyer et al., 2019), the value for flip angle was listed as 7 degrees, but the typesetting of the article used a superscript zero character rather than a degree sign, and our raw text processing turned this into "70," which is what the LLM read and reported, instead of 7° as was intended. This is an error of the preprocessing of the file, not an error introduced by the LLM itself.

The final group of mismatches are definite errors on the part of the LLM. In PMC6289814 (Kim et al., 2019), the LLM took the FOV values from a different scan. In PMC6104387 (Hahn et al., 2017), the LLM copied the FOV and the matrix values from an unrelated T2* image. Finally, in PMC6491039 (Sawyer et al., 2019), the LLM missed an explicit mention of the acquisition time.

In summary, while there are clear errors present in the LLM annotations, they are rare and the general types of errors made are not substantially different from the sorts of errors made by the human annotators.

3.2.2 Curator/Annotator detailed comparison

As described in the Section "2 Materials and methods" we have detailed information available about these annotations for the following stages of the annotation process:

1. The graduate student annotations – These are the annotations made and reviewed by the graduate student annotators. Errors here were discovered both during review and normalization stage and during the final (pre-LLM annotation) review.
2. The additional annotations made by the first author – These annotations were made during the review and normalization stage. Errors here were discovered during the final (pre-LLM annotation) review.
3. The full human system annotations – These are the annotations that exist at the end of the entire human process, specifically after all annotation, normalization, and review passes listed above. Errors here were discovered when reviewing the individual LLM mismatches (post-LLM annotation).

We compare the accuracy of these cases here. Results of these comparisons are summarized in Table 3.

The graduate student annotations required labeling 44 publications for 9 possible values. For this set, there were 9 × 44 = 396 potential annotation positions to fill in. The student annotators made 23 total errors for an accuracy of 94.2%. The additional annotations had a similar accuracy to the graduate students: three additional annotation fields across 44 publications created a total of 3 × 44 = 132 positions, and there were 7 errors, an accuracy of 94.7%. We note that for this task, the basic LLM accuracy of 95.6% is superior to either of the human annotators (94.2% and 93.9%). If the reasonable choices of the LLM above are accepted as correct, then the LLM performance becomes 97.7%, substantially better than the coordinated and time-consuming work of three human annotators.

In addition, two errors relating to the matrix values made it all the way through both student annotation and curator review and were only discovered during the detailed review of the LLM performance. This last represents the performance of the full human annotation system as described above and in the methods. This system achieved an accuracy score of 99.6% (2 errors in 12 × 44 = 528 annotation positions) with a cost of at least 60 total person-hours of work. LLM annotation of the same texts and annotation positions cost approximately 0.90 USD and took approximately 15 min. Total LLM process development time took approximately 3 h for programming, testing, etc. This comparison is not completely fair as much of the conceptual development took place during the human labeling process and normalization, but these figures provide a reasonable first estimate. It is worth noting that even if human review is used, using the LLM for the first annotation passes would still represent a significant savings.

3.3 Annotation task 3: experimental group information

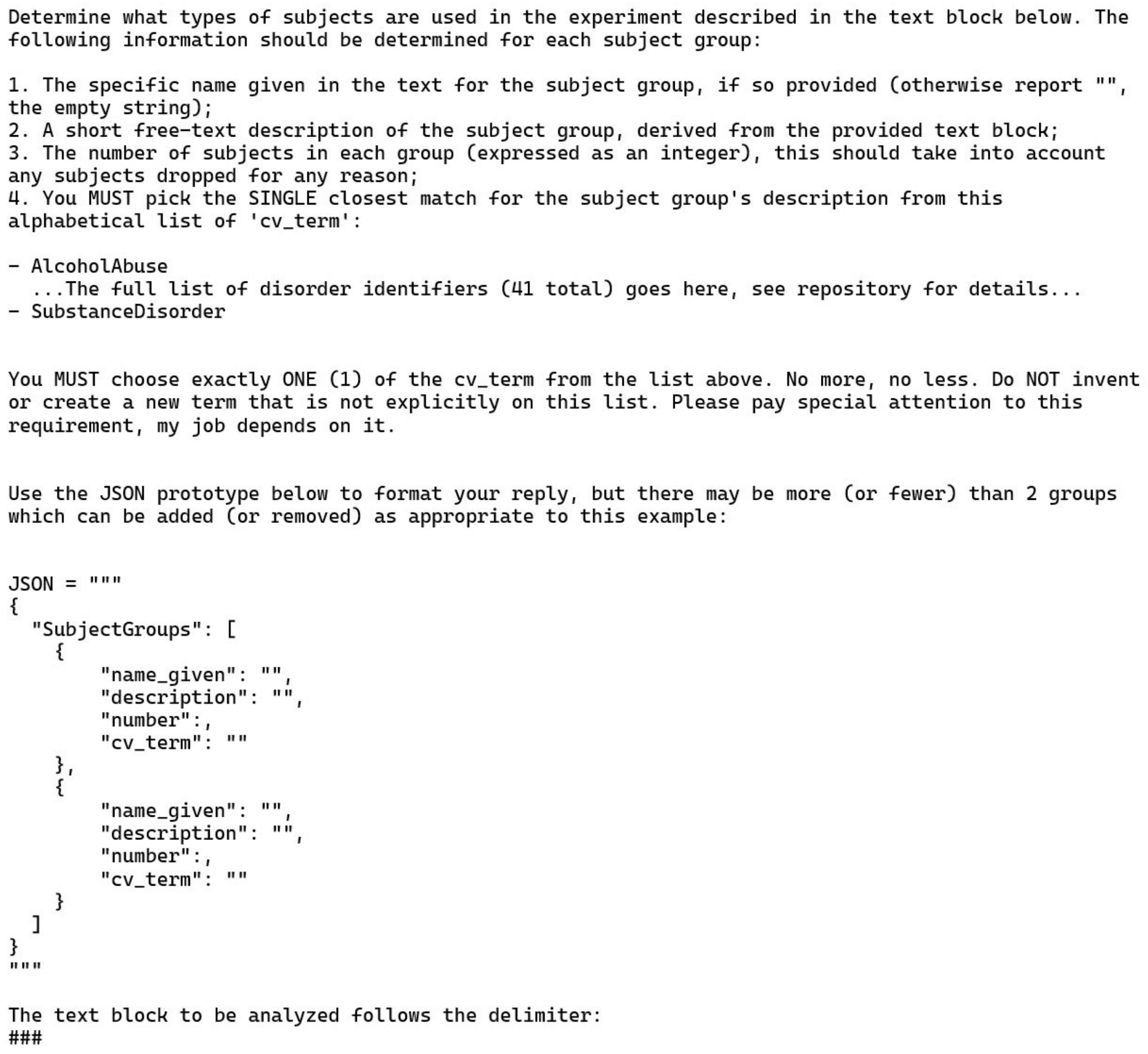
In this task the LLM was required to determine the psychiatric diagnosis and the correct final count of research participants in each experimental group present in each publication. Performance in this task is less accurate than for the other tasks, for both human annotators and the LLM. The task has the most intrinsic ambiguity as the ontology terms used are for psychiatric disorders and these often have intrinsic variability even among expert users. We note that there was variability among the ontology experts both for the exact usage of each term and which term to use for the annotation of each publication.

Each of the 30 publications used in this task had either exactly one or exactly two experimental groups (4 and 26 publications, respectively). Furthermore, each experimental group in this collection of publications had a single diagnostic label; papers with groups of combined diagnoses were excluded. This means that the LLM was required to fill 120 annotation positions, comprising 60 pairs (diagnosis and count), with each pair describing one experimental group. For the 4 publications with a single group, the

TABLE 3 Results for structural MRI parameters (task 2).

Annotator type	*LLM*	*Graduate student annotations*	*Additional annotations*	*Full human system annotations*
Accuracy	95.6% (97.7%)	94.2%	94.7%	99.6%

Accuracy of each human annotator group, annotation by LLM alone, and the complete "human system" which includes multiple review processes. See text for details of each column. The value for LLM accuracy in parentheses is the revised score for extra annotations that were correct but which did not agree with the original gold standard data.



```
Determine what types of subjects are used in the experiment described in the text block below. The
following information should be determined for each subject group:

1. The specific name given in the text for the subject group, if so provided (otherwise report "",
the empty string);
2. A short free-text description of the subject group, derived from the provided text block;
3. The number of subjects in each group (expressed as an integer), this should take into account
any subjects dropped for any reason;
4. You MUST pick the SINGLE closest match for the subject group's description from this
alphabetical list of 'cv_term':

- AlcoholAbuse
  ...The full list of disorder identifiers (41 total) goes here, see repository for details...
- SubstanceDisorder


You MUST choose exactly ONE (1) of the cv_term from the list above. No more, no less. Do NOT invent
or create a new term that is not explicitly on this list. Please pay special attention to this
requirement, my job depends on it.


Use the JSON prototype below to format your reply, but there may be more (or fewer) than 2 groups
which can be added (or removed) as appropriate to this example:


JSON = """
{
  "SubjectGroups": [
    {
        "name_given": "",
        "description": "",
        "number":,
        "cv_term": ""
    },
    {
        "name_given": "",
        "description": "",
        "number":,
        "cv_term": ""
    }
  ]
}
"""

The text block to be analyzed follows the delimiter:
###
```

FIGURE 4
Zero-shot prompt and JSON prototype for the experimental group information task (task 3). Note that the version of this prompt shown here has had the full list of ontology terms elided. This list can be found in the GitHub repository (see Section "Data availability statement"). The use of "emotional" stimuli words ("Please pay special attention to this requirement, my job depends on it.") was a previously recommended procedure that is not used as often today. See text for details.

other group should be left blank by the LLM; anything other than an empty response would be an error of equal weight. For these four groups null values should be given for both diagnosis and count.

The prompt for this task is shown in Figure 4. This prompt is like the previous examples: it is zero-shot and provides a JSON prototype to guide the LLM. The list of psychiatric diagnoses comes from the Neurobridge ontology, which is available in the Neurobridge GitHub repository as an OWL file. (see Section "Data availability statement" for details). Note that the figure elides the list of 41 diagnoses that were given to the LLM, but the full prompt is available in the GitHub repository as well. The text of the publication was appended to the end of the prompt after the "###" delimiter which follows OpenAI best practices.[4]

Additionally, to emphasize the instruction to choose just one item from the list, we added the phrase "Please pay special attention to this requirement, my job depends on it." These sorts of phrases were commonly used in prompt design (Li et al., 2023), although their importance has dropped off with more recent models.

The accuracy of the LLM in this task is 90.8%, with 11 total mismatches across 30 publications. We note that disagreements tended to bunch together by publication: 26 of the 30 publications were labeled in perfect agreement with the human results. All errors were in just four publications, as described in Section "3.3.1 Qualitative analysis" (below).

There was no direct independent analysis of human annotation performance available for comparison with the 30 publications used for LLM annotation in this task. However, as part of developing the materials for this study, we reviewed 39 different publications labeled by humans as part of our prior work in

[4] help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api

Sahoo et al. (2023). All these 39 publications were reviewed and given definitive labels by the present authors with all final decisions made by a senior author who is a co-developer of the Neurobridge ontology. These final annotation decisions were made without knowledge of either the human or LLM annotation results. While we recognize that it is impossible to resolve all ambiguity in any ontology, this approach allowed us to develop a fully expert-labeled gold standard data set that could be compared against both previous student work and LLM performance. Our review revealed many errors remaining in the human annotation results from Sahoo et al. (2023).

Specifically, in the 39 publications reviewed that were human annotated there were 13 errors found, but this was only noted at the level of an individual publication and not at the experimental group level. That is, a publication was marked as being in error if any number of mistakes of annotation were made for it. Almost all these papers were two-group research studies, although a few had more groups, so the human annotators would have had to fill in labels for 78 participant groups minimally (2 groups × 39 publications). It was rare for multiple errors to be made in a single paper. This conservatively estimates the human annotation accuracy at approximately 83.3% (13 errors across 78 annotation positions). Also note that the collection of human annotations reviewed was not selected randomly but was based on the topics of the publications. We were specifically selecting publications connected to drug abuse topics which was a subset of the original collection. There is no obvious reason to suppose that our human annotators would make more or fewer errors for other topics. Table 4 summarizes this comparison.

3.3.1 Qualitative analysis

Reviewing the mismatches between the 30-case gold standard and the LLM for this task we note the following: In the publication PMC5086261 (Forster et al., 2016), the LLM chose the annotation "substance disorder" in place of "drug abuse" and it is not clear that this difference is an actual error as these terms overlap significantly. In two cases, PMC4990879 and PMC6704377 (Chang et al., 2016; Vanes et al., 2018), the LLM appeared to ignore any control groups and split the group of interest into two, while giving each split group the same diagnostic label. We have no explanation for this strange error, as looking at the original source publication revealed no obvious reason to split these groups. In another case, PMC6215331 (Alloza et al., 2018), the LLM correctly identified one group and ignored another. The last error is hard to explain because when we took the relevant text from the paper and pasted it into a ChatGPT session (GPT-4o) with the same prompt, ChatGPT correctly identified the two groups (which happened to both be groups with "no known disorder," or control groups). As the chat interface does not report the exact model used, this might reflect a change in the underlying model, or it may be due to the stochastic nature of LLMs.

TABLE 4 Accuracy comparison between human annotators from Sahoo et al. (2023) and the LLM annotation from this study for the experimental group information task (task 3).

Labeler	Accuracy	Publication fraction correct
LLM	90.8%	26/30
Human (student annotators)	83.3%	26/39

Note that the collections of publications annotated are different and that the human annotation performance is a conservative estimate (it is possible that it undercounts human errors). See text for details.

One additional publication is also worth noting: PMC6906591 (van den Heuvel et al., 2019), although the LLM was scored as correct for this case. In this work, all the human annotators, as well as the more senior reviewers, identified two specific groups as the correct answer. This was because these two groups were the only new data collected in this publication. However, this paper has 23 total groups in it, the additional 21 being validation datasets (usually with cases and controls both, which we count as separate groups) that went unnoticed by the original human annotators. The senior reviewers noted the additional validation data but agreed that the annotators should only have listed the groups with new data, as the other data came from other publications. The LLM initially found and annotated all the groups present in the study completely correctly. It was scored only against the two groups the senior reviewers and original annotators had found. We were genuinely surprised that the LLM correctly annotated all these groups, especially as some of these groups were not human but primate data. We note that the prompt above does not limit the LLM to original data, so strictly this is not an error by the LLM but in the prompt it was given.

4 Discussion

We describe and evaluate three real-world biomedical research annotation tasks where experimental metadata from neuroimaging research publications is collected. We show that one current flagship commercial LLM can annotate this material as well as or better than human annotators. This performance is consistent with informal AI industry assessments for other, very different, metadata annotation tasks. Even in our most challenging case, the annotation of experimental group information where the textual descriptions are highly variable and the annotations themselves contain substantial ambiguity, we achieved >90% annotation accuracy with 87% of publications annotated perfectly according to a highly refined human standard. We note also that across these example tasks many of the LLM disagreements were not errors but instead reasonable and explainable responses.

Two main results are shown by this research. First, the real-world tasks here provide three new micro-benchmarks, as defined in the introduction, for the evaluation of systems that annotate free text with metadata that makes explicit ideas and concepts embedded in the publication text. Given the variable nature of annotation problems, assessing the performance of LLMs will require many more similar, real-world annotation tasks to fully calibrate our understanding of LLM annotation performance. Such benchmarks require much more intensive curation and review than is common in previous machine learning research as the evaluation of LLMs has shown new complications in the annotation task. Additionally, a broad collection of such applied benchmarks will be required due to the unique aspects of each annotation situation as there is no theoretical or statistical guarantee of results available for this sort of work. Our project demonstrates the first such micro-benchmarks, and we anticipate more and encourage others to make similar high-quality benchmarks available for use by the research community. An important aspect of this project is not just the results of our annotations, but the methodology identified as a way for others to follow in annotating their domain specific metadata extraction using current and future LLM systems.

The second result is that simple zero-shot prompts can get the current large commercial LLMs to do metadata extraction for cases similar to the ones illustrated here with high accuracy. We agree with informal industry assessments that large commercial LLMs are broadly capable and can perform human annotation tasks with only modest prompt engineering required. More challenging cases may require specialization of the LLM to the task at hand, either through few-shot prompting or some form of fine-tuning, which at a high level are similar processes (Khattab et al., 2023b; Soylu et al., 2024). This approach follows the principle that it is relatively simple to achieve a solution for the bulk of cases, while more effort is often needed to solve the edge or corner cases.

It is notable that the various publications do not always follow the field's recommendations for reporting neuroimaging parameters (Nichols et al., 2016) which hinders their interpretation. This makes the annotation problem more complicated, effectively asking the LLM or human annotator to identify what the authors intended rather than what was reported. An additional challenge is the inherent ambiguity of words, even technical vocabulary. Psychiatric labels have ambiguity both due to changing diagnostic standards and conceptualization over time as well as usage that varies internationally. That is, the underlying distribution of usage of this vocabulary is not stationary either in time or space. These issues have been noted by those developing ontologies (Mugzach et al., 2015; Larsen and Hastings, 2018). In the articles used in these results, we have already noted the issues of "drug abuse" vs. "substance abuse disorder," with different connotations across studies, or changes over time combining "substance abuse" and "substance dependence" into a single diagnostic category in the DSM-5 and subsequently in the scientific literature; the user would need to determine their own comfort with separating or combining those terms for their own purposes in combining or contrasting the datasets or the results across the studies. The choice to annotate the papers with the MRI data by types (structural, resting or task based functional scans) for example, is also a level of granularity that could vary depending on the needs of the eventual user; some users may only be interested in particular types of tasks or structural scans, which we have not evaluated. As discussed in more detail in Sahoo et al. (2023), for humans, both developing and interpreting these terminologies requires arbitrary choices about specific cases, deep knowledge of the underlying domain, and the ability to describe and compare nomenclature across studies.

We used JSON as our data serialization format. This choice was motivated by both the heavy standardization and broad adoption already present for JSON as an interchange format (Bray, 2017) and the fact that the output of formally correct JSON is something that all the major LLM producers have implemented. Because JSON is fully represented as text, it can be used both as an input and output format for LLMs. Specifically, JSON can be added directly to prompts while non-text formats cannot be used in this way. As noted in task 2, this allows additional information to be inserted into the overall prompt structure to assist the LLM in generating correct results. Finally, most computer languages and database systems support JSON as a format for data interchange making data extracted from publications broadly sharable. We strongly recommend the use of JSON in this context for all these reasons as there are no current standards for this sort of metadata extraction and exchange.

One concern is the presence of so-called hallucinations in LLM responses. These are outputs that contain untrue information that is apparently generated by the LLMs spontaneously. Generally, hallucinations appear when a LLM must recall some information that is embedded into the model itself rather than using information given in the prompt context. In the annotation task, LLM responses are confined to processing explicit text (the target to be annotated) and this reduces the likelihood of hallucinations dramatically. This approach is also in line with informal industry consensus. It can be hard to distinguish hallucinations from other types of errors, but in the work reported here hallucinations have played little role since, at least, the GPT-4 era.

Hallucinations must be defined based on applications. We define an LLM as having hallucinated only when it responds with something that does not exist (in the context of the task). We did find hallucinations of this sort in older models. For instance, those models would occasionally make up psychiatric diagnosis terms that had the form of our ontology's terms (written in camel case with a similar choice of words to our terms), but which were not in the ontology. The current GPT-4o models did not exhibit this behavior. We do not count instances of false positives in our analysis as hallucinations. For instance, the two cases of resting state imaging labels assigned by the LLM in task 1 are not hallucinations in our sense of the term because resting state imaging is a real type of imaging that was included in our terminology for that task. Others may define hallucinations differently based on their problem domain. We also note that advances have been made in enabling LLMs to "cite" their sources so one can check the responses for hallucinations in contexts where that would make sense (Gao et al., 2023; Byun et al., 2024; Huang and Chang, 2024; Wu et al., 2024).

The larger goal that drives this work is to convert the information in a neuroimaging publication into a programmatically accessible format to enable searching the literature quickly for potential MRI datasets. The work presented here represents the first steps taken toward the goal. Additional work is required to apply this more broadly to the neuroimaging research literature. For example, we focused on human MRI experimental papers, which reported datasets that might be reasonably expected to be available for additional research use. Large-scale data aggregation studies (e.g., Thompson et al., 2020; Ching et al., 2024; Ganesan et al., 2024) entail identifying the existing datasets which might be suitable for a given meta- or mega-analysis, which requires understanding the details of each study design (Turner, 2014). Our goal is to find data that can be used to support answering new questions, therefore the analysis originally performed, the specific results of the study, and the interpretations the original authors make of their experiment and analyses are not necessarily relevant to our goals. Determining the most useful aspects of publications to annotate is an ongoing research question.

We also limited the papers to certain formats. For instance, in this work we used exclusion criteria to remove publications where the information about MRI was only located in tables. We did not provide the LLM with information from the tables, figure legends, or any supplementary material. Expanding this work to cover these cases is still needed. Additionally, automated methods are needed to verify the work of the LLMs so that these processes may be expanded to work at a larger scale. This last will likely require the use of the newest "chain of thought" models. These are all additional topics for future research.

5 Conclusion

We have demonstrated that a recent commercial LLM can extract standardized information regarding neuroimaging experiments at a high level of accuracy. Given the well-known costs of human annotation, this approach is feasible for larger scale applications to the same scientific literature. Using these methods provides a possible road forward for large scale structuring of the research literature and potential facilitation of more nuanced meta-analyses.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/NeuroBridge/LLM-Metadata-Extraction-Paper-2025.

Author contributions

MT: Project administration, Methodology, Supervision, Formal analysis, Writing – original draft, Software, Data curation, Writing – review & editing, Conceptualization, Validation, Investigation. AA: Supervision, Writing – review & editing, Formal analysis. NR: Writing – review & editing, Investigation, Data curation. PG: Investigation, Data curation, Writing – review & editing. AR: Writing – review & editing, Methodology, Supervision, Conceptualization, Funding acquisition. AK: Formal analysis, Writing – review & editing. SS: Conceptualization, Writing – review & editing, Supervision. YW: Conceptualization, Writing – review & editing, Methodology. LW: Writing – review & editing, Conceptualization, Investigation, Funding acquisition. JT: Validation, Data curation, Methodology, Writing – original draft, Investigation, Funding acquisition, Conceptualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by NIDA grant R01 DA053028 "CRCNS: NeuroBridge: Connecting Big Data for Reproducible Clinical Neuroscience" (Wang, Turner, Ambite, Rajasekar, MPI).

Acknowledgments

Preliminary analyses and portions of this work were presented at the INCF Assembly 2024.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2025.1609077/full#supplementary-material

References

Agarwal, S., Laradji, I. H., Charlin, L., and Pal, C. (2024). LitLLM: A toolkit for scientific literature review. *arXiv [Preprint]* doi: 10.48550/arXiv.2402.01788

Aldeen, M., Luo, J., Lian, A., Zheng, V., Hong, A., Yetukuri, P., et al. (2023). "ChatGPT vs. Human annotators: A comprehensive analysis of ChatGPT for text annotation," in *Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA)*, (Honolulu, HI), 602–609. doi: 10.1109/ICMLA58977.2023.00089

Alloza, C., Cox, S. R., Blesa Cábez, M., Redmond, P., Whalley, H. C., Ritchie, S. J., et al. (2018). Polygenic risk score for schizophrenia and structural brain connectivity in older age: A longitudinal connectome and tractography study. *NeuroImage* 183, 884–896. doi: 10.1016/j.neuroimage.2018.08.075

Bhandari, P. (2024). A survey on prompting techniques in LLMs. *arXiv [Preprint]* doi: 10.48550/arXiv.2312.03740

Bray, T. (2017). *The JavaScript Object Notation (JSON) Data Interchange Format*. Fremont, CA: Internet Engineering Task Force. doi: 10.17487/RFC8259

Byun, C., Vasicek, P., and Seppi, K. (2024). "This reference does not exist: An exploration of LLM citation accuracy and relevance," in *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, eds S. L. Blodgett, A. C. Curry, S. Dev, M. Madaio, A. Nenkova, D. Yang, et al. (Mexico: Association for Computational Linguistics), 28–39. doi: 10.18653/v1/2024.hcinlp-1.3

Chang, H., Li, W., Li, Q., Chen, J., Zhu, J., Ye, J., et al. (2016). Regional homogeneity changes between heroin relapse and non-relapse patients under methadone maintenance treatment: A resting-state fMRI study. *BMC Neurol.* 16:145. doi: 10.1186/s12883-016-0659-3

Ching, C. R. K., Kang, M. J. Y., and Thompson, P. M. (2024). "Large-scale neuroimaging of mental illness," in *Principles and Advances in Population Neuroscience*, eds T. Paus, J. R. Brook, K. Keyes, and Z. Pausova (Cham: Springer), 371–397. doi: 10.1007/7854_2024_462

Chumin, E. J., Grecco, G. G., Dzemidzic, M., Cheng, H., Finn, P., Sporns, O., et al. (2019). Alterations in white matter microstructure and connectivity in young adults with alcohol use disorder. *Alcohol. Clin. Exp. Res.* 43, 1170–1179. doi: 10.1111/acer.14048

Comeau, D. C., Wei, C.-H., Islamaj Doğan, R., and Lu, Z. (2019). PMC text mining subset in BioC: About three million full-text articles and growing. *Bioinformatics* 35, 3533–3535. doi: 10.1093/bioinformatics/btz070

Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., et al. (2023). Is GPT-3 a good data annotator? *arXiv [Preprint]* doi: 10.48550/arXiv.2212.10450

Fernandez, R. C., Elmore, A. J., Franklin, M. J., Krishnan, S., and Tan, C. (2023). How large language models will disrupt data management. *Proc. VLDB Endow.* 16, 3302–3309. doi: 10.14778/3611479.3611527

Forster, S. E., Finn, P. R., and Brown, J. W. (2016). A preliminary study of longitudinal neuroadaptation associated with recovery from addiction. *Drug Alcohol Depend.* 168, 52–60. doi: 10.1016/j.drugalcdep.2016.08.626

Ganesan, S., Barrios, F. A., Batta, I., Bauer, C. C. C., Braver, T. S., Brewer, J. A., et al. (2024). ENIGMA-Meditation: Worldwide Consortium for Neuroscientific Investigations of Meditation Practices. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 10, 425–436. doi: 10.1016/j.bpsc.2024.10.015

Gao, T., Yen, H., Yu, J., and Chen, D. (2023). Enabling large language models to generate text with citations. *arXiv [Preprint]* doi: 10.48550/arXiv.2305.14627

Hahn, B., Harvey, A. N., Gold, J. M., Ross, T. J., and Stein, E. A. (2017). Load-dependent hyperdeactivation of the default mode network in people with schizophrenia. *Schizophr. Res.* 185, 190–196. doi: 10.1016/j.schres.2017.01.001

Ham, K. (2013). OpenRefine (version 2.5). http://openrefine.org. Free, open-source tool for cleaning and transforming data. *J. Med. Libr. Assoc. JMLA* 101, 233–234. doi: 10.3163/1536-5050.101.3.020

Hisaharo, S., Nishimura, Y., and Takahashi, A. (2024). *Optimizing LLM Inference Clusters for Enhanced Performance and Energy Efficiency*. Available online at: https://www.authorea.com/doi/full/10.36227/techrxiv.172348951.12175366?commit=edaebe5eaca88afcc9b0a05cc6c126ef44b69885 (Accessed March 6, 2025).

Hua, K., Wang, T., Li, C., Li, S., Ma, X., Li, C., et al. (2018). Abnormal degree centrality in chronic users of codeine-containing cough syrups: A resting-state functional magnetic resonance imaging study. *NeuroImage Clin.* 19, 775–781. doi: 10.1016/j.nicl.2018.06.003

Huang, J., and Chang, K. C.-C. (2024). Citation: A key to building responsible and accountable large language models. *arXiv [Preprint]* doi: 10.48550/arXiv.2307.02185

Janes, A. C., Betts, J., Jensen, J. E., and Lukas, S. E. (2016). Dorsal anterior cingulate glutamate is associated with engagement of the default mode network during exposure to smoking cues. *Drug Alcohol Depend.* 167, 75–81. doi: 10.1016/j.drugalcdep.2016.07.021

Jiang, P., Sonne, C., Li, W., You, F., and You, S. (2024). Preventing the immense increase in the life-cycle energy and carbon footprints of LLM-powered intelligent chatbots. *Engineering* 40, 202–210. doi: 10.1016/j.eng.2024.04.002

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., et al. (2023a). Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv [Preprint]* doi: 10.48550/arXiv.2212.14024

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., et al. (2023b). DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv [Preprint]* doi: 10.48550/arXiv.2310.03714

Kim, D.-J., Schnakenberg Martin, A. M., Shin, Y.-W., Jo, H. J., Cheng, H., Newman, S. D., et al. (2019). Aberrant structural–functional coupling in adult cannabis users. *Hum. Brain Mapp.* 40, 252–261. doi: 10.1002/hbm.24369

Kristensen-McLachlan, R. D., Canavan, M., Kardos, M., Jacobsen, M., and Aarøe, L. (2023). Chatbots are not reliable text annotators. *arXiv [Preprint]* doi: 10.48550/arXiv.2311.05769

Lála, J., O'Donoghue, O., Shtedritski, A., Cox, S., Rodriques, S. G., and White, A. D. (2023). PaperQA: Retrieval-augmented generative agent for scientific research. *arXiv [Preprint]* doi: 10.48550/arXiv.2312.07559

Larsen, R. R., and Hastings, J. (2018). From affective science to psychiatric disorder: Ontology as a semantic bridge. *Front. Psychiatry* 9:487. doi: 10.3389/fpsyt.2018.00487

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., et al. (2023). Large language models understand and can be enhanced by emotional stimuli. *arXiv [Preprint]* doi: 10.48550/arXiv.2307.11760

Li, Y. (2023). "A practical survey on zero-shot prompt design for in-context learning," in *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings*, (Hissar), 641–647. doi: 10.26615/978-954-452-092-2_069

Lottman, K. K., Gawne, T. J., Kraguljac, N. V., Killen, J. F., Reid, M. A., and Lahti, A. C. (2019). Examining resting-state functional connectivity in first-episode schizophrenia with 7T fMRI and MEG. *NeuroImage Clin.* 24:101959. doi: 10.1016/j.nicl.2019.101959

Maik Jablonka, K., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, D., Bran, A., et al. (2023). 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digit. Discov.* 2, 1233–1250. doi: 10.1039/D3DD00113J

Mugzach, O., Peleg, M., Bagley, S. C., Guter, S. J., Cook, E. H., and Altman, R. B. (2015). An ontology for Autism Spectrum Disorder (ASD) to infer ASD phenotypes from Autism Diagnostic Interview-Revised data. *J. Biomed. Inform.* 56, 333–347. doi: 10.1016/j.jbi.2015.06.026

Nahum, O., Calderon, N., Keller, O., Szpektor, I., and Reichart, R. (2024). Are LLMs better than reported? Detecting label errors and mitigating their effect on model performance. *arXiv [Preprint]* doi: 10.48550/arXiv.2410.18889

Nejjar, M., Zacharias, L., Stiehle, F., and Weber, I. (2024). LLMs for Science: Usage for code generation and data analysis. *arXiv [Preprint]* doi: 10.48550/arXiv.2311.16733

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2016). Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* 23, 299–303. doi: 10.1038/nn.4500

Peeperkorn, M., Kouwenhoven, T., Brown, D., and Jordanous, A. (2024). Is temperature the creativity parameter of large language models? *arXiv [Preprint]* doi: 10.48550/arXiv.2405.00492

Petrova-Antonova, D., and Tancheva, R. (2020). "Data cleaning: A case study with openrefine and trifacta wrangler," in *Quality of Information and Communications Technology. QUATIC 2020. Communications in Computer and Information Science*, eds M. Shepperd, F. Brito e Abreu, A. Rodrigues da Silva and R. Pérez-Castillo, (Cham: Springer), 32–40. doi: 10.1007/978-3-030-58793-2_3

Renze, M., and Guven, E. (2024). The effect of sampling temperature on problem solving in large language models. *arXiv [Preprint]* doi: 10.48550/arXiv.2402.05201

Reynolds, L., and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv [Preprint]* doi: 10.48550/arXiv.2102.07350

Sahoo, S. S., Plasek, J. M., Xu, H., Uzuner, Ö, Cohen, T., Yetisgen, M., et al. (2024). Large language models for biomedicine: Foundations, opportunities, challenges, and best practices. *J. Am. Med. Inform. Assoc.* 31, 2114–2124. doi: 10.1093/jamia/ocae074

Sahoo, S. S., Turner, M. D., Wang, L., Ambite, J. L., Appaji, A., Rajasekar, A., et al. (2023). NeuroBridge ontology: Computable provenance metadata to give the long tail of neuroimaging data a FAIR chance for secondary use. *Front. Neuroinformatics* 17:1216443. doi: 10.3389/fninf.2023.1216443.

Sawyer, K. S., Maleki, N., Urban, T., Marinkovic, K., Karson, S., Ruiz, S. M., et al. (2019). Alcoholism gender differences in brain responsivity to emotional stimuli. *eLife.* 8:e41723. doi: 10.7554/eLife.41723

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., et al. (2025). The Prompt Report: A systematic survey of prompt engineering techniques. *arXiv [Preprint]* doi: 10.48550/arXiv.2406.06608

Soylu, D., Potts, C., and Khattab, O. (2024). Fine-tuning and prompt optimization: Two great steps that work better together. *arXiv [Preprint]* doi: 10.48550/arXiv.2407.10930

Strubell, E., Ganesh, A., and McCallum, A. (2019). "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, eds A. Korhonen, D. Traum, and L. Màrquez (Florence: Association for Computational Linguistics), 3645–3650. doi: 10.18653/v1/P19-1355

Tan, Z., Li, D., Wang, S., Beigi, A., Jiang, B., Bhattacharjee, A., et al. (2024). Large language models for data annotation and synthesis: A survey. *arXiv [Preprint]* doi: 10.48550/arXiv.2402.13446

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nat. Med.* 29, 1930–1940. doi: 10.1038/s41591-023-02448-8

Thompson, P. M., Jahanshad, N., Ching, C. R. K., Salminen, L. E., Thomopoulos, S. I., Bright, J., et al. (2020). ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl. Psychiatry* 10, 1–28. doi: 10.1038/s41398-020-0705-1

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2006). "A review of multi-label classification methods," in *Proceedings of the 2nd ADBIS Workshop on Data Mining and Knowledge Discovery (ADMKD 2006)*, (Toronto, ON), 99–109.

Turner, J. A. (2014). The rise of large-scale imaging studies in psychiatry. *GigaScience* 3:29. doi: 10.1186/2047-217X-3-29

van den Heuvel, M. P., Scholtens, L. H., de Lange, S. C., Pijnenburg, R., Cahn, W., van Haren, N. E. M., et al. (2019). Evolutionary modifications in human brain connectivity associated with schizophrenia. *Brain* 142, 3991–4002. doi: 10.1093/brain/awz330

Vanes, L. D., Mouchlianitis, E., Collier, T., Averbeck, B. B., and Shergill, S. S. (2018). Differential neural reward mechanisms in treatment-responsive and treatment-resistant schizophrenia. *Psychol. Med.* 48, 2418–2427. doi: 10.1017/S0033291718000041

Wadhwa, S., Amir, S., and Wallace, B. C. (2024). Revisiting relation extraction in the era of large language models. *arXiv [Preprint]* doi: 10.48550/arXiv.2305.05003

Wang, L., Ambite, J. L., Appaji, A., Bijsterbosch, J., Dockes, J., Herrick, R., et al. (2023). NeuroBridge: A prototype platform for discovery of the long-tail neuroimaging data. *Front. Neuroinformatics* 17:1215261. doi: 10.3389/fninf.2023.1215261

Wang, X., Wang, Y., Ambite, J.-L., Appaji, A., Lander, H., Moore, S. M., et al. (2023). Enabling scientific reproducibility through FAIR data management: An ontology-driven deep learning approach in the neurobridge project. *AMIA Annu. Symp. Proc.* 2022, 1135–1144.

Wu, K., Wu, E., Cassasola, A., Zhang, A., Wei, K., Nguyen, T., et al. (2024). How well do LLMs cite relevant medical references? An evaluation framework and analyses. *arXiv [Preprint]* doi: 10.48550/arXiv.2402.02008

Yadav, S., Choppa, T., and Schlechtweg, D. (2024). Towards automating text annotation: A case study on semantic proximity annotation using GPT-4. *arXiv [Preprint]* doi: 10.48550/arXiv.2407.04130

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. *arXiv [Preprint]* doi: 10.48550/arXiv.2303.18223