Decoding Unstructured Text: Enhancing LLM Classification Accuracy with Redundancy and Confidence

Jörn Boehnke, Elizabeth Pontikes, Hemant K. Bhargava
Graduate School of Management, University of California Davis

July 19, 2024

Abstract

This paper develops guidelines for using Large Language Models (LLMs) in labeling and classifying complex, unstructured text. Researchers have choices of multiple configuration dimensions when employing LLMs for classification tasks. We propose a model where researchers optimize a utility function that includes two novel configuration dimensions: redundancy in classification and a confidence threshold based on the LLM self-reported confidence. The function also includes LLM version and whether prompts include detailed instructions. We propose researchers must balance increased accuracy with costs due to incomplete tasks and financial outlays. In an empirical test where we ask LLMs to classify whether press release headlines are about an acquisition, we find that applying thresholds for LLM self-reported confidence yields increases in accuracy, though with costs of incomplete tasks. Notably, we only find that redundancy increases accuracy when a strict confidence threshold is applied. For LLM classification, redundancy alone does not increase accuracy but does increase costs due to incomplete information and financial outlay. We also find that LLM version and detailed instructions shift the efficiency frontier of trade-offs between accuracy and costs of incomplete tasks when employing redundancy and confidence thresholds. Finally, we compare LLM classification results for different configurations to redundancy in human worker classification. We draw on our model and empirical results to develop guidelines for researchers and practitioners using LLM classification to create structured data from unstructured text.

1 Introduction

There are many settings in scholarly research and applied work where large amounts of unstructured data exist, too diverse to aggregate directly, yet containing elements needed for high-value structured data. Researchers have conducted search and labeling tasks on such data for decades, usually relying on human workers such as undergraduate students or crowdsourced labelers ("mechanical turks" sourced over the Internet), the latter of whom may lack expertise but are more cost-effective (e.g., Geiger et al., 2021; Cegin et al., 2023; Li et al., 2023). More recently, considerations of cost and expertise have motivated researchers to seek automated solutions such as machine learning and natural language processing (NLP) algorithms, an approach drastically amplified after the advent of large language models (LLMs) and Generative AI (Zhikai Chen et al., 2023; Kolluri et al., 2019). The idea is that LLMs can provide more expertise at a lower cost, opening opportunities for researchers to develop structured data more quickly and easily than ever before. However, there remain outstanding questions about this approach. While researchers have years of experience guiding them in effectively using human workers, employing LLMs for these tasks is still in the nascent stages.

This paper develops guidelines for the effective use of LLMs in classification tasks using unstructured data. Our guidelines are based on experiments where we asked LLMs to classify press release headlines. As we iterated through various designs and observed LLM performance matched to "ground truth" created by the authors as subject matter experts, we recognized that any practitioner who wishes to deploy LLMs for classification tasks would confront a series of decision problems. The contributions of this paper are to articulate the decision problems and provide guidelines based on empirical tests that are specific to each decision problem. We develop novel decision points and solutions that practitioners should consider when using any LLM (or more generally, an AI tool) for labelling unstructured data.

Typically, the practitioner has a choice among multiple *configurations*, and seeks to optimize the cost-quality trade-off along some efficient frontier. Two straightforward configuration dimensions are: (1) the LLM version and (2) instructions required for accurate output (Del Arco et al., 2024).[1] We also theorize and empirically study whether *redundancy* in classification and LLM self-reported *confidence* increase accuracy – and if so, how they affect the researcher's cost-quality trade-off and efficient frontier. Redundancy in human responses is well studied and broadly incorporated for human worker classification tasks to increase accuracy (Agley et al., 2022; Alyakoob and Rahman, 2022). However, to our knowledge, whether redun-

[1] Prompt engineering is also an important aspect of using an LLM effectively, and indeed we experimented with multiple prompts. However, enough is already written on this topic in both the scholarly and general press (see, e.g., Sahoo et al. (2024) and Liu et al. (2023)), and we choose to focus our research on other less-studied dimensions.

dancy increases accuracy for LLM classification has not been examined. This is critical information given the financial costs for LLMs to duplicate a task. Furthermore, we theorize that asking LLMs to self-report confidence and applying a confidence threshold will further result in accuracy increases, a dimension that, to our knowledge, has neither been studied nor broadly employed in LLM classification.

Our guidelines examine how a practitioner might set the configuration dimensions $(r, d; v, i)$ for a labeling project. Each configuration choice will affect two, generally conflicting, outcome measures: accuracy $a(r, d; v, i)$ and cost $C(r, d; v, i)$, where the cost measure comprises both the direct financial costs $x(r, d; v, i)$ of applying the dimensions, and a delayed cost $C(n_u(r, d; v, i))$ for handling *unresolved* labeling tasks. Tasks remain unresolved when the LLM results fail to meet the required confidence threshold d for all r repetitions. Intuitively, higher accuracy implies higher cost (both direct and delayed), hence it is vital for the practitioner to pay special attention to the nature of non-linearity in accuracy (i.e., whether there are diminishing or increasing marginal returns for accuracy) when defining their two-dimensional utility function $U(r, d; v, i)$ over the configuration dimensions.

Our empirical analysis reveals that, as expected, more advanced LLM versions and more detailed instructions increase accuracy. Perhaps more intriguingly, these dimensions also reduce the costs of incomplete tasks, thus shifting the *efficiency frontier* for the accuracy-cost trade-off when employing redundancy and confidence. Our analysis further suggests that the practitioner should include a *blend* of redundancy and confidence to realize higher accuracy, i.e., a task is deemed *resolved* only when r repetitions exceed the confidence threshold d. Notably, we do not find evidence of increased accuracy when applying redundancy alone (unlike when employing human workers). This provides an important guideline for practitioners weighing the cost-accuracy trade-off when using LLMs. Redundant classification, in the absence of applying a strict confidence threshold, increases costs without increasing accuracy.

Based on our experiences, we emphasize that in most applications there will be some tasks that are quite nuanced and difficult to perform, for both human labellers and an LLM, and may require more costly expert review. Thus, the LLM can be used as a tool for a scalable approach for *isolating a small fraction of difficult tasks*, which can then be forwarded to experts, while performing relatively easier tasks at a low cost and with a machine's "willingness" to perform repetitive work. This is consistent with the idea that humans and AI can be an effective and efficient combination for cognitive work (Fügener et al., 2022).

Our guidelines were based on extensive experiments and analysis involving data from a large corpus of corporate press releases (over 27 million stories sourced from Lexis/Nexis archives spanning three decades). The current study stemmed from a related investigation, where our intent was to examine small acquisitions by "big tech" companies such as Amazon, Facebook, and Google – ones that were not required to be reported and thus were difficult to track, but that are typically announced in press releases. Our first step was to identify acquisitions of small private companies from press release headlines. Since press releases cover hundreds of topics (not just acquisition related events), and because even an acquisition headline could be written in hundreds of ways, a vital element involved inferring acquisition-related actions from the press release headlines and text. We found that discerning that a headline described an acquisition (or not) was difficult, as evident in Table 1. We needed an automated, reliable way to examine these releases for evidence of corporate acquisitions, their nature, intent, and post-acquisition events – and to be inclusive of acquisitions of small companies and start-ups often missed by government databases and existing commercially available databases.

Headline Text	Comment
BWI - Completion of Winery Purchase purchase	Appears to represent an acquisition, but not clear if it is a property or a company
Vantex Acquires the Lac Fortune West Property	Property acquisition, not a corporate acquisition
SYB - Sigma consortium's superior bid for Symbion businesses 1/1	Suggests an effort to acquire, and is "about" acquisition, although outcome is not clear
BPC - Sale of Hartland Cables Business	Could be the sale of a company or a business line
Crosshair Announces Termination Of Purchase and Sale Agreement with Strathmore Minerals Corp.	Does not indicate what was intended to be purchased
JDSU Completes Sale of Hologram Business to OpSec Security	This is likely the sale of a business unit but could be an independent company.
Orrick Advises Instagram on Acquisition by Facebook	Does not convey if an acquisition is in progress or it is about a potential acquisition.
Acquires Significant Ownership Stake In Network	Indicates an acquisition but not whether it is majority control.
LAF - Acquisition & Financing of the Rapu Rapu	This could be a company acquisition or property.
Acquires Tervita's Drilling Fluids Business And Forms Strategic Alliance	Unclear whether this is an acquisition of a company or a division within a company.

Table 1: Sample Press Release Headlines Potentially About a Corporate Acquisition.

This prompted us to use LLMs for classification, inputting press release headlines and asking the LLM to output whether the headline was about a corporate acquisition (and if so, for which entities). These proved to be daunting tasks which required substantial expertise to perform correctly because, similar to BioNER work (Zhou et al., 2021), acquisition-related headlines have a lot of nuances: acquisitions are stated in multiple ways and acquisition-related words do not always refer to a corporate acquisition. We uncovered more challenges than we expected in realizing necessary accuracy levels for our structured data – challenges that using the latest LLM version and crafting targeted instructions did not adequately address.

Thus, we embarked on a series of experiments that led to systematic guidelines for using Generative AI tools to label unstructured text. We also compared LLM results to human worker classification, based on data from an initiative where we recruited and trained over a dozen undergraduate student research assistants (RAs) to classify the headlines as a benchmark for assessing LLM accuracy.

We draw on our results to create guidelines for any researcher or practitioner who employs Generative AI tools for a labeling project. Our results suggest that by following these guidelines researchers can employ LLMs to achieve high accuracy even for moderately complex classification tasks that require background information or special handling.

2 Emerging Literature on LLMs for Labelling

Our work contributes to the growing area of research on the use of Generative AI for labeling and for making comparisons between human and machine techniques. Several recent studies show promising results for advanced LLMs. Cegin et al. (2023) report that for "paraphrase generation for intent classification," GPT creates more robust and diverse answers as compared to human labelers. Ringel (2023) experiments with using GPT-4 as a surrogate for human expertise in identifying marketing mix variables in consumers' posts on Twitter, and finds GPT labels – unlike crowdsourced labels – are in high agreement with expert labels. Le Mens et al. (2023) find that GPT-4 without training closely tracks human judgment in classifying book descriptions as similar to a genre. Li et al. (2023) find that with suitable prompt engineering, Generative AI models can match the performance of human surveys in generating perceptual maps, and do so more cost effectively.

Other studies show more nuanced effects. Zenan Chen and Chan (2023) tasked expert and non-expert users to write ad copy either on their own or using LLMs. They find improved outcomes when the LLM served as a sounding board to comment on human-created content, but there is a detrimental performance when used as a ghostwriter with primary responsibility for the content. Relatedly, Brynjolfsson et al. (2023) examine the use of Generative AI for customer service agents and show substantial productivity gains for less experienced agents but little to negative gains for highly experienced ones.

Previous studies compare human vs. machine performance, but it is an open question whether and when researchers can effectively apply LLMs to create structured data, especially for complex tasks. An example of complicated but important labeling is *Biomedical Named Entity Recognition* (BioNER), a data labeling task that is vital for scaling and automating downstream biomedical natural language processing tasks (Zhou et al., 2021). It is non-trivial and difficult because of "...various ways of naming biomedical entities, ambiguities caused by the frequent occurrences of abbreviations, and new entities constantly and rapidly reported in scientific publications." Our research group faced a similar challenge in labeling corporate press release headlines.

3 LLM Design for Labeling: Configurations and Objectives

In a project where an LLM is deployed for large-scale labeling of unstructured data, a typical practitioner will care about both the *accuracy* of the LLM's work and its *cost*, with cost comprising both immediate costs of execution as well as a delayed cost of dealing with tasks that the LLM was unable to resolve. In general, higher accuracy will result in higher execution cost and, as we discuss below, may also create higher delayed costs. What is critical, however, is that the practitioner has multiple ways to deploy the LLM, which vary in the accuracy-cost trade-offs they present. We focus on alternative ways of deployment from different combinations along the following four dimensions.

- **Version of the tool (v).** In 2023, we considered GPT-3.5 and GPT-4, the latter with a higher cost and a promise of better quality.
- **Level of instructions given to the LLM (i).** Ideally, more detailed instructions should improve quality with relatively little increase in cost.
- **Redundancy level (r).** Assign the same task multiple (r) times (with the LLM set at a high "creative" level to enable variety in answers), then consider the task as resolved only upon agreement among the r repetitions, and compute accuracy $a(r; v, i)$ based on the performance of this subset of tasks against ground truth. This approach requires defining a redundancy level and a resolution rule (e.g., unanimity).
- **Confidence threshold (d).** Ask the LLM to self-report *confidence*, accept only those responses that exceed a defined *confidence threshold* d, and compute accuracy $a(d; v, i)$ based on the performance of this subset of tasks against ground truth.

An LLM configuration is a specific combination of $(r, d; v, i)$, and the number of combinations depends on the choice set within each dimension. Each configuration imposes an immediate execution cost $x(r, d; v, i)$, and delivers an associated accuracy level $a(r, d; v, i) \in [0, 1]$ for tasks it can resolve, but it leaves a subset $n_u(r, d; v, i) \in [0, 1]$ of unresolved tasks because of failure to achieve the confidence threshold d for all r repetitions. These unresolved tasks impose a delayed cost $C(n_u(r, d; v, i))$ representing unfinished tasks or the cost of labeling via expensive human experts. These functions can be estimated by running all configurations over a sample of tasks. We propose practitioners should define their utility function for accuracy $U(a(r, d; v, i))$, along with execution and unresolved-task cost functions. Utility for accuracy should be framed with attention to the nature of non-linearity (i.e., whether there are diminishing or increasing marginal returns for accuracy). For instance, diminishing marginal returns can be captured by a function such as $\sqrt{a}$; increasing returns can be reflected via a^2; and a logistic-like function $\frac{2}{1+e^{-x}} - 1$ can be used to get increasing returns with an asymptotic maximum. Likewise, there is a need to estimate the cost function for unresolved tasks.

The practitioner needs to pick a best-fit configuration over the space of configurations. This best-fit configuration would maximize net utility ($U(a(r, d; v, i)) - w_1 x(r, d; v, i) - w_2 C(n_u(r, d; v, i))$), where w_i are relative weights assigned to costs. Alternatively, it could maximize accuracy subject to a cost (budget) constraint, or minimize cost for a desired level of accuracy. Since the specific optimization goals, and cost and accuracy functions, are practitioner-specific, we do not make specific recommendations on which configuration to use. Instead, we leverage our experimental work to describe trade-offs by computing an efficient frontier between accuracy and costs based on percent of unresolved tasks.

A key contribution centers on the role of redundancy in increasing LLM classification accuracy. When human workers are utilized for labeling complex unstructured data, affordable workers – undergraduate students with extra time or strangers sourced over the Internet using tools such as Prolific or AMT – often have limited expertise on the topic (Agley et al., 2022; Alyakoob and Rahman, 2022). Hence, it is common to assign each labeling task to multiple workers under the assumption that an identical label from multiple (n) researchers implies a higher confidence that the label is correct. For instance, Alyakoob and Rahman (2022) initially assigned each labeling task to 3 workers; in case of disagreement it was assigned to another 2 (total 5); and if at least 4 out of 5 responses weren't identical, another 5 subjects were sought; if 7 out of 10 agreed, the task was assigned that label and otherwise considered unresolved. Similarly, professional news organizations require at least 2 convergent sources to publish a claim. In the U.S., the jury system requires unanimity among 12 jurors to render a verdict in a criminal case, otherwise resulting in a hung jury and possible retrial.[2] Radiology readings of complex images often require multiple radiologists or a mix between machine and human experts.

[2]The requirement is relaxed to a $\frac{5}{6}th$ agreement for some civil trials, with the caveat that a panel with only 6 jurors must produce a unanimous outcome (Andres v. United States, 333 U.S. 740, 748 (1948)).

The idea that redundancy leads to increased quality draws from the French intellectual Marquis de Concordet's demonstration, two and a half centuries ago, that "there are situations in which it is advisable to entrust a decision to a group of individuals of lesser competence than to a single individual of greater competence" (Boland, 1989). When redundancy is implemented by assigning a task to multiple human subjects, it is reasonable to assume that each subject's response on a task is independent and that subjects have identical probability p of providing an accurate response. Then, if the same task is assigned to m subjects, the probability that all m agree and provide the *correct* response is p^m; that all m agree and provide an *incorrect* response is $(1-p)^m$; and the task remains unresolved with probability $1-p^m-(1-p)^m$. Further, the *conditional probability* (i.e., conditional on all m subjects providing the same response) of a correct vs. incorrect response is $\frac{p^m}{p^m+(1-p)^m}$. Finally, if the labeling project involves n tasks, one would have an expected value of $n(1-p^m-(1-p)^m)$ unresolved tasks under redundancy level m. These relationships create an internal consistency check for our labeling experiments and empirical analysis.

With this background of redundancy in human labeling projects, we asked if redundancy also increases accuracy when an LLM does the labeling. One might initially think redundancy is unnecessary with LLMs because a machine, given the same input, should produce the same output each time. This determinism would imply that multiple iterations of the same task by an LLM would yield identical responses, negating the benefits of redundancy realized with human labelers. However, LLMs can be configured to introduce variability through a parameter called "temperature," which controls the randomness of the model's predictions. A lower temperature results in more deterministic outputs, while a higher temperature increases randomness, allowing the model to generate different responses to the same prompt. It is an open question how similar this is to having multiple human labelers evaluate the same task, in terms of whether LLM redundancy also improves accuracy.

Another important contribution is that we investigate whether prompting an LLM to self-report its confidence can be used to increase accuracy. We reason that prompting a model for its confidence can help filter out less reliable answers, thereby improving the overall accuracy of the task at hand. High-confidence responses typically mean the model finds strong alignment with familiar patterns, making these responses more dependable. This is particularly useful in applications where precision is crucial, such as data classification or decision support systems. By using confidence assessments, researchers can focus on high-confidence answers for critical tasks, while flagging low-confidence responses for further review, thus optimizing the balance between automation efficiency and accuracy. Therefore, we include a confidence threshold in our model, theorizing that conditioning on self-reported confidence should improve accuracy (while incurring a cost of incomplete tasks).

A novel element our analysis uncovers is that combining *redundancy* with self-reported *confidence* further enhances accuracy. By leveraging multiple responses generated at a higher temperature and applying a confidence threshold, we select only the most confident and varied responses. This combined strategy mitigates the limitations of both methods when used alone, providing a more robust framework for accurate and efficient data labeling. This ensures that tasks are resolved with high reliability, optimizing the balance between automation efficiency and accuracy.

4 Data and Methodology

We developed our guidelines from our observations of LLM performance classifying unstructured data based on the configuration dimensions described above.

4.1 Lexis Nexis Dataset of Press Releases

The data for our study is from a comprehensive text corpus comprising over 27 million press release headlines from Lexis-Nexis. This corpus represents a broad spectrum of corporate communications spread over several decades, providing a large and varied initial set of articles for our analysis.

We seek to evaluate the performance of LLMs in *differentiating between headlines pertaining to corporate acquisitions and mergers, and those that do not*. With 27 million headlines, a random selection large enough to yield a substantial number of acquisition-related headlines from human or LLM labelers is cost-prohibitive. Therefore, we first apply automated methods for an initial filter to generate a set of headlines for review that over-sampled those likely to be acquisition related.

This task is addressed through a two-stage process. In the first stage, we employ a pretrained Bidirectional Encoder Representations from Transformers (BERT) named entity recognition model, combined with part-of-speech (POS) tagging heuristics. This approach focuses on identifying key elements indicative of acquisitions, such as organizational names and relevant linguistic markers. We consider a headline as potentially related to an acquisition based on the detection of specific verbs, nouns, company names, and terminologies typically associated with acquisition activities. The second stage involves using Word2Vec for generating detailed headline embeddings to grasp the semantic context more effectively. We supplement this with a standard BERT classifier, enhanced by data augmentation techniques to address class imbalances.

With this methodology, we identify a focused set of headlines that includes roughly 50% likely acquisition-related headlines and 50% randomly selected headlines. Given that acquisition-related headlines are rare in the original dataset, this approach creates a balanced, manageable, and realistic dataset.

4.2 Empirical Analysis

We ask an LLM to classify the set of press release headlines as being about an acquisition, or not (see Appendix C for details). We focus on evaluating *accuracy* by comparing *LLM classification* to *expert classification*, and *costs* based on unresolved tasks, as a function of the configuration dimensions (as outlined in section 3).[3]

We first define our dependent and independent variables, and then describe data collection for *expert classification* and *LLM classification*. For additional analyses, we compare LLM classification accuracy to *human worker classification* based on assessments of undergraduate RAs.

4.2.1 Dependent Variables

- **Accuracy ($a(r, d; v, i)$).** The correlation between LLM and expert classification for headlines accepted based on the independent variables, where the classification task is whether a headline is about an acquisition or not.

- **Cost of incomplete tasks $C(n_u(r, d; v, i))$.** $n_u(r, d; v, i)$ is the number of headlines left unresolved based on the given configuration or independent variables.

4.2.2 Independent Variables

- **Version of the LLM (v).** A categorical variable, with 1 indicating the LLM classifier is GPT-4, and 0 indicating the classifier is GPT-3.5.

- **Level of instructions (i).** We employ two prompts, one with very basic directions of inputs and outputs, and another that includes detailed instructions describing what constitutes an acquisition and examples. This is a categorical variable with 1 indicating the LLM was given detailed instructions, and 0 indicating basic instructions.

- **Redundancy level (r).** This is the number of times the LLM was asked to perform the same classification task (with the LLM set at a high "creative" level), between 1 and 25 times. Redundant

[3] Financial costs can then be calculated based on costs to running the LLM at the desired redundancy level and employing human workers or expert classification to fill in incomplete tasks.

classification of a headline required defining a resolution rule for accepting the LLM response and assigning classification. We use unanimous agreement, only including headlines with 100% consistent classification as acquisition or not across all trials.[4]

- **Confidence threshold (c).** This is the threshold of self-reported confidence required to accept the LLM response. Again, we need to employ an aggregation rule and we use average confidence over redundant trials. The LLM reports confidence as a percentage from 0 to 100. We test five confidence thresholds: (1) all headlines (no threshold), (2) 85%, (3) 90%, (4) 95%, and (5) 100% confidence (details below).

4.2.3 Estimation

We use OLS regression to assess whether, and the extent to which, our theorized configurations impact accuracy and cost due to incomplete tasks, as theorized in our utility functions.

$$a = \beta_1 * v + \beta_2 * i + \beta_3 * r + \beta_4 * c$$

$$n_u = \gamma_1 * v + \gamma_2 * i + \gamma_3 * r + \gamma_4 * c$$

With these results, we then provide guidance for how a researcher can assess the cost-quality trade-off along an efficient frontier. If both β and γ are positive, this configuration dimension presents a trade-off, and the researcher must weigh the value of increased accuracy with incurring increased costs. If β is positive and γ is negative, this represents a shift in the efficiency frontier, such that the configuration results in accuracy increases and reductions in costs due to incomplete tasks.

4.2.4 Expert Classification

To assess accuracy, we compare classification by LLMs to expert classification as determined by the authors of this paper. Two authors independently reviewed each headline for a sample of 1,155 headlines and agreed on the classification for 1,022 headlines. We conduct our analysis on the 1,022 headlines with expert agreement. This excludes the subset of headlines (133) where the underlying meaning is ambiguous.

[4] We only include consistently classified headlines because otherwise average confidence is not meaningful. The LLMs are highly consistent in their classification; with redundancy of 25 trials, between 75% and 95% of headlines are always classified as acquisition or not (depending on the version and level of instructions). We do not find substantial differences in accuracy based on consistent classification across redundant trials. Appendix A provides details.

4.2.5 LLM Classification

For LLM classification, we study classification from two versions: GPT-3.5 and GPT-4. We also devised two distinct experimental setups: a "detailed instructions" prompt and a "basic instructions" prompt, resulting in the first four configurations (see Appendix C for prompt language). All prompts ask the LLM to report its confidence in its classification as a percentage between 0 and 100.

For each of the four version/instruction configurations, we ran the prompts 25 times ($k = 25$) to generate 25 separate classification responses. We consider classification for the k requested runs for the respective redundancy level ($k = 1$ through $k = 25$).

We evaluate accuracy by computing the correlation between expert classification and LLM classification for the headlines retained based on the four independent variables. For instance, using GPT-3.5 with basic instructions, a confidence level of 85, and a redundancy of 10, the LLM classifies headlines as acquisitions or not based on consistent results across 10 trials, with an average reported confidence of at least 85. We compute the correlation of LLM and expert classification for this set of headlines, and the correlation, number of unresolved headlines, and respective configuration dimensions are one observation.

We repeat the entire set of $k = 1$ through $k = 25$ redundant *trials* over multiple *runs*. We employ 11 runs for GPT-3.5 and 5 runs for GPT-4, with fewer runs for GPT-4 due to cost considerations. This results in 4,000 observations.[5] Table 2 summarizes our approach. For all configurations we use the parameter temperature $= 1.0$, which controls the randomness of the generated responses. A temperature of 1.0 indicates a higher level of randomness, allowing for more varied and creative responses. This setting was chosen to increase the LLMs' capability to generate diverse interpretations of the headlines.

	Basic Instructions	Detailed Instructions
GPT-3.5	11 trials	11 trials
GPT-4	5 trials	5 trials

Table 2: Number of trials performed under four configurations. Each trial involved 25 repetitions.

The text of the prompts are in Appendix C. The prompt with *detailed instructions* provides definitions and examples to explain how we want the LLM to classify headlines as acquisitions (coded as "A"), mergers (coded as "M"), or neither (coded as "N"), along with its confidence in the classification, and to identify the names of the companies involved, if applicable.[6] This prompt includes examples of each type, additional instructions on specific cases (e.g., partial acquisitions, acquisitions of assets not constituting a company

[5] 2 instruction configurations, 5 confidence thresholds, 25 redundancy levels, 11 runs for GPT-3.5 and 5 runs for GPT-4: 2x5x25x(11+5) = 4,000.

[6] For our analysis, we combine the acquisition and merger classifications.

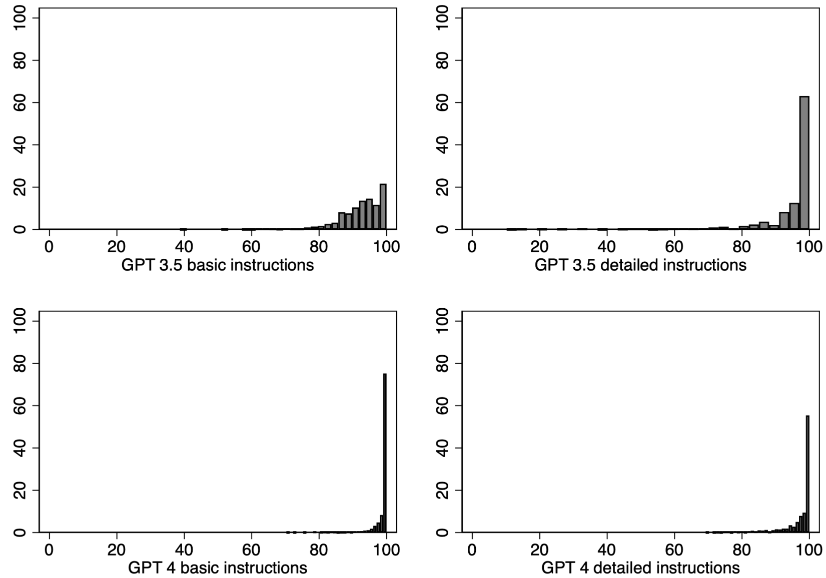

Figure 1: Distribution of LLM self-reported confidence for our base prompt, by condition.

acquisition), and guidelines for identifying company names involved in these events. The aim is to provide the LLM with comprehensive contextual understanding before classification. The prompt with *basic instructions* presents a more straightforward task to the LLMs without additional context or examples. It instructs the LLM to read a headline and classify it as an acquisition, merger, or neither, its confidence in the classification, and to identify the names of the companies involved, if applicable. This prompt tests the LLMs' inherent capabilities in headline classification without detailed instructions, examples, or other context.

We first inspect the distributions of the LLMs' self-reported confidence in its classifications. Representative distributions from one run for each condition are presented in figure 1. These show the LLMs' self-reported confidence for this prompt is highly skewed toward 100%, which informs the confidence thresholds for our independent variable (see section 4.2.2).

4.2.6 Human Worker Classification

In additional analyses, we compare LLM classification accuracy to human workers. This provides a baseline for assessing the levels of LLM accuracy that can be achieved with our guidelines. It also provides a potential pathway for resolving the incomplete tasks.

We recruited 13 undergraduate research assistants from a West Coast university to read a random sample of the headlines selected by our algorithm and label them as describing a company merger and acquisition (whether it is impending or already completed), or not. We ask them to classify each headline as either: (1) about a company merger or acquisition, (2) about an acquisition of property, (3) not about a merger or acquisition, or (4) unclear/unsure.[7] We also asked them to enter the company names for the acquirer or acquiree into a text box, if applicable. Unfortunately, we did not collect data on RA self-reported confidence because we conducted this data collection before we theorized about this configuration dimension for LLM classification. The RAs attended a two-hour instructional session where we explained criteria for classifying headlines into the above categories. During the session, the RAs classified a set of test headlines and then we presented the correct classifications and answered questions.

The RAs worked remotely. We created an interface using a Google spreadsheet where the RAs could fetch a set of 50 headlines at a time to classify. The interface provided a drop-down box with the four classification choices and columns to copy and paste the acquirer and acquiree names when applicable. When the RAs finished classifying the 50 headlines they would click the "submit" button that recorded their answers in our database. The RAs then had the option of retrieving another set of 50 headlines. Each headline was reviewed by between one to eight RAs.

5 Results

5.1 LLM Classification

We conduct our analysis on the 1,022 headlines where there was expert agreement. As described in section 4.2.3, we use OLS regression to analyze whether the configuration dimensions predict (i) classification accuracy, measured by the correlation between LLM and expert classification, and (ii) costs incurred later due to incomplete responses, measured by the number of headlines not accepted. We also run additional regressions to consider the interaction between two of the dimensions, redundancy and confidence threshold. Table 3 summarizes the statistical results from all regressions.

Redundancy and confidence thresholds: Results in table 3 show an accuracy-cost trade-off for employing redundancy and confidence thresholds. Every unit of increased redundancy increases the accuracy correlation by 0.001 (column 1) such that redundancy of 25 results in a .025 increased correlation. The cost of incomplete tasks also increases with each unit of redundancy by 5.5 (column 3) such that redundancy of

[7] As described above, some headlines use terms like 'acquire' to describe acquiring property, rather than a corporate acquisition. We thought it might be useful for subsequent algorithm development if our training data separately classified property acquisition. For the purposes of this paper, this category is grouped with 'not about a merger or acquisition.'

	Corr. Model	Corr. Model (I)	Cost Model	Cost Model (I)
Intercept	0.680^{***}	0.623^{***}	608.760^{***}	497.199^{***}
	(0.003)	(0.004)	(5.281)	(7.722)
Confidence Threshold 0	-0.193^{***}	-0.130^{***}	-452.639^{***}	-318.624^{***}
	(0.003)	(0.005)	(5.361)	(10.565)
Confidence Threshold 0.85	-0.199^{***}	-0.127^{***}	-393.015^{***}	-237.026^{***}
	(0.003)	(0.005)	(5.361)	(10.565)
Confidence Threshold 0.90	-0.179^{***}	-0.107^{***}	-336.183^{***}	-190.265^{***}
	(0.003)	(0.005)	(5.361)	(10.565)
Confidence Threshold 0.95	-0.135^{***}	-0.058^{***}	-222.656^{***}	-100.774^{***}
	(0.003)	(0.005)	(5.361)	(10.565)
Version: 4.0	0.062^{***}	0.062^{***}	-189.233^{***}	-189.233^{***}
	(0.002)	(0.002)	(3.657)	(3.496)
Instructions: Detailed	0.039^{***}	0.039^{***}	-82.018^{***}	-82.018^{***}
	(0.002)	(0.002)	(3.390)	(3.241)
Redundancy	0.001^{***}	0.006^{***}	5.546^{***}	14.127^{***}
	(0.000)	(0.000)	(0.235)	(0.503)
Confidence Threshold 0 x Redundancy		-0.005^{***}		-10.309^{***}
		(0.000)		(0.711)
Confidence Threshold 0.85 x Redundancy		-0.006^{***}		-11.999^{***}
		(0.000)		(0.711)
Confidence Threshold 0.90 x Redundancy		-0.006^{***}		-11.224^{***}
		(0.000)		(0.711)
Confidence Threshold 0.95x Redundancy		-0.006^{***}		-9.376^{***}
		(0.000)		(0.711)
R^2	0.698	0.724	0.761	0.781
Adj. R^2	0.697	0.723	0.760	0.781
Num. obs.	4000	4000	4000	4000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: OLS regressions to evaluate effects of configuration dimensions on LLM classification accuracy ("Corr. Model") and "inverse cost" (number of headlines included, "Cost Model"). Columns 2 and 4 add the interaction between redundancy and confidence threshold ("I"). Confidence threshold effects are compared to employing a 100% threshold (baseline) and detailed instructions are compared with minimal (baseline).

25 results in an incomplete task cost of 139. There is also an accuracy-cost trade-off to applying a higher confidence threshold: employing a 100% confidence threshold increases the accuracy correlation by .135 compared to a 95% threshold, and by .19 compared to no threshold. The trade-off is that costs also increase by 223 and 452, respectively.

LLM version and level of instructions: Results also show that both the advanced LLM version (GPT-4) and detailed instructions *increase accuracy* while *reducing costs* of unresolved tasks. The advanced LLM increases the accuracy by .062 (column 1) while also decreasing the costs due to incomplete tasks by 189 (column 3). Providing detailed instructions increases accuracy by .039 (column 1) with an accompanying cost decrease of 82.

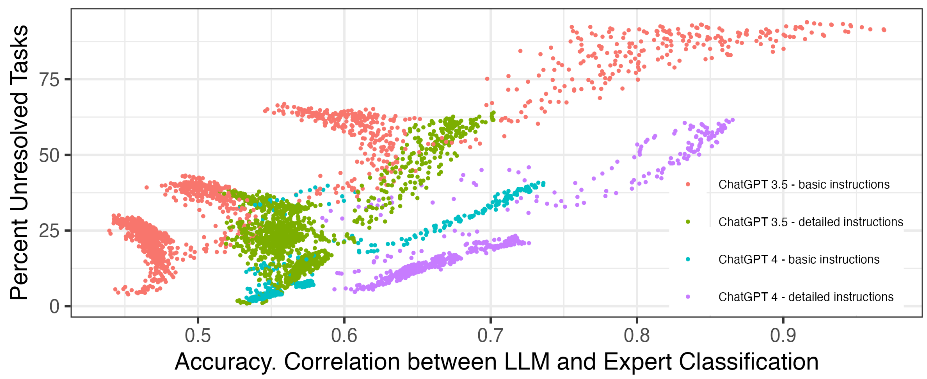

Figure 2: Efficiency frontier. Correlation accuracy and unresolved tasks for each trial and run, by LLM version and instructions.

Together, these results indicate that the advanced LLM and instructions shift the efficiency frontier in terms of the accuracy and cost of unresolved task trade-offs. Figure 2 plots accuracy and unresolved tasks for each observation. The color of the data point indicates the first two configuration dimensions: whether GPT-3.5 or GPT-4 was used and if the prompt contained detailed instructions. The graph presents an overall cost-accuracy trade-off, with higher accuracy resulting in higher costs (more unresolved tasks), which is due to employing different confidence thresholds and redundancy (variance is also due to stochasticity across runs). The graph also shows that version and instructions shift the efficiency frontier: there is a larger increase in cost from unresolved tasks per increment of accuracy for the less advanced LLM and basic instruction prompts. For example, with GPT-4 with detailed instructions, accuracy levels increase to over .8 correlation while losing half the tasks, whereas for GPT-3.5 with basic instructions, high accuracy comes at a very high cost of retaining less than 10% of tasks. Also interesting to note: only GPT-3.5 with basic instructions yields such large sacrifices in cost of incomplete observations to realize exceptionally high accuracy – for this version redundancy and confidence thresholds result in correlations that approach 1, but with only a handful of tasks resolved.

Using the advanced LLM and detailed instructions to shift the efficiency frontier does incur a *financial cost*. The cost comparison for OpenAI API access to (standard) GPT models is as follows:

Standard Models:

- GPT-3.5: $1 to $1.50 per 1 million input tokens and $2 per 1 million output tokens.[8]

- GPT-4: $30 per 1 million input tokens and $60 per 1 million output tokens.

Thus, GPT-4 is approximately 20 to 30 times more expensive for both input and output tokens compared to GPT-3.5. In our case, for 1,000 tasks, using GPT-4 with detailed instructions would incur an additional $970 expense.

Figure 3 plots the relationship between accuracy and redundancy by confidence threshold for classification by GPT-4 with detailed instructions.[9] Markers plot the average accuracy over all runs for the respective redundancy level and confidence threshold, and error bars are 95% confidence intervals. The graph reveals that accuracy is fairly flat as redundancy increases except when the confidence threshold is 100% – that is, when we only include responses where the LLM has full confidence in all redundant trials.

Interactions: Based on the trends revealed in figure 3, we investigate whether interactions between redundancy and confidence thresholds affect accuracy and costs. Table 3 presents effects. Column 2 reveals an interaction such that redundancy only increases accuracy when there is a high confidence threshold (at 100% confidence, the baseline condition), and the size of the redundancy effect increases by 6 times, to .006. For the 100% confidence threshold, redundancy of 25 increases correlation accuracy by .15 compared to no redundancy. When we run the estimation excluding the 100% confidence conditions from our risk set, the effect of redundancy decreases by two orders of magnitude and is no longer significant ($\beta = .00007; p = .35$).

This result indicates that for LLM classification *redundancy alone does not improve accuracy*. Unlike human worker classification, running redundant trials without conditioning on confidence does not provide benefits for the additional financial cost. Moreover, although there is also an interaction between redundancy and confidence threshold for the costs of incomplete tasks, the coefficients do not net to zero (see table 3 column 4), and the estimation excluding the 100% confidence conditions from our risk set yields an effect that is smaller but still significant both statistically and economically ($\beta = 3.4; p < .001$). This means,

[8] Think of tokens as words. They are roughly equivalent.
[9] Plots for other versions and instructions show similar trends.

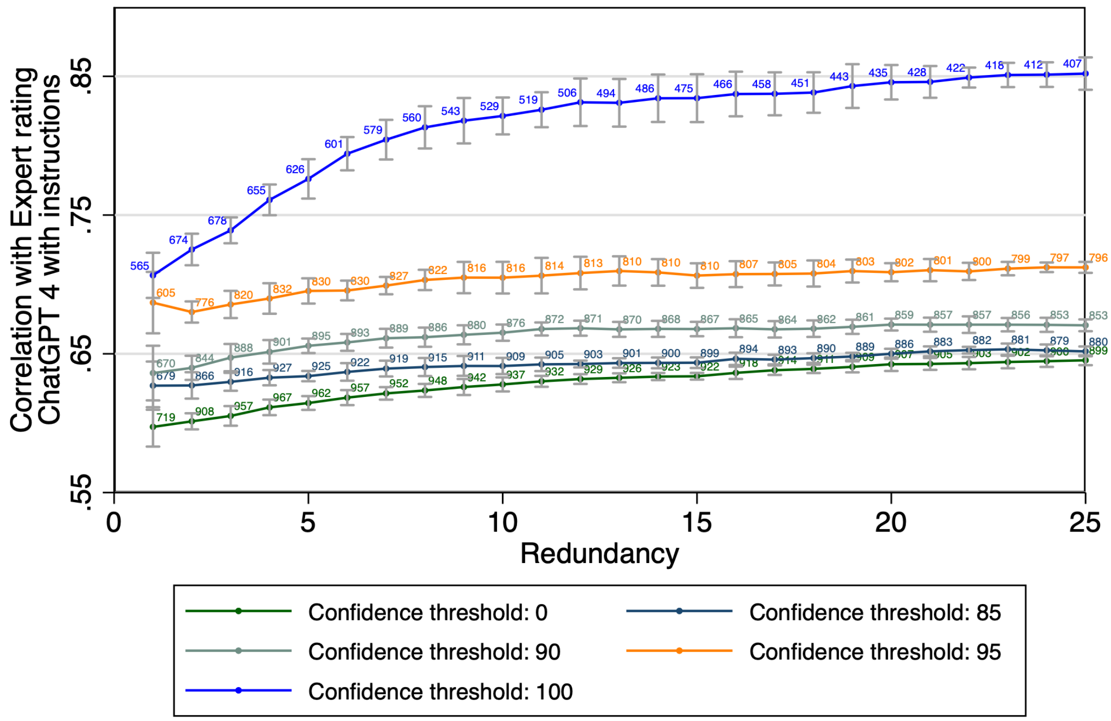

Figure 3: Accuracy by redundant runs for each confidence threshold for classification by GPT-4 with detailed instructions. Averages taken over five runs at each redundancy level. Error bars 95% CI. The average number of resolved (included) tasks are indicated at each datapoint, out of 1,022 requested. Tasks are excluded both because of stochasticity in questions the LLM answers as well as conditioning on consistent classification and confidence threshold.

No. RAs	1	2	3	4	5	6	7	8
No. Headlines	167	265	274	175	92	36	12	1

Table 4: Headline Frequency for Number of RA Reviewers

absent a high confidence threshold, redundancy yields no increase in accuracy but does increase both costs of incomplete tasks and financial costs of running the redundant tasks.

Robustness to language used in prompts: We test the robustness of our results to using alternative language in the text of the prompts, described in Appendix B. We find results are similar to those reported in table 3.

5.2 Comparing LLM and Human Worker Classification

One question might be how LLM classification accuracy compares to employing human workers to classify headlines. To this end, we asked human workers to complete the same classification task, as described in section 4.2.6. Table 4 shows the number of headlines reviewed by RA redundancy (the number of RAs that reviewed each headline), for the set of 1,022 headlines. We aggregate responses in a comparable manner to LLM classification, coding a headline as being about an acquisition if all RAs code it as such, otherwise the headline is coded as neither.

Figure 4 shows the correlation between human worker and expert classification by worker redundancy.[10] As expected, for human workers, accuracy increases with redundancy, providing a contrast to the finding for LLM classification described above. Redundancy of 2 or 3 workers leads to modest accuracy, with correlations around .5 or .6 and comparable to LLM classification by GPT-3.5 with detailed instructions or GPT-4 with basic instructions. Redundancy of 5 or 6 workers yields correlations above .7, comparable to those realized by GPT-4 with detailed instructions.

6 Guidelines

As researchers increasingly use Generative AI in labeling and classification tasks to create structured data sets there is a need for guidelines for best practices. Our findings provide important insights as to how LLMs can be best used in these tasks. They reveal areas where best practices that work with human labelers do not directly extend to LLMs (for example in terms of redundant classification). We develop guidelines based on our findings for researchers to best utilize LLMs for classification tasks, and also identify areas where LLMs may be less reliable or consistent.

[10] Headlines reviewed by 6-8 RAs are collapsed into a single group due to low numbers (see table 4).

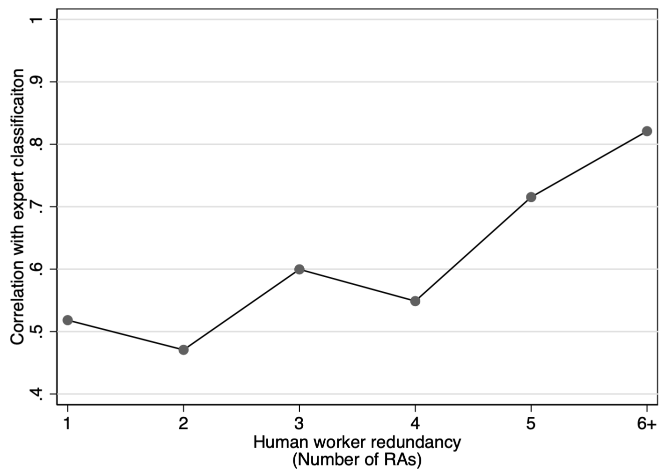

Figure 4: Correlation between human worker and expert classification by human worker redundancy (number of RAs to review a headline)

Self-Reported Confidence

Researchers should employ "self-reported confidence," pick a confidence threshold that fits the needs of their project, and use the results to identify the subset of their data that needs deeper human evaluation. A key finding from our empirical analysis is that researchers can use LLM self-reported confidence to increase classification accuracy. Of course, there is an inherent trade-off to having an LLM classify a large number of objects (quantity) versus having high accuracy in each classification (quality). Researchers therefore need to examine and pick a confidence threshold for accepting accuracy based on the needs of their project and the confidence distribution. Moreover, our findings that self-reported confidence improves accuracy means researchers can identify the subset of their data that can be satisfactorily classified by an LLM as opposed to what needs to be reviewed by a human. This should lead to higher accuracy at lower cost.

Redundancy with Multiple Trials

Unlike human classification, where multiple subjects increases accuracy, there is little value in deploying redundant trials with an LLM for classification only. A second important finding is that redundancy alone in LLM classification across multiple trials only leads to small improvements in accuracy (at best). In this way, LLMs are different from human classifiers. Figure 4 shows that human worker redundancy increases accuracy, in line with our theoretical predictions and historical research. For LLMs, however, we find little evidence that redundancy increases accuracy in the absence of a confidence threshold (see table 3 and figure 3). Based on these findings, we do not recommend that researchers incur the costs of asking an LLM to perform the same classification over multiple trials without a confidence threshold.

There is value in deploying redundant trials with an LLM to compute average self-reported confidence. At the same time, our findings show substantial increases in accuracy when considering classifications with high LLM self-reported confidence averaged across redundant trials (see table 3 and figure 3). For the highest classification accuracy, we recommend researchers ask the LLM to self-report confidence with redundancy *across multiple trials*, and implement a confidence threshold based on classifications that are both consistently classified and have high average self-reported confidence across the trials. At the same time, redundancy and a confidence threshold also increase costs of incomplete tasks. Since the efficiency frontier shifts with an advanced LLM and detailed instructions, we recommend researchers use the combination of high confidence and redundancy with those configuration dimensions. This combination also can be used in contexts where it is adequate to study a subset of data from unstructured text, or in combination with expert review for the subset of incomplete tasks.

Instructions

For nuanced classification tasks that require expertise, researchers should include detailed instructions in their prompts. Our results also suggest that, for classification tasks that are somewhat nuanced or require special handling, accuracy improves with reduced costs of incomplete tasks when the LLM is provided with detailed instructions. Our specific task, in asking whether a headline is classified as about an acquisition, was by no means difficult, but did require some background knowledge. In our case, we found that providing detailed instructions improved classification accuracy, for both the basic and advanced LLM (see figure 2, table 3). We believe many classification tasks for which researchers employ human and machine labelers are of a comparable level of difficulty and require similar handling and expertise. Based on our results, we suggest researchers create detailed prompts with clear and concise instructions, including definitions and examples, if the classification requires some level of expertise.

Version

Advanced LLMs, in our case GPT-4, provide higher accuracy with reduced costs due to incomplete tasks. At the same time, advanced LLMs also incur higher financial costs (see section 5). Evaluating this trade-off will depend on the researcher's utility function in terms of their value of high accuracy and lowering incomplete tasks. The results we present provide information for a data-driven approach for a researcher to answer this question.

7 Conclusion

The use of LLMs and Generative AI in labeling and classification tasks is in the nascent stages and will likely increase dramatically over the coming years. This creates a pressing need for guidelines and insights for best practices when using LLM classification. We propose guidelines for researchers or practitioners who wish to use an LLM for a complex labeling task, and care about balancing accuracy and costs. We set out four configuration dimensions. Two of these – redundancy and confidence – provide novel insights that, to our knowledge, have not previously been empirically tested for LLM classification.

Our empirical analysis provides important insights regarding redundancy and confidence. First, our results provide no evidence that redundancy in LLM classification increases accuracy in the absence of a confidence threshold. This is unlike human worker redundancy, and provides an important guideline to researchers to best allocate resources. Second, we introduce a novel dimension – self-reported confidence – and find that accuracy increases with stricter confidence levels. Third, we find that *redundancy in self-reported confidence* yields the most accurate results.

Our model also highlights that increasing accuracy by applying redundancy and confidence thresholds incurs the trade-off of costs based on incomplete tasks. Our model and empirical analysis provides valuable information for researchers to define their utility functions to balance this trade-off, given the unique needs of their project.

We also find that both advanced LLMs and including detailed instructions in the prompt *both* increase accuracy and reduce costs of incomplete tasks, such that these configurations shift the *efficiency frontier*. That these increase accuracy increases is in line with previous research, but the finding that they also reduce the cost of incomplete tasks is, to our knowledge, a novel contribution.

Overall, this study provides important guidelines as researchers are increasingly turning to LLMs and Generative AI, rather than using human labellers, to classify unstructured text in order to develop high-value structured data.

A Appendix: LLM Consistency in Classification Across Redundant Trials

We investigate whether consistent LLM classification across redundant trials increases accuracy. We define *consistency* as a ratio: the number of trials where the LLM classifies a headline as acquisition or neither (whichever is larger) divided by the total number of trials. Consistency ranges from .5 to 1.

We find that LLMs are highly consistent in their classification. Table 5 presents the range over trials for the percent of headlines with fully consistent classification (consistency = 1) for redundancy of 25, by condition. Even with high redundancy, the LLM is fully consistent in its classification for the overwhelming majority of headlines.

	Minimum	Maximum
GPT-3.5 basic instructions	76%	78%
GPT-3.5 detailed instructions	82%	84%
GPT-4 basic instructions	96%	97%
GPT-4 detailed instructions	88%	88%

Table 5: Minimum and maximum percent of headlines with consistency = 1 for all trials by condition.

Figure 5 presents distributions of consistency for each condition, and shows the distributions are highly skewed, with very few headlines resulting in consistency below .85 for any condition. These suggest there is not enough variance in consistency outcomes for this metric to be useful in classifying unstructured text.

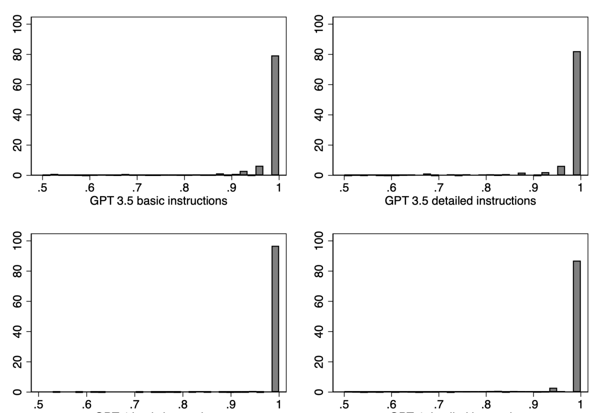

Figure 5: Distribution of consistency for redundancy of 25 for one representative trial, by condition.

We also test whether there are accuracy differences based on applying consistency thresholds, using a similar estimation approach as described in section 4.2.3, except that we include headlines based on *consistent* classification (rather than average self-reported confidence). We code LLM classification based on the

	Consistency Model
(Intercept)	$0.472\ (0.001)^{***}$
Consistency Threshold 50	$-0.017\ (0.000)^{***}$
Consistency Threshold 85	$-0.004\ (0.000)^{***}$
Consistency Threshold 95	$-0.000\ (0.000)$
Version: 4.0	$0.063\ (0.000)^{***}$
Instructions: Detailed	$0.086\ (0.000)^{***}$
Redundancy	$0.000\ (0.000)^{***}$
R^2	0.966
Adj. R^2	0.966
Num. obs.	3200

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6: OLS regression to evaluate effects of consistency.

majority across trials.[11]

Table 6 reports regression results. We find that conditioning on *consistent* classification does not meaningfully increase accuracy. Headlines with fully consistent classification (consistency = 1) only have a .004 increase in correlation compared to headlines with .85 consistency. Conditioning on full consistency (1) versus all of the data (above .5) yields a very modest increase in accuracy, a .017 increased correlation. This is 10x lower than the improvement in accuracy from self-reported confidence (see table 3). This analysis suggests LLM redundancy does not appreciably increase accuracy when conditioning only on consistent classification responses.

B Appendix: Results with Alternative Language in Prompts

One question might be whether our results depend on the language we use in our prompt. To test this, we run three alternative prompts. The alternatives do not include detailed instructions, for considerations of cost and ease of comparability. Our objective is to test whether results replicate, which we can do by focusing on the basic instructions conditions. Appendix C includes the text of the alternative prompts and describes how we developed them.

To analyze results, we employ the same method as with our base prompt (see section 4.2.5). First, we inspect the distribution of the LLMs' self-reported confidence (a percentage between 0% and 100%). Representative distributions for each version and alternative prompt are presented in figure 6. These show that the distribution varies depending on the language of the prompt, with prompt 1 and 3 presenting less

[11] We exclude the few headlines with undecided classification, or consistency exactly .5.

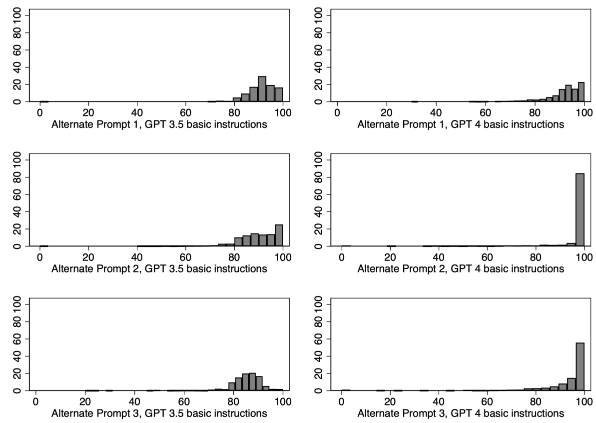

Figure 6: Distribution of LLM self-reported confidence for the three alternative prompts, by version.

skew as compared to the base prompt (see basic instructions conditions from figure 1), and prompt 2 showing a similar skew. The distributions informed the confidence thresholds for our independent variables. We chose confidence thresholds for each prompt to balance having enough data in each segment with having enough of a fine-grained threshold to interpret and compare results.[12] The varied skew presented in the self-reported confidence distributions highlight the importance of inspecting the LLM responses, and possibly testing accuracy for a subset of data, before determining the optimal confidence threshold.

Tables 7, 8, and 9 present effects of redundancy and self-reported confidence on accuracy, based on correlation with expert classification, using OLS regression for Alternative Prompts 1, 2, and 3, respectively. These regressions are identical to those presented in table 3 except the data do not include the "detailed instructions" condition (so there are 2,000 observations). Results show the same pattern: accuracy increases with self-reported confidence, and redundancy increases accuracy only for the highest confidence threshold (baseline).

[12]For the main prompt with basic instructions, about 10% and and 60% of headlines have confidence 100, the highest threshold, for GPT-3.5 and GPT-4, respectively. For Alternative Prompt 1, about 25% and 40% of headlines have confidence 95 and above, the highest threshold, for GPT-3.5 and GPT-4, respectively. For Alternative Prompt 2, about 10% and 70% of headlines have confidence 95 and above, the highest threshold, for GPT-3.5 and GPT-4, respectively. For Alternative Prompt 3, about 20% and 75% of headlines have confidence 90 and above, the highest threshold, for GPT-3.5 and GPT-4, respectively.

	Prompt 1 Model
(Intercept)	$0.655\ (0.003)^{***}$
Confidence Threshold: 0	$-0.161\ (0.005)^{***}$
Confidence Threshold: 80	$-0.200\ (0.005)^{***}$
Confidence Threshold: 85	$-0.192\ (0.005)^{***}$
Confidence Threshold: 90	$-0.151\ (0.005)^{***}$
Redundancy	$0.001\ (0.000)^{**}$
Version: 4.0	$0.120\ (0.002)^{***}$
Confidence Threshold 0 x Redundancy	$-0.000\ (0.000)$
Confidence Threshold 80 x Redundancy	$-0.001\ (0.000)^{**}$
Confidence Threshold 85 x Redundancy	$-0.002\ (0.000)^{***}$
Confidence Threshold 90 x Redundancy	$-0.001\ (0.000)^{***}$
R^2	0.896
Adj. R^2	0.896
Num. obs.	2000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 7: Alternative Prompt 1. OLS regression to evaluate effects of configuration dimensions on LLM classification accuracy (correlation) for Alternative Prompt 1. Confidence threshold effects are compared to employing a baseline of 95% and above.

	Prompt 2 Model
(Intercept)	$0.543\ (0.003)^{***}$
Confidence Threshold 0	$-0.125\ (0.004)^{***}$
Confidence Threshold 85	$-0.123\ (0.004)^{***}$
Confidence Threshold 90	$-0.095\ (0.004)^{***}$
Confidence Threshold 95	$-0.048\ (0.004)^{***}$
Redundancy	$0.001\ (0.000)^{***}$
Version: 4.0	$0.117\ (0.001)^{***}$
Confidence Threshold 0 x Redundancy	$-0.001\ (0.000)^{***}$
Confidence Threshold 85 x Redundancy	$-0.002\ (0.000)^{***}$
Confidence Threshold 90 x Redundancy	$-0.002\ (0.000)^{***}$
Confidence Threshold 95 x Redundancy	$-0.004\ (0.000)^{***}$
R^2	0.900
Adj. R^2	0.899
Num. obs.	2000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 8: Alternative Prompt 2. OLS regression to evaluate effects of configuration dimensions on LLM classification accuracy (Correlation) for Alternative Prompt 2. Confidence threshold effects are compared to employing a baseline of 100%.

	Prompt 3 Model
(Intercept)	$0.534\ (0.003)^{***}$
Confidence Threshold 0	$-0.045\ (0.004)^{***}$
Confidence Threshold 75	$-0.050\ (0.004)^{***}$
Confidence Threshold 80	$-0.073\ (0.004)^{***}$
Confidence Threshold 85	$-0.070\ (0.004)^{***}$
redundancy	$0.003\ (0.000)^{***}$
version: 4	$0.031\ (0.001)^{***}$
Confidence Threshold 0 x redundancy	$-0.001\ (0.000)^{***}$
Confidence Threshold 75 x redundancy	$-0.001\ (0.000)^{***}$
Confidence Threshold 80 x redundancy	$-0.002\ (0.000)^{***}$
Confidence Threshold 85 x redundancy	$-0.003\ (0.000)^{***}$
R^2	0.714
Adj. R^2	0.713
Num. obs.	2000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 9: Alternative Prompt 3. OLS regression to evaluate effects of configuration dimensions on LLM classification accuracy (Correlation) for Alternative Prompt 1. Confidence threshold effects are compared to employing a baseline of 90% and above.

C Appendix: LLM Execution - Prompts

The development of our prompt followed an iterative process involving multiple trials and refinements using ChatGPT. Initial versions of the prompt led to misinterpretations of the confidence measure by the LLM, necessitating continuous improvements. Providing explicit instructions on how to interpret and report confidence levels was crucial to ensure the model accurately understood the task. Each iteration aimed to clarify ambiguities and enhance the accuracy and reliability of the LLM's classification and confidence reporting. This iterative approach ensured that the final prompt was robust and capable of consistently delivering meaningful results.

The basic instructions prompt provides straightforward task instructions without additional context or examples, asking the LLM to classify headlines simply. In contrast, the detailed instructions prompt includes comprehensive definitions, examples, and specific guidelines to ensure the LLM understands nuanced classification criteria, enhancing accuracy and consistency.

Prompt - Basic Instructions

[SYSTEM] You are a helpful acquisition/merger analyst.

[USER] Read and analyze the following headline. Create a table with 4 columns.

In column 1, put ”A” if the headline describes a company acquisition, ”M” if the headline describes a company merger, or ”N” if the headline describes neither a company acquisition nor a merger.

In column 2, indicate your confidence in the previous classification (”A”, ”M”, or ”N”) in percent (value between 0% and 100%).

In column 3, identify the acquiring company (if acquisition), one of the merged companies (if merger), or ”NA” (if neither).

In column 4, identify the acquired company (if acquisition), the other merged company (if merger), or ”NA” (if neither).

Don't provide reasoning/comments. Just provide the markdown table.

This is the headline:

[HEADLINE]

Prompt - Detailed Instructions

[SYSTEM] You are a helpful acquisition/merger analyst.

[USER] Your job is to classify if a headline is about a company acquisition (”A”), a company merger (”M”), or neither a company acquisition nor a merger (”N”).

An acquisition is when one company buys another company. A merger is when two companies join together as equals.

Before you start, carefully review the following examples and instructions.

For example, this headline refers to a company acquisition: ”BEWiSynbra acquires the recycling company EcoFill.” One company, BEWiSynbra, is acquiring another, EcoFill. We code it as acquisition (”A”).

For example, this headline refers to a company merger: ”Kroger and Albertsons Companies Announce Definitive Merger Agreement”. The two companies, Kroger and Albertsons, agreed to merge. We code it as merger (”M”).

For example, this headline is not about an acquisition nor merger: ”-Nevada Geothermal Power Inc. Announces Gordon Bloomquist Stepping Down as Director”. We code it as neither (”N”).

For example, this headline refers to an acquisition, but it is unclear whether it is about a particular company acquiring another: ”Determination for taxation purposes of acquisition cost of the shares of Qt Group Plc, established through partial demerger from Digia Plc”. We code it as neither (”N”).

Sometimes a headline refers to acquiring assets or property, but not a company acquisition. This should be coded as neither (”N”).

For example, this headline refers to a company acquiring a set of programs: ”Medivir strengthens its clinical pipeline by entering into agreement to acquire a portfolio of clinical stage oncology programs”. We code it as neither (”N”).

For example, this headline refers to a company acquiring property: ”Arizona Metals Corp to Acquire Additional Private Lands at its Kay Mine Project”. We code it as neither (”N”).

Additional information: (1) Headlines that announce that an acquisition or merger is about to take place or already completed should be coded as acquisition (”A”) or merger (”M”).

(2) Headlines that announce that an acquisition or merger is not happening should be coded as acquisition (”A”) or merger (”M”).

(3) Some acquisitions are partial, meaning a company is acquiring a percentage of another company. We code these headlines as acquisition (”A”) if more than 50% were acquired, and as neither (”N”) if 50% or less were acquired or if the percentage is not clear from the headline. (4) There may be some headlines that are not straightforward in a different way. If the headline cannot be classified as company acquisition (”A”) or company merger (”M”), code it as neither (”N”).

Company Names: If the headline was coded as acquisition or merger, capture the names of the companies involved. In an acquisition, one company buys another. Sometimes the headline does not name one or both companies. For example, the headline ”DDM provides additional information related to its Hungarian acquisition” reports the name of the acquiring company (DDM) but not the name of the acquired company. In this case, leave the acquired company blank. In the case of a merger, two companies are coming together (or merging), but one is not buying another. In this case, enter the names of the companies if they are mentioned in the headline. If only one company is mentioned, enter only one name. If both companies are mentioned, enter both names.

[USER] Now start to classify headlines.

Read and analyze the following headline. Create a table with 4 columns.

In column 1, put ”A” if the headline describes a company acquisition, ”M” if the headline describes a company merger, or ”N” if the headline describes neither a company acquisition nor a merger.

In column 2, indicate your confidence in the previous classification (”A”, ”M”, or ”N”) in percent (value between 0% and 100%).

In column 3, identify the acquiring company (if acquisition), one of the merged companies (if merger), or ”NA” (if neither).

In column 4, identify the acquired company (if acquisition), the other merged company (if merger), or ”NA” (if neither).

Don’t provide reasoning/comments. Just provide the markdown table.

This is the headline:

[HEADLINE]

Alternative Prompts

To ensure the robustness of our prompt design, we asked ChatGPT to generate alternative versions of the prompt based on a variety of objectives: “Simplified Instruction with Clear Steps,” “Direct and Concise Format,” and “Structured and Detailed Instructions.” This involved instructing ChatGPT to create different prompt formulations that maintained the core instructions but varied in language and structure. By comparing the performance of these alternative prompts, we aimed to evaluate whether the specific wording of the prompt influenced the LLM’s accuracy and confidence reporting. This additional testing helped verify that our results were not overly dependent on the particular phrasing used and that the guidelines developed were broadly applicable.

Alternative Prompt 1: Simplified Instruction with Clear Steps

[SYSTEM] You are a helpful acquisition/merger analyst.

[USER] Analyze the given headline to determine if it's about a company acquisition (A), a merger (M), or neither (N).

Create a markdown table with 4 columns:

1) Classification (A, M, N)

2) Confidence Level (0-100%)

3) Acquiring/Merged Company (or 'NA')

4) Acquired/Other Merged Company (or 'NA')

No additional comments needed.

Headline:

[HEADLINE]

Alternative Prompt 2: Direct and Concise Format

[SYSTEM] You are a helpful acquisition/merger analyst.

[USER] Given this headline, fill out a 4-column markdown table:

Column 1: Type (A for acquisition, M for merger, N for neither)

Column 2: Confidence (0-100%)

Column 3: Acquiring or Merged Company (or 'NA')

Column 4: Acquired or Other Merged Company (or 'NA')

Avoid explanations. Just fill the table.

Headline:

[HEADLINE]

Alternative Prompt 3: Structured and Detailed Instructions

[SYSTEM] You are a helpful acquisition/merger analyst.

[USER] For the following headline, create a markdown table to classify and detail the news. The table should have 4 columns:

1) Classification: 'A' for acquisition, 'M' for merger, 'N' for neither.

2) Confidence Level: Indicate your certainty in percentage (0-100%).

3) Company Involved: Name the acquiring or merged company, or 'NA'.

4) Counterpart Company: Name the acquired or other merged company, or 'NA'.

Provide only the table without any commentary.

Headline:

[HEADLINE]

References

Agley, Jon, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo (2022). "Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7". In: *Behavior research methods* 54.2, pp. 885–897.

Alyakoob, Mohammed and Mohammad Saifur Rahman (2022). "Market Design Choices, Racial Discrimination, and Equitable Micro-Entrepreneurship in Digital Marketplaces".

Boland, Philip J (1989). "Majority systems and the Condorcet jury theorem". In: *Journal of the Royal Statistical Society Series D: The Statistician* 38.3, pp. 181–189.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond (2023). *Generative AI at work*. Tech. rep. National Bureau of Economic Research.

Cegin, Jan, Jakub Simko, and Peter Brusilovsky (2023). "ChatGPT to Replace Crowdsourcing of Paraphrases for Intent Classification: Higher Diversity and Comparable Model Robustness". In: *arXiv preprint arXiv:2305.12947*.

Chen, Zenan and Jason Chan (2023). "Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise". In: *Available at SSRN 4575598*.

Chen, Zhikai, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang (2023). "Label-free node classification on graphs with large language models (LLMs)". In: *arXiv preprint arXiv:2310.04668*.

Del Arco, Flor Miriam Plaza, Debora Nozza, and Dirk Hovy (2024). "Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation". In: *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives)@ LREC-COLING 2024*, pp. 19–30.

Fügener, Andreas, Jörn Grahl, Alok Gupta, and Wolfgang Ketter (2022). “Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation”. In: *Information Systems Research* 33.2, pp. 678–696.

Geiger, R Stuart, Dominique Cope, Jamie Ip, Marsha Lotosh, Aayush Shah, Jenny Weng, and Rebekah Tang (2021). “” Garbage In, Garbage Out” Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data?” In: *arXiv preprint arXiv:2107.02278*.

Kolluri, Johnson, Dr Shaik Razia, and Soumya Ranjan Nayak (2019). “Text classification using machine learning and deep learning models”. In: *International Conference on Artificial Intelligence in Manufacturing & Renewable Energy (ICAIMRE)*.

Le Mens, Faruk EnesGael, Balazs Kovacs, Michael Hannan, and Guillem Pros (2023). “Uncovering the semantics of concepts using GPT-4”. In: *PNAS* 120.49, pp. 885–897.

Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary (2023). “Language Models for Automated Market Research: A New Way to Generate Perceptual Maps”.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig (2023). “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing”. In: *ACM Computing Surveys* 55.9, pp. 1–35.

Ringel, Daniel (2023). “Creating Synthetic Experts with Generative Artificial Intelligence”.

Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha (2024). “A systematic survey of prompt engineering in large language models: Techniques and applications”. In: *arXiv preprint arXiv:2402.07927*.

Zhou, H., Z. Liu, C. Lang, Y. Xu, Y. Lin, and J. Hou (2021). “Improving the recall of biomedical named entity recognition with label re-correction and knowledge distillation”. In: *BMC bioinformatics* 22 (1).