# Confidence Over Redundancy: Operational Guidelines for Text Classification with Large Language Models

Jörn Boehnke Elizabeth Pontikes Hemant K. Bhargava

Graduate School of Management, University of California, Davis

**Abstract.** Organizations increasingly delegate operational work to large language models, including the classification of unstructured text such as clinical notes, regulatory filings, corporate communications. This paper develops empirically grounded guidelines for that practice and proposes a decision framework that treats classification-workflow design as an operational decision, addressing an open question in the use of large language models for classification: which cases should the model decide, which should be routed to human review, and at what cost? In the framework, a decision maker trades accuracy on resolved tasks against the cost of unresolved tasks and the computational cost of repeated model runs. We evaluate two primary configuration levers—self-reported confidence thresholds and redundancy (repeated runs of the same task)—in a controlled computational experiment spanning two applied classification tasks, corporate-acquisition press-release headlines and clickbait news headlines, and five large language models. We find that confidence thresholds reliably and substantially increase classification accuracy, albeit at the cost of more unresolved tasks, thus presenting practitioners an operational decision problem in optimizing the LLM-human workflow. Redundancy, although well established as effective at improving accuracy under human labeling, does not meaningfully improve accuracy while still increasing unresolved tasks. Combining a simple model of redundancy with the empirical evidence, we posit that this discrepancy stems from errors that are positively correlated within and across models—plausibly reflecting shared training data and architectures—violating the independence assumption that underlies the accuracy benefits of redundancy in human labeling. We translate these results into guidelines for a workflow design to create structured data from unstructured text: partition the corpus by self-reported confidence, accept model classifications for high-confidence cases, and route the remainder to human review, raising accuracy on the automated subset while concentrating human review where it adds the most value.

**Keywords:** text classification; large language models; quality control; human-machine routing; unresolved tasks

## 1 Introduction

Artificial intelligence (AI) systems, particularly large language models (LLMs), are increasingly used for large scale classification and labeling tasks to transform unstructured text into structured data (Dai and Swaminathan, 2026; Mithas et al., 2022). From clinical notes to regulatory filings to corporate communications, business operations generate large amounts of unstructured textual data that contain valuable insights but are not amenable to traditional quantitative analysis (Koleck et al., 2019; Chen et al., 2012; Li and Hitt, 2008; Liu and Lopez, 2016; Abrahams et al., 2015; Feng and Shanthikumar, 2018). For decades, practitioners and researchers have relied on human labelers—often undergraduate students or crowd-sourced workers—to classify text into predefined categories. This human-centered approach has proven effective in many contexts, but suffers from well-documented limitations, including cost, inconsistency, and scalability constraints (Geiger et al., 2021; Khare et al., 2015).

The emergence of LLMs has transformed how organizations and researchers approach this challenge. AI systems offer the promise of automated text classification on an unprecedented scale and at potentially lower cost than human alternatives (Chen et al., 2024; Kolluri et al., 2019; Cegin et al., 2023; Li et al., 2024b; Terwiesch et al., 2026). However, the practical adoption of LLMs for classification tasks has proceeded largely without the theoretical foundations and empirical guidelines that inform human labeling practices. In contrast to human labeling, where decades of research have established best practices—including worker selection, training protocols, and quality assurance mechanisms—comparable frameworks for LLM-based classification remain underdeveloped. This gap is particularly concerning given the known limitations of LLMs, including their propensity to generate plausible but incorrect outputs and their sensitivity to prompt design and configuration choices.

This paper develops empirically grounded guidelines for using LLMs to classify unstructured text, treating their configuration as a workflow design problem. We focus on the later stage of the workflow when LLMs are put into practice and used by researchers and organizations to create structured data. The question underlying our research is: *given that LLMs will perform imperfectly, especially at classification tasks that have inherent ambiguity, how should an organization configure its classification workflow—which cases does the LLM decide, which are routed to human review—and at what cost?*

We conduct extensive computational experiments comparing two accessible mechanisms for quality control: *redundancy*, which requires agreement across repeated classifications, and *self-reported confidence*, which filters classifications using a numeric score elicited in the prompt. Both are simple levers that practitioners can readily implement in real-world LLM classification pipelines. Redundancy draws on established human-labeling practices (Ipeirotis et al., 2010; Agley et al., 2022); confidence filtering builds on selective prediction and LLM confidence elicitation (Geifman and El-Yaniv, 2017; Hendrickx et al., 2024; Tian et al., 2023; Xiong et al., 2024). We systematically evaluate how these mechanisms perform across models, tasks, and implementation choices. Our focus is the operational choice between them: which should practitioners prioritize when improving the quality of accepted labels also creates unresolved work and additional execution costs?

Our study contributes to the literature by providing a research-based framework based on controlled experiments in two empirical contexts. Our first study context involves classification of press releases as about corporate acquisitions or not based on headline text. This is an example of a business research challenge, where scholars seek first to identify, from a large corpus, which are the relevant documents for a topic of study (for example studying acquisition patterns). Accurately labeling which texts comprise the corpus is complex and ambiguous task, as examples in Table 1 illustrate. Our second context involves the classification of news headlines as clickbait or not, an important task for organizations and researchers who seek to filter relevant news. This also presents labeling ambiguity; closely related news- and content-screening classification problems have been studied as operational quality problems in operations management (Lee and Ram, 2026; Ng et al., 2023; Boehnke et al., 2024). Headlines may use attention-grabbing phrasing without being deceptive, or may appear straightforward while omitting critical context, as Table 1 illustrates. Both classification contexts involve moderate complexity and inherent ambiguity, making them representative of the classification challenges researchers commonly face when coding unstructured text.

Our focus on employing self-reported confidence and redundancy to increase accuracy draws from and

**Table 1.** Sample headlines for the two classification tasks. Panel A shows press release headlines potentially about a corporate acquisition; the task to the LLM is to determine whether each headline represents a corporate acquisition (or merger) or not. These headlines are drawn from the 1,022-headline expert-coded analysis sample, and the Classification column gives the expert ground-truth label against which LLM accuracy is scored. Panel B shows news headlines labeled clickbait or straight news; the task to the LLM is to determine whether each headline is clickbait. These headlines are drawn from the 1,000-headline analysis sample, and the Classification column gives the ground-truth label against which LLM accuracy is scored; these labels follow the publishing outlet, not independent annotation.

**Panel A: Press Release Headlines (Acquisition Task)**

| Press Release Headline | Classification | Comment |
|---|---|---|
| BWI - Acquisition of Beston Pacific by Challen | Acquisition | A terse wire-service headline; once unpacked, one company (Challen) acquires another (Beston Pacific). |
| Coram Completes Sale of Rx Unit | Neither | The sale of a business unit, not of a free-standing company. |
| Barry Callebaut finalizes sale of U.S. subsidiary | Acquisition | By contrast, a subsidiary is itself a company, so the sale of one is a corporate acquisition. |
| Industrivarden Acquires 15% Stake in Ossur | Neither | A 15% stake is a minority position, short of the control that defines an acquisition. |
| Atlas Acquires Option to Purchase Fluorite Property in Mexico | Neither | An option rather than a purchase, and over a property rather than a company. |
| Telia Purchases U.K. Assets of 360 Networks | Neither | A purchase of assets, not of the company that holds them. |

**Panel B: News Headlines (Clickbait Task)**

| News Headline | Classification | Comment |
|---|---|---|
| 18 Facts You Can't Go Another Day Without Knowing. #4 Will Change The Way You Think | Clickbait | A prototypical numbered listicle whose hook is a pure curiosity gap, with no information about the content. |
| Chhattisgarh: Two injured in pressure bomb blast in Sukma district | News | A prototypical factual report with a place-name dateline and no rhetorical hook. |
| A Rich Guy Bought A Water Tower From 1938. What Happened Next Is Ridiculously Awesome | Clickbait | The headline withholds its outcome behind a curiosity-gap teaser ("what happened next") and a vague superlative. |
| Tumour-seeking salmonella may combat deadly brain cancer | News | A research finding stated plainly and directly, with no rhetorical hook. |
| 20 Animals You Won't Believe Actually Exist. One of Them Will Haunt You Forever. | Clickbait | A numbered listicle that pairs a "you won't believe" dare with an emotional teaser. |
| Goldman Sachs stalls plan to move jobs to UK amid Brexit uncertainty | News | A factual business report naming the company, the action, and the context, with no sensational framing. |

contributes to a large, interdisciplinary body of research. The idea of asking LLMs to self-report their estimated confidence is based in established practice in human-performed labeling when content is ambiguous or complex. However, human self-reported confidence is notoriously misaligned with actual accuracy, a phenomenon extensively documented across behavioral economics, cognitive psychology, and management science (Moore and Healy, 2008; Ipeirotis et al., 2010). For LLMs, there is some reason to have faith in confidence self-reporting: when generating a classification or other response, an LLM has internal signals about how likely that answer is to be correct. Although the internal probability signal may not extrapolate to an external-facing confidence estimate (i.e., may not have cardinal value), it does have ordinal or directional

value, so that responses tagged with high confidence are likely to be more reliable than those tagged with low confidence. Accepting answers only above a confidence threshold aligns with selective prediction (Geifman and El-Yaniv, 2017) and with LLM-specific work on hallucination mitigation through response rejection (Xu et al., 2024; Cao, 2024). In fact, recent work suggests LLMs can provide meaningful self-evaluations of correctness, with verbalized confidence helping rank predictions by reliability (Kadavath et al., 2022; Tian et al., 2023). Whether this translates to accuracy gains in applied classification tasks remains an open question that we address empirically.

The motivation to investigate *redundancy* in LLM labeling reflects the fact that the use of redundancy is one of the most established principles in human labeling. Practitioners use multiple labelers for the same classification task and then apply majority decision rules to aggregate their judgments (Ipeirotis et al., 2010; Agley et al., 2022; Alyakoob and Rahman, 2026; Nair and Huchzermeier, 2024). This approach has strong theoretical foundations: Condorcet's Jury Theorem demonstrates that aggregating decisions from multiple fallible agents can outperform any single agent (Condorcet, 1785), and decades of empirical work confirms redundancy reliably increases accuracy for human labelers (Boland, 1989; Ipeirotis et al., 2010). For instance, in a binary classification task where individual labelers achieve 55% accuracy, employing three labelers can increase accuracy to 58% under a simple majority rule (2 out of 3 agree) and 65% under unanimity. However, these accuracy gains are often accompanied by costs from leaving some tasks unresolved, with approximately 75% of tasks failing to achieve unanimity.[1] In this study, we ask whether principles of redundancy, developed for human labelers, apply to LLM-based classification systems.

When using LLMs for classification, redundancy can take two forms: multiple classifications by the same LLM (*within-LLM redundancy*), and using multiple different LLMs for the same task (*across-LLM redundancy*). To date, there is limited evidence on how either form affects accuracy when in practice-based classification tasks aimed at creating structured data for study. There are conceptual reasons to be skeptical of both. Within-LLM redundancy suffers from the likely failure of independence (the first i in i.i.d.—independent and identically distributed), whereas across-LLM redundancy has the potential failure of identical in i.i.d. While recent work shows that aggregating predictions across LLMs can improve accuracy in forecasting and prediction market contexts (Schoenegger et al., 2024), it remains an open question whether such gains extend to classification tasks, where errors stem from ambiguity in interpreting unstructured text rather than uncertainty about future events. Moreover, across-LLM redundancy introduces practical and conceptual costs: different LLMs require tailored prompts and configurations for optimal performance, which makes comparisons difficult, and imposes higher requirements on cost and expertise. Accordingly, we focus on within-LLM redundancy as our primary analysis—a controlled design that isolates the effects of stochastic variation while holding model architecture and configuration constant—and conduct supplementary analyses of across-LLM redundancy to assess its comparative performance.

Our central finding, which we establish through experiments across two distinct classification contexts, is that confidence thresholds are a reliable and substantial lever for improving LLM classification accuracy, whereas redundancy alone—a cornerstone of human labeling practice—provides no meaningful benefit for LLMs. Our findings about the benefits of confidence thresholds compliment recent work which suggests the plausibility of LLMs to uncover confidence in their outputs (Kadavath et al., 2022). Our experiments

[1] With a majority decision rule there are no unresolved tasks but also smaller accuracy gains.

leverage this technical understanding and examine whether confidence thresholds can benefit practical work in text classification. Our findings about redundancy warn practitioners against the temptation to mimic this lever, which succeeds when humans do classification (Sheng et al., 2008; Snow et al., 2008) and is considered plausible even for LLMs especially with ensemble of models (Wang et al., 2023, 2024). Conditioning on high self-reported confidence yields substantial accuracy gains—up to 29% improvement—with a clear trade-off in unresolved tasks, providing evidence that higher confidence reflects more certain internal probability signals and thus predicts more accurate classifications. In contrast, within-LLM redundancy does not consistently deliver higher accuracy even while imposing higher computational and unresolved task costs—a striking finding given that redundancy is a natural and tempting intervention for researchers familiar with human labeling practice. This is likely because positive correlation across repeated runs means systematic biases persist rather than average away. Across-LLM redundancy similarly shows little benefit, which possibly reflects shared architectures and overlapping training data across models.

Our central finding is that the two quality-control approaches perform very differently. Self-reported confidence provides a useful and consistent way to improve the quality of accepted classifications, whereas redundancy does not deliver comparable gains. In the main pooled estimates, moving from the lowest to the highest confidence threshold considered in each task increases accuracy correlation by 0.156 for acquisition classification and 0.183 for clickbait classification, while leaving more tasks unresolved. Repeated classification produces less consistent accuracy gains and requires additional model runs. These findings challenge the routine transfer of redundancy-based quality control from human labeling to LLMs. They support using confidence filtering as the starting point for workflow design, with repeated classification used when its incremental benefits justify its costs.

Our contribution is to provide an empirical basis for this approach and a decision framework for implementing it. The framework jointly evaluates the quality of accepted labels, unresolved workload, and execution requirements. It enables practitioners to choose acceptance rules around review capacity and application-specific costs, rather than maximizing classification accuracy in isolation. The resulting partition-and-route workflow accepts classifications that meet the chosen standard and directs the remainder to further review, connecting our findings to gatekeeper and targeted-automation designs in operations management (Ahsen et al., 2026; Dai and Singh, 2026; Imdahl et al., 2026). The same framework can be updated as models change by re-estimating these trade-offs on the target task.

## 2 Related Literature

We build our framework on three streams of literature: (1) human-centered labeling and its established best practices, (2) redundancy in both human and LLM labeling, and (3) confidence estimation in LLMs and its potential as a practical accuracy lever.

### 2.1 Labeling by Humans and Machines

Researchers have long relied on human coders—graduate and undergraduate assistants—to classify unstructured text: online product reviews (Humphreys, 2010), strategic change narratives (Sonenshein, 2010), book reviews (Kovács and Sharkey, 2014), corporate communications (Fiss and Zajac, 2006), legitimation narratives in multinational mergers (Vaara and Tienari, 2011), and themes in postmerger integration interviews

(Monin et al., 2013). Such efforts are labor-intensive, costly, and subject to human error.

LLMs promise to classify unstructured text at scale, potentially with higher accuracy and lower cost. Earlier transformer models like BERT (Devlin et al., 2019) perform well on classification tasks with labeled training data and unambiguous categories, but the tasks researchers face in practice often involve contextual reasoning, domain knowledge, and inherent ambiguity—conditions under which LLMs offer distinct advantages. Several studies confirm these advantages: Cegin et al. (2023) report that GPT creates more robust and diverse paraphrases than human labelers for intent classification; Ringel (2023) finds that GPT-4 labels of marketing mix variables in consumers' Twitter posts—unlike crowd-sourced 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; and Li et al. (2024b) find that with suitable prompt engineering, generative AI models match the performance of human surveys in generating perceptual maps, at lower cost.[2]

### 2.2 Redundancy in Human Labeling

Human labeling of complex, unstructured data typically draws on undergraduate research assistants or paid crowdworkers whose topical expertise is limited (Snow et al., 2008). Redundancy is the standard quality-control response: each task is assigned to multiple labelers, and a decision rule selects the accepted label. Alyakoob and Rahman (2026), for example, used a sequential protocol that starts with 3 workers per task and escalates to up to 10 when disagreements persist, accepting a label only when a sufficient supermajority agrees. The idea descends from Condorcet's demonstration 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): if each labeler is independently correct with some probability, the chance that multiple labelers all make the same mistake is lower than the chance that any one of them does, so requiring agreement makes the accepted answer more likely to be right. The gain comes at a cost—tasks on which labelers disagree go unresolved—and both the accuracy gain and the unresolved share grow with the required level of agreement. These relationships assume labeler errors are independent and identically distributed (i.i.d.) across labelers and tasks, which is reasonable for humans working independently but, as we discuss below, less certain for LLM redundancy. This theoretical foundation provides the benchmark against which we evaluate LLM redundancy.

### 2.3 Redundancy in Labeling by LLMs

Whether redundancy similarly increases accuracy in LLM labeling is an open question, and early evidence is mixed. Le Mens et al. (2023) use multiple GPT-4 queries to classify book descriptions—a task involving meaningful subjectivity—and find only limited accuracy improvements relative to a single run. In contrast, Li et al. (2024a) show that accuracy scales with redundancy through sampling-and-voting in tasks with objectively verifiable answers—but this finding that may not extend to tasks where errors stem from ambiguity rather than randomness. Aggregating predictions across LLMs can improve accuracy in forecasting contexts (Schoenegger et al., 2024), and Plaza-del Arco et al. (2024) connect the wisdom-of-crowds logic directly to

[2]Because prompt design—few-shot exemplars, chain-of-thought prompts, and structured output requirements—materially shapes LLM behavior on classification tasks (Brown et al., 2020; Wei et al., 2022; Sahoo et al., 2024), our experiments include prompting style as a controlled treatment, directly contrasting zero-shot prompting (no worked examples) with few-shot prompting (with worked examples) within each task.

crowds of instruction-tuned language models in annotation settings, and cross-model agreement is already used as an informal validity check in applied empirical work (Bodoh-Creed et al., 2026), but systematic evidence for interpretive classification tasks remains limited. Historically, algorithms produced fixed results, obviating the need for redundancy. However, LLMs introduce stochasticity through parameters such as temperature, so repeated runs of the same prompt can disagree, creating the potential for aggregation gains that parallel humans—but only to the extent that it generates independent variation rather than correlated errors.

Ensemble theory sharpens this question. Aggregation improves accuracy only to the extent that the aggregated classifiers make diverse—weakly correlated—errors; combining reduces variance but cannot remove bias that the classifiers share (Zhou, 2025). The same logic governs when human and machine judgments complement each other in operational forecasting (Nair and Huchzermeier, 2024). LLMs that share transformer architectures and overlapping training corpora are natural candidates for shared bias, so ensemble theory predicts limited gains from repeated runs and from cross-model ensembles alike. A central question is therefore whether, by how much, and at what cost in unresolved tasks LLM redundancy improves classification accuracy in practice. We focus on within-LLM redundancy—the same model queried repeatedly under fixed prompts and configurations, isolating the effect of stochastic variation—with across-LLM redundancy as a secondary comparison.

### 2.4 Self-Reported Confidence

Compared to redundancy, self-reported confidence has received less attention in human labeling, but it holds particular promise for LLMs, which can draw on internal probability signals to assess the reliability of their own outputs. Communicating confidence can increase trust in LLM outputs (Foroughifar et al., 2025), and selective classification establishes that abstaining on low-confidence cases raises accuracy on the retained set whenever the confidence signal is valid (Geifman and El-Yaniv, 2017; Hendrickx et al., 2024). Language models can indeed provide meaningful self-evaluations of correctness, with verbalized confidence helping rank predictions by reliability (Kadavath et al., 2022; Tian et al., 2023). A related machine-learning literature studies confidence calibration—the alignment of predicted probabilities with observed accuracy (Geng et al., 2024)—and how best to elicit confidence statements (Xiong et al., 2024); Ni et al. (2024) find that probabilistic confidence is generally more reliable than verbalized confidence. However, verbalized confidence (asking the model to report a numeric estimate) is more accessible to applied researchers who do not have access to model internals. It is an open question whether such practitioners can increase accuracy (and minimize costs of unresolved tasks) by filtering on self-reported verbalized confidence when using LLMs to label unstructured data.

## 3 A Decision Framework for Balancing Accuracy and Costs

In a project where an LLM is deployed for large-scale labeling of unstructured data, a practitioner aims to balance maximizing *accuracy* of the LLM's work while minimizing its *cost*. We develop a two-part decision framework to navigate these trade-offs. Section 3.1 derives formal expressions for accuracy and completion rates (or, inversely, unresolved tasks) under i.i.d. assumptions for both redundancy and confidence thresholds, then characterizes how positive correlation across LLM runs causes observed performance to deviate from

those predictions. We test these directional predictions empirically via per-LLM split-sample regressions in Section 5. Section 3.2 develops the practitioner-facing decision framework, mapping configuration choices to the accuracy-cost trade-offs each implies.

Our two primary configuration dimensions are redundancy level ($r$) and confidence threshold ($d$). Both operate by selecting on likely classification accuracy: redundancy through requiring response consistency across repeated runs, and confidence through the LLM's self-reported certainty. Applying either lever creates a subset of tasks where the signal is strong enough to accept the classification, leaving an unresolved subset that requires further review. We treat unresolved tasks as a primary cost, alongside the immediate execution costs of running additional LLM queries.

- **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. Compute accuracy $a(r, \cdot)$ 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*, and accept only those responses that exceed a defined *confidence threshold* $d$. Compute accuracy $a(\cdot, d)$ based on the performance of this subset of tasks against ground truth.

Additional configuration dimensions—LLM choice and prompting style—are not primary optimization parameters but are included as controls.

### 3.1 Redundancy: Accuracy, Completion, and Correlation Structure

Suppose that a task is performed $r$ times, a response is accepted only when the LLM's self-reported confidence is at least $d$, and the task is considered complete when at least $m(> r/2)$ of the $r$ responses are identical. Let $p$ denote the probability that a $d$-confidence response is accurate and $X \sim \text{Binomial}(r, p)$ the number of correct responses, with $B(r, k) = \binom{r}{k} p^k (1-p)^{r-k}$ denoting the probability of exactly $k$ correct responses. We write $C(r, p, m)$ for the probability that the task is completed and $A(r, p, m)$ for the conditional accuracy of the consensus response given completion. Under independent and identically distributed (i.i.d.) responses across the $r$ runs, $C$ and $A$ take the closed forms given in Eq. 1 below. The rules for special types of majority can be obtained by substitution (for simple majority, $m = \lceil (r+1)/2 \rceil$; for two-thirds majority, replace $m$ with $\lceil 2r/3 \rceil$). If the independence assumption truly holds, then (1) $C$ is strictly decreasing in $r$, and rapidly (convex, and with higher curvature for higher $m$), (2) $A$ is increasing in $r$ for all $p > 1/2$, but gains diminish as $r$ increases, hence, (3) the trade-off between $C$ and $A$ becomes more severe as $r$ increases.

| | $C$ (completion rate) | $A$ (conditional accuracy) | |
|---|---|---|---|
| General majority: | $C_{\text{maj}} = \sum_{k=m}^{r} B(r,k) + \sum_{k=0}^{r-m} B(r,k)$ | $A_{\text{maj}} = \left(\sum_{k=m}^{r} B(r,k)\right) \Big/ C_{\text{maj}}$ | (1a) |
| Unanimity: | $C_{\text{un}} = p^r + (1-p)^r$ | $A_{\text{un}} = \dfrac{p^r}{p^r + (1-p)^r}$ | (1b) |

However, when applying redundancy with LLMs, we posit that multiple runs for the same LLM will force at least some positive correlation across runs, causing empirical outcomes to differ from the predictions of Eq. 1.[3] An LLM with moderate correlation across repeat queries (but not close to 1) should deliver less variation in both $C$ and $A$, while still satisfying the properties (rapidly decreasing $C$, concave increase in $A$, and increasingly severe trade-off). Note that the decrease in $C$ will persist more even when $A$ fails to increase much (for example under unanimity even one deviant response will prevent completion, whereas $A$ is a ratio), thus worsening the accuracy-cost trade-off from redundancy. Very high correlation will fail to produce gains in $A$ while still causing some decrease in $C$. At the extreme, a correlation close to 1 will render both $A$ and $C$ relatively flat, as repeated runs become nearly identical and the agreement requirement loses its discriminating power.

These computations and Eq. 1 can be applied to make inferences from empirical observations as follows. The $p$ rate for each LLM can be approximated from single-run performance (i.e., $A \approx p$ when $r = 1$). Then, inverting the properties (1–3) above, a slower rate of increase in $A$ and rate of decrease in $C$ (than predicted in Eq. 1), combined with lower curvature, will indicate moderate positive correlation across multiple runs. For higher correlation, $A$ will be relatively unchanged but $C$ will still exhibit some decrease. A relatively flat $C$ and $A$ will signal that multiple runs of an LLM have very high correlation, rendering redundancy meaningless. The confidence threshold enters these expressions through $p$: if LLMs report confidence with meaningful ordinal value, then $p$—and with it $A$—should be increasing in $d$, a prediction Section 5 tests directly.

Section 5 uses the i.i.d. expressions above as a directional benchmark. The correlation structure derived here implies a testable prediction: the redundancy coefficient should be more positive for an LLM whose repeated runs are less correlated, and near zero or negative for an LLM whose runs are strongly correlated. Testing it requires a prior about which LLMs carry the weaker correlation, which we obtain by holding the model family fixed and contrasting a reasoning model with a standard instruction-following model from that same family: a reasoning model follows a more variable deliberation path before committing to an answer, so its errors should be less correlated across repeated runs. In our set of LLMs this contrast is Qwen 3 (a reasoning model) against Qwen 2.5 (a standard instruction-following model), both from the Qwen family, so that the reasoning capability—rather than differences in architecture, developer, or training corpus—is the salient difference. Accordingly, that section reports per-LLM split-sample regressions (Tables 4 and 5) and reads the Qwen 2.5 vs. Qwen 3 contrast—not a coefficient spread across architecturally unrelated LLMs—as the directional test of this prediction.

### 3.2 Decision Framework

An LLM configuration is a specific combination of $(r, d)$, and the number of combinations depends on the choice set within each dimension. Each configuration imposes an immediate execution cost $x(r, d)$, and delivers an associated accuracy level $a(r, d) \in [0, 1]$ for tasks it can resolve, but it leaves a number $n_u(r, d) \geq 0$ of unresolved tasks (a count, matching the dependent variable in Section 4.3) because the classification does not satisfy both the confidence threshold $d$ and the decision rule across the $r$ repetitions. These unresolved tasks impose a delayed cost $DC(n_u(r, d))$ that aggregates several organizationally distinct components: the direct cost of human review for routed cases, the opportunity cost of the expert attention that

[3] High temperature will reduce correlation but also decrease $p$.

review consumes, the latency of waiting for resolution, and—where unresolved tasks are closed by a fallback rule rather than by review—the expected cost of misclassification, which is typically asymmetric between false positives and false negatives. The execution cost $x(r, d)$ likewise bundles compute or per-query fees with wall-clock time, both of which scale with $r$. These functions can be estimated by running all configurations over a sample of tasks.

We propose practitioners should define their utility function for accuracy $U(a(r, d))$, along with execution and unresolved-task cost functions. Utility for accuracy should be determined considering 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 $\frac{2}{1+e^{-a}} - 1$ can be used to get increasing returns with an asymptotic maximum.

The practitioner needs to pick a best-fit configuration over the space of configurations. This best-fit configuration should maximize net utility $(U(a(r, d)) - w_1 x(r, d) - w_2 DC(n_u(r, d)))$, where $w_j$ 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 results to characterize these trade-offs empirically, providing the inputs practitioners need to define their own utility functions and select a configuration that best fits their needs.

A worked illustration shows how these weights move the optimum. On the acquisition task, tightening the threshold from $d = 95\%$ to $d = 100\%$ buys a 0.096 gain in accuracy correlation at the price of roughly 158 additional unresolved tasks routed to review (Table 2, columns 1 and 4). At a per-review cost of $c$ per routed case, the tighter threshold is preferred whenever the utility gain from higher accuracy on the resolved set exceeds $158\,c$ plus the associated delay costs. An organization with asymmetric error costs—say, a compliance screen in which false negatives are expensive—will rationally accept a much larger review bill than one labeling data for exploratory analysis, so the optimal $d$ shifts with the error-cost ratio even though the accuracy-coverage curve is unchanged. In this sense the confidence threshold plays the role of an acceptance-sampling parameter in quality control: it sets which units the automated process may pass and which are held for inspection.

## 4 Experiment Design and Data

We test our ideas using two labeling tasks on two text datasets that are characteristic of how organizations and researchers use LLMs to structure data from text. The first task is to classify press release headlines drawn from a LexisNexis corpus as being about an acquisition or not. The second task is to classify a set of news headlines from the Clickbait-Detector corpus (Mathur, 2020) as clickbait or news. Both tasks are moderately complex applied labeling problems that involve short headline-length text and binary classification outcomes. Classifying headlines, rather than full text, is a typical task for researchers who aim to sort through large corpora of data to identify a subset of data for further study and do not have access to or the necessary computational resources for the full text. We ask each LLM to classify each headline under the conditions outlined above.

### 4.1 Lexis Nexis Dataset of Press Releases

The data for the first task is drawn from a large-scale corpus of press release headlines obtained from LexisNexis under a commercial data license (similar content is available through LexisNexis products such as Nexis Uni), comprising over 27 million observations from major press release distributors such as PR Newswire and Business Wire. This corpus spans several decades and represents a broad spectrum of corporate communications.

We evaluate the performance of LLMs in classifying whether a headline pertains to a corporate acquisition (or not). With 27 million headlines, drawing a random sample large enough to yield a sufficient number of acquisition-related observations for LLM labeling is cost-prohibitive. We therefore apply automated methods as an initial filter to construct a set of headlines that oversamples those likely to be acquisition-related.[4]

We construct a balanced sample of roughly 50% acquisition-related and 50% non-acquisition-related headlines.

The prompt for the classification task instructs the LLM to classify whether each headline is about an acquisition or not and to report a corresponding confidence score (0–100), based on a definition of acquisition-related activity and a standardized output format (see D for full details).

### 4.2 Clickbait Headlines

The data for the second task is drawn from the Clickbait-Detector corpus published by Mathur (2020). This dataset contains 11,717 English-language headlines classified as clickbait (5,637; 48.1%) or legitimate news (6,080; 51.9%). Headlines were collected from specific online outlets, with labels determined by outlet identity: clickbait headlines originate from outlets such as BuzzFeed and The Huffington Post, while news headlines come from outlets such as The Wall Street Journal and The Economist. The dataset contains headlines only, without article bodies or timestamps.

From this corpus, we drew a random sample of 1,000 headlines (489 clickbait, 511 news) for our analysis. The clickbait classification task involves identifying whether a headline employs sensationalized or misleading language designed to attract clicks, a judgment that carries inherent labeling ambiguity: headlines may use attention-grabbing phrasing without being deceptive, or may appear straightforward while omitting critical context.

The prompt for the classification task instructs the LLM to classify whether each headline is clickbait or not and to report a corresponding confidence score (0–100), based on a definition of clickbait and a standardized output format (see D for full details).

### 4.3 Empirical Analysis

We evaluate *accuracy*, measured as the Pearson correlation between *LLM classification* and *ground truth classification*, and *costs* based on unresolved tasks, as a function of the configuration dimensions outlined in

[4]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 capture the semantic context more effectively. We supplement this with a standard BERT classifier, enhanced by data augmentation techniques to address class imbalances.

Section 3. The unit of analysis is a configuration that determines the LLM classification, from which we compute *accuracy* and *unresolved tasks* (dependent variables) as a function of *redundancy* and *confidence thresholds* (independent variables), along with additional conditions used as control variables. Configurations sharing the same LLM or temperature are handled through fixed effects in the regression specifications.

*Dependent Variables.*

- **Accuracy ($a(r, d)$)** is measured as the Pearson correlation between LLM classification and ground truth classification for headlines included based on the independent variables. We use correlation rather than binary accuracy (proportion correct) because our classification tasks involve ordinal confidence signals and varying sample compositions across conditions; correlation captures the strength of linear agreement and is less sensitive to marginal distribution shifts. Both vectors entering the Pearson correlation are binary—the LLM-aggregated classification (under the unanimity or majority decision rule) and the binary expert label—on the subsample retained by the confidence filter; in cells with insufficient response variation the correlation is undefined and the configuration is dropped. For brevity, we refer to this measure as "accuracy" throughout, where higher correlation indicates more accurate classification.[5]

  For the acquisition classification task, ground truth was based on expert classification as determined by the authors of this paper. Two authors independently reviewed each headline in a sample of 1,155 headlines and agreed on the classification for 1,022. We conduct our analysis on the 1,022 headlines with expert agreement.

  For the clickbait classification task, ground truth labels are from the Mathur (2020) corpus, where classification is determined by the publishing outlet: every headline carries the fixed label—clickbait or news—of the outlet that published it. We use a sample of 1,000 headlines for our analysis. This provides a scalable, reproducible ground truth consistent with prior work using this corpus. It is possible that outlet-based labeling introduces noise, but such noise is unlikely to systematically vary with our experimental conditions.

- **Number of unresolved tasks ($n_u(r, d)$).** $n_u(r, d)$ is the number of headlines left unresolved based on the given configuration of independent and control variables.

*Independent Variables.*

- **Redundancy level ($r$)** is the number of times the LLM is asked to perform the same classification task, ranging from 1 to 25. Implementing redundancy requires an aggregation rule to resolve potentially differing classifications across repeated trials. We consider two such rules and conduct parallel analyses under each. In the unanimous specification, we retain only headlines for which the classification is identical across all trials. In the majority specification (reported in the appendix), we assign classification based on a two-thirds (67%) majority across trials. Results are robust to the choice of aggregation rule.

[5]Alternative inter-rater reliability metrics such as Cohen's kappa or Krippendorff's alpha are sensitive to marginal class distributions, which vary systematically across our experimental conditions as confidence thresholds and redundancy levels change the subset of retained observations. F1 score and binary accuracy face similar limitations in this setting. Results are qualitatively unchanged under Fisher Z-transformed specifications (see B).

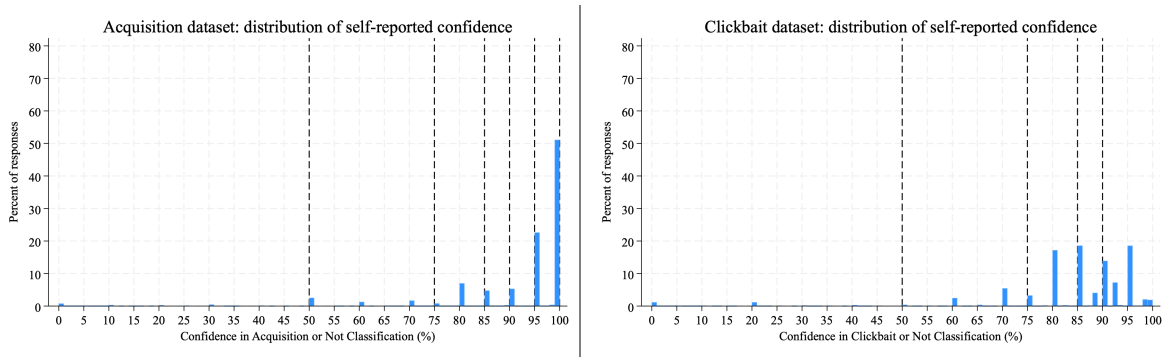


**Figure 1. Distribution of self-reported confidence across tasks.** The left panel shows the acquisition dataset and the right panel shows the clickbait dataset. Distributions are based on 652,229 classifications for acquisition and 439,512 classifications for clickbait across all LLMs, temperatures, prompting styles, and trials. Confidence is highly concentrated at the upper bound in the acquisition task, with 51% of classifications reporting exactly 100% confidence, whereas the clickbait task has only 1.7% of responses at 100% and exhibits a more dispersed distribution. Vertical dashed lines indicate the confidence thresholds used in each analysis.

- **Confidence threshold** ($d$)**.** This is the threshold of self-reported confidence required to accept an LLM classification. In the unanimous specification, a headline is retained only when the *minimum* self-reported confidence across all $r$ trials meets or exceeds $d$—that is, every individual run must clear the threshold. This ensures that the confidence filter is applied consistently across all repeated trials, analogous to the unanimity requirement for classification agreement. In the majority specification, we apply the same minimum-confidence criterion over the subset of trials corresponding to the majority classification. The LLM reports confidence as a percentage from 0 to 100.

  We select confidence thresholds based on both absolute levels and the empirical distribution, which balances the need for meaningful distinctions with sufficient observations in each cell. The distribution of self-reported confidence for each task is shown in Figure 1. The distributions differ across tasks: confidence is highly concentrated at the upper bound in the acquisition setting, whereas it is more dispersed in the clickbait setting. This difference informs the choice of thresholds used in each analysis.

  For the acquisition headlines, we use the following confidence thresholds: (1) 50%, (2) 75%, (3) 85%, (4) 90%, (5) 95%, and (6) 100%. For the clickbait headlines, we use the following thresholds: (1) 50%, (2) 75%, (3) 85%, and (4) at least 90%. We exclude the small number of observations with reported confidence below 50%, as these may reflect inconsistent or reversed coding of the classification. Results are not sensitive to this exclusion.

*Control Variables.*

We test effects of redundancy and confidence thresholds under a number of conditions to establish robustness and generalizability.

- **LLM.** A categorical variable indicating the LLM used for classification. We use five LLMs in the acquisition analysis: Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2.5-7B-Instruct (Yang et al., 2024), Qwen-3-4B (Yang et al., 2025), Gemma-3n-E4B-IT (Gemma Team et al., 2025), and GPT-4 (OpenAI et al., 2023). For the clickbait analysis, we use Llama-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, and Qwen-3-4B. For brevity, we refer to these models as Llama 3.1, Qwen 2.5, Qwen 3, Gemma 3n,

and GPT-4 throughout.[6]

- **Temperature.** The temperature parameter controls the randomness of the LLM's output distribution during text generation. Lower temperatures produce more deterministic outputs, while higher temperatures introduce greater stochastic variation. We test three temperature settings: 0.5, 1.0, and 2.0. Higher temperature is expected to increase variability across trials, which may interact with both redundancy and confidence thresholds in classification accuracy.

- **Prompting Style.** We test two prompting styles drawn from the prompt-design literature (Brown et al., 2020; Wei et al., 2022; Sahoo et al., 2024): *zero-shot prompting*, in which the prompt specifies only inputs, outputs, and class labels, with no worked examples; and *few-shot prompting*, which keeps the same structure but adds task-specific definitions, worked examples, and edge-case rules. These worked examples are illustrative; none is a headline from the classified samples. We code this as a binary indicator *Few-shot*, taking value 1 for few-shot prompting and 0 for zero-shot prompting. Full prompt text is provided in D.

*Estimation.*

We use OLS regression to assess how redundancy and confidence thresholds relate to classification accuracy and unresolved tasks. Let $j$ index a configuration. The main specifications are:

$$a_j = \alpha_a + \beta_r r_j + \beta_d d_j + \mathbf{X}_j \boldsymbol{\delta}_a + \varepsilon_{a,j} \tag{2}$$

$$n_{u,j} = \alpha_u + \gamma_r r_j + \gamma_d d_j + \mathbf{X}_j \boldsymbol{\delta}_u + \varepsilon_{u,j} \tag{3}$$

where $a_j$ is accuracy, measured as the Pearson correlation between LLM classification and ground-truth classification for configuration $j$, and $n_{u,j}$ is the number of unresolved tasks under configuration $j$. The variable $r_j$ denotes redundancy level, $d_j$ denotes the confidence threshold, and $\mathbf{X}_j$ is a vector of control variables, including indicators for LLM, temperature, and prompting style. Standard errors are heteroskedasticity-robust (HC1); we do not cluster, since each row of the regression sample is one unique configuration—defined by redundancy level ($r$), confidence threshold ($d$), LLM, temperature, and prompting style—and configurations are independent under the experimental design. As a robustness check, we also estimate all specifications using Fisher Z-transformed correlations (Fisher, 1915) as the dependent variable, which addresses the bounded distribution and heteroskedastic variance of Pearson correlations. Results are reported in B and are consistent with the main findings throughout.

## 5 Results

We present our main results on how redundancy and confidence thresholds affect classification accuracy and unresolved tasks. Figure 2 provides a visual summary of the trade-offs under the unanimity decision rule for

[6]Our primary analysis focuses on open-weight models (Llama 3.1, Qwen 2.5, Qwen 3, Gemma 3n) for reproducibility: any researcher can download the same model weights and replicate our experiments. We include GPT-4 for the acquisition task at temperature 1.0 as a commercial API benchmark to assess whether our findings generalize beyond open-weight models. For the clickbait task, Gemma 3n did not provide any variation in its classifications and could not be analyzed, so it is excluded. These design choices reflect practical constraints and do not affect the qualitative patterns of our results.

both acquisition and clickbait tasks, holding temperature and prompting style fixed for visual clarity (see caption for details). Tables 2 and 3 report the corresponding regression estimates across all experimental conditions. Results under the majority decision rule are reported in Online Appendix Section A; Fisher Z-transformed robustness checks are reported in Online Appendix Section B.

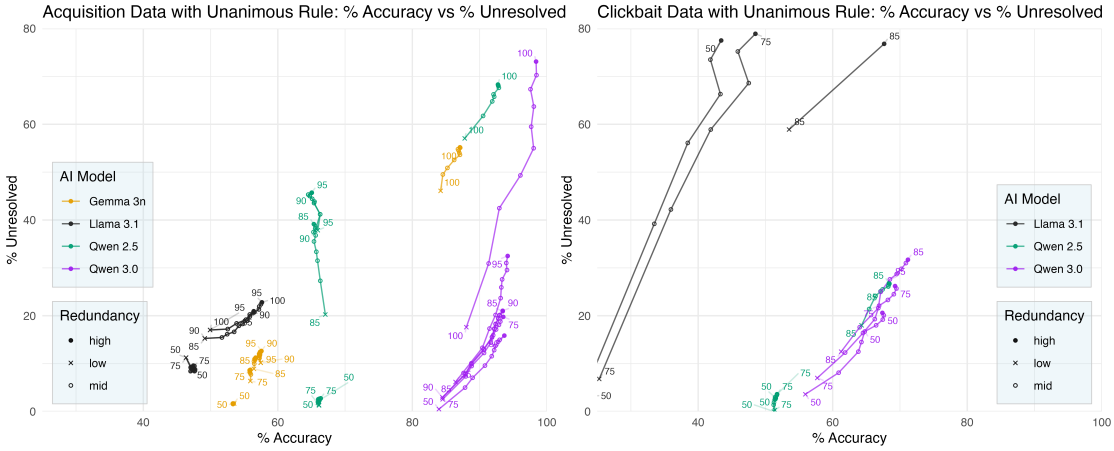


**Figure 2.** Effect of redundancy level and confidence threshold on resulting accuracy and unresolved tasks, under the unanimous agreement decision rule, including configurations with N ≥ 200, shown for temperature = 0.5 and few-shot prompting to hold control conditions fixed and facilitate visual comparison across LLMs; regression results in Tables 2 and 3 cover all temperature and prompting conditions. The left panel shows the acquisition task and the right panel shows the clickbait task. The numbers along the dots indicate the minimum confidence level required to accept an answer, and all dots with the same confidence threshold and varying redundancy are connected by line segments.

Figure 2 illustrates how redundancy and confidence thresholds affect accuracy and unresolved tasks for the acquisition (left panel) and clickbait (right panel) classification tasks. A key pattern is that increasing redundancy primarily raises the number of unresolved tasks with little corresponding improvement in accuracy, as reflected in the largely vertical alignment of points at each confidence threshold. In contrast, increasing the confidence threshold shifts outcomes toward higher accuracy, albeit at the cost of more unresolved tasks. While there is some variation across models and tasks, these patterns are broadly consistent.

Our regression results provide a statistical assessment of these effects. Table 2 presents regression results for the acquisition classification task, and Table 3 for the clickbait task, under the unanimous aggregation rule.[7] We restrict our analysis to configurations that retain a sufficient number of classification tasks to ensure reliable estimation and practical relevance: $N \geq 300$ (columns 1, 4) and $N \geq 200$ (columns 3, 6), out of a total of 1,022. We also include an estimation that further restricts the data to configurations with redundancy $\leq 10$ to assess whether results are driven by high-redundancy configurations where nonlinearities or diminishing returns may emerge (columns 2, 5). Results are not sensitive to these choices.

### 5.1 Effect of Redundancy

For the acquisition classification task, the estimated effect representing increase in accuracy is effectively zero (Table 2, columns 1–3). In contrast, redundancy is clearly associated with an increase in unresolved tasks (Table 2, columns 4–6). For example, from the estimation where N ≥ 300 (column 4), redundancy increase of 10 leads to 47 additional unresolved tasks, or approximately 15% of the mean (322), with no increase in

[7]The corresponding majority-rule results are in A, Tables A1 and A2.

**Table 2.** Regression results for the acquisition task under unanimous agreement.

| | *Dep. Var.: Correlation* | | | *Dep. Var.: Number Unresolved* | | |
|---|---|---|---|---|---|---|
| | (1) N ≥300 | (2) N≥300; redundancy ≤10 | (3) N≥200 | (4) N ≥300 | (5) N≥300; redundancy ≤10 | (6) N≥200 |
| Confidence = 50% | −0.156*** | −0.148*** | −0.155*** | −433.314*** | −429.893*** | −484.441*** |
| | (0.011) | (0.014) | (0.010) | (11.736) | (16.450) | (10.706) |
| Confidence = 75% | −0.172*** | −0.163*** | −0.171*** | −351.976*** | −352.938*** | −397.606*** |
| | (0.011) | (0.014) | (0.010) | (11.704) | (16.219) | (10.588) |
| Confidence = 85% | −0.148*** | −0.136*** | −0.146*** | −243.388*** | −247.594*** | −281.483*** |
| | (0.011) | (0.014) | (0.010) | (11.504) | (15.931) | (10.293) |
| Confidence = 90% | −0.144*** | −0.128*** | −0.138*** | −230.549*** | −229.637*** | −253.956*** |
| | (0.011) | (0.014) | (0.010) | (11.697) | (16.317) | (10.620) |
| Confidence = 95% | −0.096*** | −0.094*** | −0.098*** | −158.248*** | −165.106*** | −176.628*** |
| | (0.011) | (0.014) | (0.010) | (11.890) | (16.394) | (10.769) |
| Redundancy | 0.000 | 0.001 | 0.000 | 4.695*** | 9.831*** | 5.097*** |
| | (0.000) | (0.001) | (0.000) | (0.314) | (1.090) | (0.322) |
| GPT-4 | 0.253*** | 0.217*** | 0.250*** | −121.589*** | −133.188*** | −129.349*** |
| | (0.007) | (0.010) | (0.007) | (9.287) | (14.647) | (9.746) |
| Llama 3.1 | 0.041*** | 0.032*** | 0.050*** | 114.073*** | 86.202*** | 135.100*** |
| | (0.007) | (0.009) | (0.007) | (6.690) | (9.320) | (6.826) |
| Qwen 2.5 | 0.122*** | 0.138*** | 0.127*** | 139.517*** | 126.921*** | 172.380*** |
| | (0.005) | (0.007) | (0.006) | (7.018) | (10.273) | (6.692) |
| Qwen 3 | 0.368*** | 0.361*** | 0.372*** | 28.691*** | −32.688*** | 34.239*** |
| | (0.004) | (0.006) | (0.004) | (5.034) | (7.857) | (5.399) |
| Temperature = 1.0 | −0.032*** | −0.016*** | −0.025*** | 68.633*** | 54.534*** | 76.460*** |
| | (0.004) | (0.006) | (0.004) | (5.088) | (7.532) | (5.256) |
| Temperature = 2.0 | −0.074*** | −0.059*** | −0.084*** | 150.559*** | 142.222*** | 169.371*** |
| | (0.005) | (0.007) | (0.006) | (6.064) | (8.629) | (6.192) |
| Few-shot | 0.283*** | 0.259*** | 0.278*** | −118.014*** | −122.768*** | −134.452*** |
| | (0.004) | (0.005) | (0.004) | (4.273) | (6.393) | (4.469) |
| Constant | 0.410*** | 0.405*** | 0.407*** | 494.429*** | 505.623*** | 527.194*** |
| | (0.012) | (0.016) | (0.011) | (12.469) | (17.044) | (11.110) |
| Observations | 3,080 | 1,447 | 3,315 | 3,080 | 1,447 | 3,315 |
| $R^2$ | 0.804 | 0.812 | 0.779 | 0.590 | 0.598 | 0.641 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

accuracy (column 1). For the clickbait task, the effect of redundancy on accuracy is more precisely estimated but remains modest in magnitude. In Table 3 column 1, a redundancy increase of 10 is associated with a 0.03 increase in correlation, corresponding to a 4.6% improvement relative to the mean (0.65), alongside an increase of 60 unresolved tasks (column 4), or approximately 20% of the mean (302).

Prior literature and practice on use of human labelers has presented overwhelming justification for redundancy in classification work (Sheng et al., 2008; Snow et al., 2008). Not surprisingly, researchers have also posited the plausibility of the same for LLMs either with repeated runs of an LLM with varied temperature setting (Wang et al., 2023) or a mixture of models (Wang et al., 2024). In contrast to these, we find no evidence that redundancy improves classification accuracy; and worse, it increases cost.

**Finding 1.** Increasing redundancy in classification tasks yields at most modest changes in accuracy, while consistently increasing the number of unresolved tasks.

**Table 3.** Regression results for the clickbait task under unanimous agreement.

| | *Dep. Var.: Correlation* | | | *Dep. Var.: Number Unresolved* | | |
|---|---|---|---|---|---|---|
| | (1) N ≥300 | (2) N≥300; redundancy ≤10 | (3) N≥200 | (4) N ≥300 | (5) N≥300; redundancy ≤10 | (6) N≥200 |
| Confidence = 50% | −0.183*** | −0.182*** | −0.177*** | −417.322*** | −410.935*** | −441.126*** |
| | (0.009) | (0.012) | (0.008) | (12.999) | (18.390) | (12.451) |
| Confidence = 75% | −0.147*** | −0.141*** | −0.134*** | −330.544*** | −309.847*** | −328.224*** |
| | (0.009) | (0.013) | (0.008) | (14.014) | (19.678) | (13.470) |
| Confidence = 85% | −0.085*** | −0.078*** | −0.074*** | −196.693*** | −170.858*** | −172.197*** |
| | (0.008) | (0.012) | (0.008) | (14.685) | (21.296) | (14.426) |
| Redundancy | 0.003*** | 0.008*** | 0.003*** | 6.041*** | 21.703*** | 8.472*** |
| | (0.000) | (0.001) | (0.000) | (0.461) | (1.956) | (0.549) |
| Qwen 2.5 | 0.295*** | 0.289*** | 0.238*** | −344.425*** | −369.494*** | −430.685*** |
| | (0.021) | (0.021) | (0.020) | (23.817) | (22.785) | (24.145) |
| Qwen 3 | 0.371*** | 0.344*** | 0.321*** | −146.188*** | −210.180*** | −205.537*** |
| | (0.021) | (0.021) | (0.020) | (24.782) | (23.518) | (24.477) |
| Temperature = 1.0 | 0.013*** | 0.001 | 0.012** | 50.672*** | 29.868** | 41.411*** |
| | (0.005) | (0.008) | (0.005) | (8.478) | (13.711) | (9.099) |
| Temperature = 2.0 | 0.033*** | 0.028*** | 0.029*** | 110.488*** | 87.387*** | 126.859*** |
| | (0.005) | (0.008) | (0.005) | (8.973) | (14.185) | (10.198) |
| Few-shot | −0.027*** | −0.023*** | −0.035*** | −69.592*** | −76.372*** | −94.955*** |
| | (0.004) | (0.006) | (0.004) | (7.928) | (11.349) | (7.750) |
| Constant | 0.431*** | 0.419*** | 0.478*** | 746.044*** | 712.788*** | 823.924*** |
| | (0.023) | (0.026) | (0.022) | (25.990) | (27.412) | (26.005) |
| Observations | 973 | 463 | 1,101 | 973 | 463 | 1,101 |
| $R^2$ | 0.773 | 0.798 | 0.744 | 0.702 | 0.672 | 0.708 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

### 5.2 Effect of Self-Reported Confidence

With the dismal performance of one plausible lever for improving accuracy—redundancy—and growing interest in the use of self-reported confidence in LLM answers (Kadavath et al., 2022) albeit with concerns of overconfidence (Xiong et al., 2024), we investigate the accuracy gain and cost loss of segmenting results on self-reported confidence scores. We find larger and more consistent increases in accuracy when conditioning on self-reported confidence, accompanied by substantial increases in unresolved tasks. For the acquisition task, the omitted category corresponds to classifications made with 100% confidence. From Table 2 column 1, compared to a confidence threshold that simply excludes inconsistent coding (all classifications above 50%), classifications with 100% confidence are more accurate, with a 0.156 increase in correlation, on average, which corresponds to a 29% increase relative to the mean correlation (0.54). Even moving from 95% confidence to 100% yields higher accuracy, with a 0.096 increase in correlation, an 18% increase relative to the mean.

These improvements come at a cost in terms of unresolved tasks. Relative to 100% confidence, 95% confidence is associated with approximately 158 fewer unresolved tasks (49% of the mean), while 50% confidence is associated with approximately 433 fewer unresolved tasks (134% of the mean), indicating that higher confidence thresholds substantially reduce coverage (Table 2 column 4). Employing a majority classification rule yields similar effects in magnitude and significance (see Appendix Table A1).

For the clickbait task, the same pattern holds, although the available thresholds differ due to the underlying distribution of confidence (Figure 1). Relative to the highest threshold (90%), classifications at 50% confidence

are less accurate by 0.183 correlation (28% below the mean of 0.65), while classifications at 85% confidence are less accurate by 0.085 correlation (13% below the mean, Table 3 column 1).

These accuracy gains come at the cost of unresolved tasks: relative to the highest threshold of 90% confidence, 85% confidence is associated with approximately 197 fewer unresolved tasks (65% of the mean of 302), while 50% confidence is associated with approximately 417 fewer unresolved tasks (138% of the mean, Table 3 column 4). Figure 3 provides a visual illustration of this relationship across LLMs, showing particularly sharp gains at the highest thresholds.

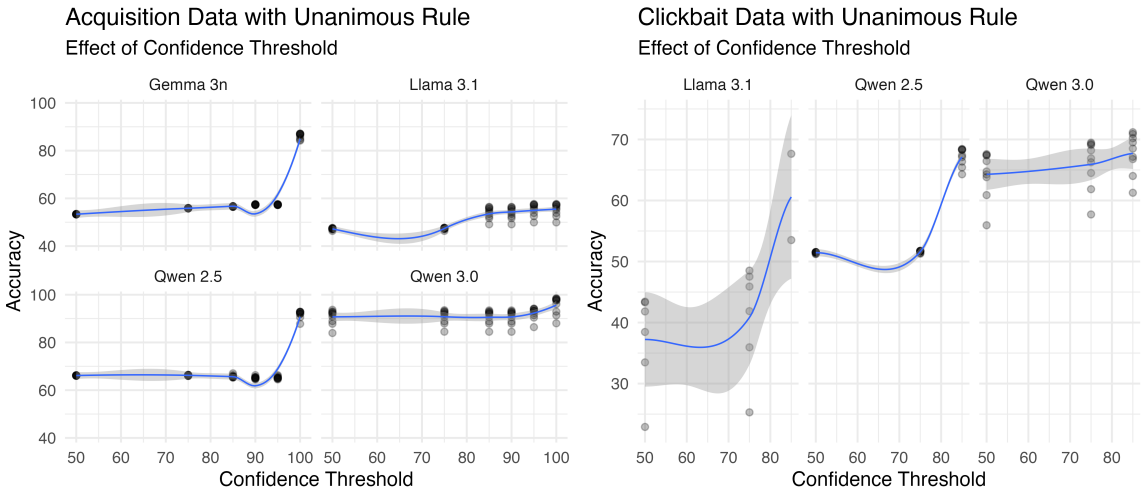


**Figure 3.** Accuracy by confidence threshold for each LLM, under the unanimous agreement decision rule, including configurations with N ≥ 200, shown for temperature = 0.5 and few-shot prompting to hold control conditions fixed and facilitate visual comparison across LLMs. The left panel shows the acquisition task (Gemma 3n, Llama 3.1, Qwen 2.5, Qwen 3), and the right panel shows the clickbait task (Llama 3.1, Qwen 2.5, Qwen 3). Smoothed trend lines with 95% confidence intervals illustrate that accuracy generally increases with the self-reported confidence threshold, with especially sharp gains at the highest thresholds.

Taken together, the results indicate that conditioning on self-reported confidence reliably improves classification accuracy, with a clear trade-off between accuracy and unresolved tasks. The magnitude and precision of these effects are substantially larger than those observed for redundancy. Unlike human worker classification, running redundant trials without conditioning on confidence does not justify the additional cost. A contour plot of the accuracy surface over redundancy and confidence (Appendix Figure A1) shows how accuracy variation is driven primarily by the confidence threshold, not redundancy; the companion surface for unresolved tasks (Appendix Figure A2) shows that both levers raise the unresolved-task share.

**Finding 2.** Conditioning on self-reported confidence yields improvements in classification accuracy, but at the cost of reduced coverage through increased unresolved tasks.

**Effects of Control Variables**: Both accuracy and the cost of unresolved tasks also vary with additional configuration factors included as controls in our analyses. Few-shot prompting shifts the efficiency frontier on the acquisition task, increasing accuracy correlation by 0.283 (SE = 0.004, $p < 0.01$) while reducing unresolved tasks. On the clickbait task it does not shift the frontier: accuracy correlation falls by 0.027 (SE = 0.004, $p < 0.001$) while unresolved tasks also fall by roughly 70 (SE = 7.9, $p < 0.001$)—a movement along the frontier rather than a shift. Temperature has mild, task-dependent effects and does not substantially shift

the frontier. Among LLMs, Qwen 3 achieves the highest accuracy on both tasks. Better LLM choice shifts the frontier in the clickbait task but has more mixed effects in the acquisition task.

**Per-LLM split-sample regressions**: Having established that confidence thresholds are the primary accuracy lever, we return to redundancy—which yields few accuracy gains—to examine the within-LLM error correlation that our decision framework in Section 3 identifies as the likely mechanism. The framework's key prediction, from Equation 1, is that the redundancy coefficient should be more positive for LLMs with weaker within-LLM error correlation across repeated runs, and closer to zero (or negative) for LLMs with strong within-LLM correlation. As motivated in Section 3.1, the Qwen 2.5 vs. Qwen 3 contrast—an instruction-following and a reasoning model from the same family—supplies the required prior about correlation structure, with the reasoning model expected to show the more positive redundancy coefficient. We re-estimate the main accuracy specification separately for each LLM.[8]

Tables 4 and 5 report per-LLM split-sample regressions for the acquisition and clickbait tasks respectively. We include all LLMs for completeness but focus on the Qwen 2.5 vs. Qwen 3 comparison as our directional test.[9] For the acquisition task (Table 4), the Qwen 2.5 vs. Qwen 3 comparison confirms the directional prediction: Qwen 3 shows a positive redundancy coefficient ($r = 0.004$, $p < 0.01$) while Qwen 2.5 is significantly negative ($r = -0.009$, $p < 0.01$). For the clickbait task (Table 5), the same directional pattern holds: Qwen 3 shows a larger positive coefficient ($r = 0.013$, $p < 0.01$) than Qwen 2.5 ($r = 0.005$, $p < 0.01$). Results for the remaining LLMs are reported in the tables for completeness.

The Qwen 2.5 sign-flip across tasks—negative in acquisition but positive in clickbait—likely reflects the different knowledge demands of each task. Acquisition classification requires specialized knowledge about corporate transactions that Qwen 2.5 may lack, causing it to make the same mistakes repeatedly across runs. Clickbait classification, by contrast, relies on everyday linguistic cues that the model handles more variably, so errors are less systematic and redundancy provides modest benefit.

Taken together, the Qwen 2.5 vs. Qwen 3 comparison provides directional evidence that within-LLM correlation governs redundancy gains, consistent with our framework's prediction, and contributes a mechanistic explanation for the near-zero average effect documented in Finding 1.

**Across-LLM Redundancy**: The above results show limited gains from within-LLM redundancy. A natural question is whether combining predictions from different LLMs—or across-LLM redundancy—yields more substantial improvements. Different LLMs may provide more diverse assessments, potentially reducing correlated errors and making redundancy more effective at improving accuracy compared to within-LLM redundancy. We therefore examine classification accuracy under across-LLM redundancy. To isolate across-LLM effects, we use a single classification from each model per headline, rather than multiple runs per model. We report accuracy for pairwise and three-way LLM ensembles under both majority and unanimity

[8]We restrict to redundancy $\leq 10$ because the split-sample regressions reflect the range over which most observed gains accrue. Acquisition ground truth is balanced by construction; the clickbait corpus is also approximately balanced (489 clickbait / 511 news). Pearson correlation and binary accuracy therefore co-move on both tasks, supporting the directional reading.

[9]The Qwen 2.5 vs. Qwen 3 pair holds the model family fixed but is not a clean single-variable manipulation: Qwen 2.5 is a 7B-parameter model and Qwen 3 a 4B-parameter model, and the two were trained on different-vintage corpora. The parameter-count gap, however, runs against a scale-based account of the contrast—the *smaller* model (Qwen 3) shows the *more* positive redundancy coefficient—so we read the reasoning capability, not scale, as the most plausible driver, while treating the comparison as directional rather than dispositive.

**Table 4.** Per-LLM regression results for the acquisition task, redundancy ≤ 10 and N ≥ 300, under unanimous agreement.

| | (1) GPT-4 | (2) Gemma 3n | (3) Llama 3.1 | (4) Qwen 2.5 | (5) Qwen 3 |
|---|---|---|---|---|---|
| Confidence = 50% | −0.132*** | −0.180*** | −0.190*** | −0.148*** | −0.041*** |
| | (0.008) | (0.023) | (0.011) | (0.019) | (0.009) |
| Confidence = 75% | −0.122*** | −0.162*** | −0.168*** | −0.262*** | −0.038*** |
| | (0.008) | (0.023) | (0.010) | (0.018) | (0.008) |
| Confidence = 85% | −0.102*** | −0.177*** | −0.062*** | −0.247*** | −0.035*** |
| | (0.007) | (0.022) | (0.009) | (0.017) | (0.008) |
| Confidence = 90% | −0.095*** | −0.167*** | −0.060*** | −0.254*** | −0.034*** |
| | (0.007) | (0.022) | (0.009) | (0.019) | (0.008) |
| Confidence = 95% | −0.037*** | −0.164*** | −0.006 | −0.229*** | 0.022** |
| | (0.008) | (0.022) | (0.011) | (0.018) | (0.009) |
| Redundancy | 0.006*** | −0.001 | 0.004*** | −0.009*** | 0.004*** |
| | (0.001) | (0.001) | (0.001) | (0.002) | (0.001) |
| Temperature = 1.0 | | 0.005 | −0.017*** | −0.034*** | −0.004 |
| | | (0.007) | (0.005) | (0.010) | (0.003) |
| Temperature = 2.0 | | −0.006 | −0.157*** | −0.042*** | −0.012** |
| | | (0.006) | (0.008) | (0.012) | (0.005) |
| Few-shot | 0.097*** | 0.402*** | 0.061*** | 0.298*** | 0.341*** |
| | (0.004) | (0.006) | (0.005) | (0.012) | (0.004) |
| Constant | 0.630*** | 0.348*** | 0.527*** | 0.641*** | 0.588*** |
| | (0.008) | (0.024) | (0.010) | (0.019) | (0.008) |
| Observations | 120 | 337 | 300 | 252 | 316 |
| $R^2$ | 0.930 | 0.949 | 0.867 | 0.852 | 0.971 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

**Table 5.** Per-LLM regression results for the clickbait task, redundancy ≤ 10 and N ≥ 300, under unanimous agreement.

| | (1) Llama 3.1 | (2) Qwen 2.5 | (3) Qwen 3 |
|---|---|---|---|
| Confidence = 50% | −0.159 | −0.285*** | −0.124*** |
| | (0.101) | (0.007) | (0.007) |
| Confidence = 75% | −0.115 | −0.278*** | −0.049*** |
| | (0.102) | (0.007) | (0.006) |
| Confidence = 85% | −0.043 | −0.186*** | −0.023*** |
| | (0.127) | (0.008) | (0.006) |
| Redundancy | −0.004 | 0.005*** | 0.013*** |
| | (0.011) | (0.001) | (0.001) |
| Temperature = 1.0 | −0.029 | 0.015*** | 0.005 |
| | (0.047) | (0.004) | (0.006) |
| Temperature = 2.0 | 0.125** | 0.033*** | 0.004 |
| | (0.058) | (0.005) | (0.006) |
| Few-shot | 0.053 | −0.057*** | −0.031*** |
| | (0.040) | (0.004) | (0.005) |
| Constant | 0.371*** | 0.843*** | 0.691*** |
| | (0.106) | (0.009) | (0.007) |
| Observations | 54 | 202 | 207 |
| $R^2$ | 0.379 | 0.912 | 0.736 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

decision rules, at different confidence thresholds, and compare to individual LLM accuracy (Tables 6 and 7).[10]

For the acquisition task, we examine both pairwise and three-way aggregation under unanimity and majority decision rules. Most approaches do not outperform the best individual LLM. One pairwise ensemble (Qwen 2.5 + Gemma 3n, conf $\geq$ 50%) modestly exceeds the best individual LLM (0.681 vs. 0.661 for Qwen 2.5), but at the cost of reducing resolved tasks from 1,009 to 831. Voting across all three LLMs, using either unanimity or majority aggregation, does not outperform the best individual model at any confidence threshold.

For the clickbait task, ensemble classification yields only small improvements over the best individual LLM at low confidence thresholds (0.525 vs. 0.513 for Qwen 2.5), while substantially reducing resolved tasks. At the highest threshold (85%), the ensemble achieves substantially higher accuracy than any individual LLM (0.813 vs. 0.643 for Qwen 2.5), but on only 290 of 1,000 observations. Since the pattern does not replicate in the acquisition analysis, this is likely task-specific rather than a general property of across-LLM redundancy.

Taken together, results provide limited support for across-LLM redundancy improving classification accuracy. The acquisition findings show little benefit across configurations, while the clickbait findings are suggestive but not robust. The limited gains across both tasks likely reflect correlated errors across models due to shared deep-learning architectures and overlapping training data—though the clickbait finding at the highest confidence threshold suggests this correlation may break down under sufficiently strict filtering.

This pattern of correlated errors across LLMs connects to a broader debate in the ensemble learning literature about whether LLM diversity is sufficient for meaningful aggregation gains (Schoenegger et al., 2024), and suggests that architectural similarity may be a fundamental constraint on across-LLM redundancy benefits in the current generation of models. Whether this pattern generalizes to other tasks and model combinations remains an open question, but in conjunction with the within-LLM findings, these results suggest that redundancy—whether within or across LLMs—is not a reliable lever for improving classification accuracy in interpretive tasks, at least for the current generation of models.

**Table 6.** Pairwise and three-way ensemble correlations vs. individual LLM correlations on the acquisition task ($N$ = 1,022); resolved-task counts in parentheses.

| | *Individual LLMs* | | | *Pairwise Ensembles* | | | *Voting Methods* | |
|---|---|---|---|---|---|---|---|---|
| | Gemma 3n | Llama 3.1 | Qwen 2.5 | Qwen 2.5 + Gemma 3n | Qwen 2.5 + Llama 3.1 | Llama 3.1 + Gemma 3n | Majority | Unanimous |
| Conf. $\geq$ 50% | 0.535 | 0.464 | 0.661 | 0.681 | 0.656 | 0.508 | 0.564 | 0.646 |
| | (1,005) | (907) | (1,009) | (831) | (693) | (813) | (1,017) | (667) |
| Conf. $\geq$ 85% | 0.565 | 0.491 | 0.671 | 0.668 | 0.646 | 0.545 | 0.609 | 0.635 |
| | (931) | (866) | (815) | (704) | (604) | (752) | (779) | (581) |
| Conf. $\geq$ 95% | 0.575 | 0.499 | 0.659 | 0.653 | 0.632 | 0.560 | 0.656 | 0.626 |
| | (918) | (848) | (635) | (618) | (525) | (730) | (624) | (516) |

Number of classification tasks meeting the confidence threshold in parentheses.

**Financial and Operational Costs**: Our model and empirical analysis focus on the trade-off between accuracy and the cost of unresolved tasks. Because practitioners must also weigh direct financial costs when selecting a configuration, we outline the key cost implications of LLM choice and redundancy level.

Open-weight models (Llama 3.1, Qwen 2.5, Qwen 3, and Gemma 3n) require GPU infrastructure but do not incur per-query fees, whereas commercial API models such as GPT-4 are priced per token.

[10]We exclude Qwen 3 from the across-LLM analysis because it is a reasoning model, whereas the other models are standard instruction-following models. We do not find across-LLM redundancy gains when combining Qwen 3 with the other LLMs.

**Table 7.** Pairwise ensemble correlations vs. individual LLM correlations on the clickbait task ($N$ = 1,000); resolved-task counts in parentheses.

| | Llama 3.1 | Qwen 2.5 | Ensemble |
|---|---|---|---|
| Conf. ≥ 50% | 0.229 | 0.513 | 0.525 |
| | (967) | (1,000) | (611) |
| Conf. ≥ 75% | 0.253 | 0.514 | 0.537 |
| | (932) | (997) | (601) |
| Conf. ≥ 85% | 0.535 | 0.643 | 0.813 |
| | (411) | (820) | (290) |

Number of classification tasks meeting the confidence threshold in parentheses.

Redundancy also imposes direct computational and financial costs that compound with $r$. For open-weight models, wall-clock time scales linearly with $r$ on a fixed GPU; for commercial API models, per-query fees accumulate proportionally. Given our main finding that redundancy delivers at most modest accuracy gains over a single run, the incremental cost of additional runs—which at the scale of typical research corpora runs to tens or hundreds of thousands of additional queries—is rarely justified.

## 6 Guidelines

Our results yield four actionable guidelines for researchers and practitioners at the end stage of an operational workflow, using LLMs to great structured data from unstructured text; one for each configuration dimension in our decision framework. Casting empirical results as deployment guidance follows an established practice-guidance genre in operations management (De-Arteaga et al., 2022).

The main takeaway, that confidence thresholds, not redundancy, are the primary accuracy lever, inverts a long-standing principle from the human-labeling playbook. The unifying logic is that confidence and redundancy trade accuracy against unresolved tasks, while LLM choice and few-shot prompting can shift the efficiency frontier outward in some settings—raising accuracy and reducing unresolved tasks simultaneously—though the extent depends on the task.

### Self-Reported Confidence

*Use self-reported confidence as the starting point for automated acceptance. Choose a model- and task-specific threshold on an expert-labeled pilot, and route cases below that threshold for further review.* Our experiments show that confidence is the primary accuracy lever. On the acquisition task, raising the threshold from $d$=50% to $d$=100% improves accuracy by 29% over the mean, and even the last step from 95% to 100% yields an 18% gain (Table 2). The clickbait task shows the same monotone pattern at its task-appropriate thresholds (Table 3). The lever is also straightforward to operationalize, and can be done in two steps. First, inspect the empirical distribution of self-reported confidence on a small sample before fixing a threshold. As our two tasks illustrate, confidence distributions vary substantially across classification contexts (Figure 1). Setting the threshold against the empirical distribution—considering both absolute levels and the distribution of confidence scores—rather than picking a fixed level a priori, targets the subset the practitioner is willing to commit to LLM classification. Second, use the threshold to partition the corpus: accept the LLM classification for high-confidence cases and route low-confidence ones for human review—the division of labor that targeted-automation and gatekeeper designs formalize in operations settings (Imdahl et al., 2026; Dai and

Singh, 2026). Because accuracy rises monotonically with the threshold, this hybrid pipeline delivers higher overall accuracy at lower cost than either an LLM-only or human-only approach.

**Redundancy with Multiple Trials**

*Do not adopt repeated classification as the default quality-control response. Start with a single-run baseline and add repetition only when its incremental gains justify the additional model calls and unresolved work.* Redundancy alone does not meaningfully improve LLM classification accuracy, even though it consistently increases costs. This is in sharp contrast to the human-labeling playbook, where redundancy is the cornerstone of quality control. The Condorcet-style logic that justifies redundancy for human labelers requires independent and identically distributed responses, an assumption that does not hold for LLMs.

Per-LLM split-sample regressions in Section 5 (Tables 4 and 5) provide evidence that the gap stems from positive within-LLM correlation: when an LLM misclassifies a task, repeated runs tend to produce the same error, so systematic mistakes persist rather than average out. Any gains appear to plateau at low redundancy levels while the cost of unresolved tasks continues to grow with $r$.

Across-LLM redundancy is similarly limited, likely because the open-weight models we test share deep-learning architectures and overlapping training data, leading to correlated errors across models. Combining predictions across LLMs does not reliably outperform the best individual LLM and adds setup, prompt-calibration, and unresolved-task costs. The one exception—a single clickbait ensemble at the strictest threshold, resolving fewer than a third of tasks—we treat as task-specific (Section 5).

**Prompting Style**

*For domain-specific labeling, start with explicit definitions, worked examples, and edge-case guidance. Evaluate their effect on both classification quality and unresolved work.* For the acquisition task, few-shot prompting raises accuracy *and* reduces unresolved tasks (Tables 2 and A1)—a frontier shift, not a movement along the frontier. For the clickbait task the effect on accuracy is near zero, plausibly because the classification task (sensational vs. informative phrasing) is intuitive and widely encountered in training data, so the LLM can apply it without worked examples (Tables 3 and A2). The contrast supports a clean rule: treat few-shot prompts as the default for any classification task that requires domain knowledge or expert disambiguation, and zero-shot prompts only when the labeling cue is plainly self-evident from the input.

**LLM Choice**

*Choose a high-performing model that meets the workflow's execution and reproducibility requirements. For reproducibility-sensitive applications, begin with open-weight models and use commercial models as benchmarks.* LLM choice affects both accuracy and unresolved tasks, though effects vary by task (Tables 2 and 3). Among the five LLMs we tested, Qwen 3—an open-weight model—achieved the highest accuracy on both the acquisition and clickbait tasks. GPT-4 performed competitively on the acquisition task but is available only through a commercial API subject to per-query cost and version drift, both of which compromise reproducibility. We therefore recommend treating open-weight models as the default and commercial API models as benchmarks: in our experiments the open-weight set already contained the accuracy frontier, and the reproducibility risk of commercial APIs is not justified by any accuracy delta we observed for routine use.

## 7 Conclusion

Decoding unstructured text with LLMs is now standard practice in organizations and in applied research, but it has proceeded largely without the empirical guidelines that govern human labeling. This paper asks the workflow designer's question directly: given imperfect LLM classification, which cases should the LLM decide, which should be routed to human review, and at what cost? The two configuration levers we study parameterize exactly this routing decision. We answer by developing a decision framework in which practitioners trade accuracy of resolved tasks against the cost of unresolved tasks and the computational cost of repeated LLM runs, and we test it over redundancy ($r$) and self-reported confidence threshold ($d$) across two applied classification tasks (corporate-acquisition and clickbait headlines) and five LLMs. We establish an empirical basis for confidence-first quality control and translate it into a decision framework for configuring LLM-based labeling workflows.

Two findings stand out, in order of practical importance. *First, self-reported confidence thresholds are the primary accuracy lever.* On the acquisition task, raising the threshold from $d$=50% to $d$=100% improves accuracy by 29% over the mean; raising from 95% to 100% alone yields an 18% gain. *Second, redundancy alone does not transfer the usual advantages familiar from the human-labeling playbook.* We find no meaningful accuracy improvements from within-LLM redundancy in our primary analysis, nor from across-LLM redundancy in our supplementary analysis. Our framework attributes these null results to the failure of the i.i.d. assumption that underwrites the Condorcet-Jury logic behind human-labeling redundancy, and our per-LLM comparison provides directional evidence for this prediction. When repeated runs are positively correlated within an LLM, misclassifications tend to persist rather than average out across runs. Correlated errors may also explain the limited benefits of across-LLM redundancy, as the open-weight models we test share deep-learning architectures and overlapping training data, so error correlations may carry across models. We read the evidence of correlated errors within and across LLMs as a methodological finding in its own right, and it points to open questions about whether targeted interventions—in LLM tuning, configuration control, or prompt design—could reduce within-LLM error correlation and restore the benefits of redundancy.

Our framework makes the cost side of the accuracy-cost trade-off explicit. For commercial API models, scaling redundancy imposes per-query costs that compound with $r$ while delivering near-zero accuracy gains; for open-weight models, wall-clock time scales linearly with $r$ on a fixed GPU. Since gains, if any, appear to plateau at low redundancy levels, additional runs are rarely worth the cost. Beyond redundancy and confidence, LLM choice and few-shot prompting also affect accuracy and unresolved tasks, though effects vary by task. In the acquisition task, few-shot prompting raises accuracy while also reducing unresolved tasks—a frontier shift rather than a movement along it. Temperature, by contrast, has at most a mild and task-dependent effect. Together, our decision framework and empirical results let organizations and researchers balance the three quantities that matter: accuracy on resolved tasks, computational cost, and the downstream cost of unresolved tasks.

Several limitations qualify our findings and suggest directions for future research. Both tasks are binary classification of short headlines; whether confidence thresholds remain the dominant accuracy lever in longer documents, multi-class settings, or tasks with qualitatively different ambiguity structures is an open question for future work. The clickbait ground truth is outlet-based rather than content-annotated—headlines

from outlets such as BuzzFeed are labeled clickbait regardless of individual content—which introduces measurement noise but is unlikely to correlate systematically with our experimental dimensions. Our main accuracy metric is Pearson correlation with ground truth. We chose this metric over other options for distributional reasons discussed in Section 4.3, and results are robust to Fisher Z-transformed specifications (B). Correlation does not distinguish error types, however, and other measures such as precision, recall, and F1 may matter differently in applications with asymmetric misclassification costs. Finally, we use verbalized confidence because it is accessible to applied researchers without model-internals access; probabilistic confidence may yield tighter thresholds where internals are available.

Our evidence for the correlation mechanism—the per-LLM split-sample regressions of Section 5—has two limitations that make it directional rather than conclusive. First, the framework of Section 3.1 is stated in binary accuracy while the regressions use Pearson correlation, so the analysis tests its directional predictions rather than matching the i.i.d. equations numerically; in our balanced samples the two metrics co-move, supporting the directional reading. Second, the test rests on one theoretically motivated within-family contrast: Qwen 3, the reasoning model, shows the more positive redundancy coefficient in both tasks, as predicted, but a single model pair is limited in scope. A more purposeful, controlled study—recomputed on a binary-accuracy basis and spanning a wider set of model contrasts—could turn this suggestive evidence into a conclusive test.

Our estimates describe the specific models we test, at one point in a fast-moving technology; advances in prompting, confidence calibration, fine-tuning, or access to token-level probabilities could shift the coefficients we report. What is built to travel is the decision framework: the accuracy-coverage-cost logic and the partition-and-route design carry over unchanged as the model set evolves, with the empirical surfaces re-estimated on a pilot sample.

The practical lesson is to organize LLM-based labeling around confidence rather than repeated classification by default. This runs against the routine use of redundancy in human labeling: confidence thresholds provide a useful lever for determining which classifications to accept, whereas additional runs should be used only when pilot evidence shows that their incremental gains justify their computational and unresolved-task costs. Model choice and task guidance provide additional levers that affect where this accuracy-cost frontier lies. Together, these choices determine both the quality of automated classifications and the workload left for further review, and our framework provides a way to configure them around the requirements of the application.

## References


**Abrahams, Alan S, Weiguo Fan, G Alan Wang, Zhongju Zhang, and Jian Jiao**, "An Integrated Text Analytic Framework for Product Defect Discovery," *Production and Operations Management*, 2015, *24* (6), 975–990.

**Agley, Jon, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo**, "Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7," *Behavior Research Methods*, 2022, *54* (2), 885–897.

**Ahsen, Mehmet E., Radha Mookerjee, and Mehmet U. S. Ayvaci**, "Will machines take over? Algorithms for human–machine collaborative decision making in healthcare," *Production and Operations Management*, 2026. Advance online publication.

**Alyakoob, Mohammed and Mohammad Rahman**, "Market design choices, racial discrimination, and equitable microentrepreneurship in digital marketplaces," *Management Science*, 2026, *72* (3), 1878–1903.

**Bodoh-Creed, Aaron, Jörn Boehnke, and Brent Hickman**, "Market frictions versus seller signaling: Explaining deviations from the law of one price," *Quantitative Marketing and Economics*, 2026. Forthcoming.

**Boehnke, Jörn, Pantelis Loupos, and Ying Gu**, "Social drug dealing: How peer-to-peer fintech platforms have transformed illicit drug markets," *Annals of Operations Research*, 2024, *335* (2), 645–663.

**Boland, Philip J**, "Majority systems and the Condorcet jury theorem," *Journal of the Royal Statistical Society Series D: The Statistician*, 1989, *38* (3), 181–189.

**Brown, Tom B., Benjamin Mann, Nick Ryder et al.**, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, 2020, *33*, 1877–1901.

**Cao, Lang**, "Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism," in "Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing" 2024, pp. 3628–3646.

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

**Chen, Hsinchun, Roger H. L. Chiang, and Veda C. Storey**, "Business intelligence and analytics: From big data to big impact," *MIS Quarterly*, 2012, *36* (4), 1165–1188.

**Chen, Zhikai, Haitao Mao, Hongzhi Wen et al.**, "Label-free node classification on graphs with large language models (LLMs)," in "International Conference on Learning Representations (ICLR)" 2024.

**Dai, Tinglong and Jayashankar M Swaminathan**, "Artificial Intelligence and Operations: A Foundational Framework of Emerging Research and Practice," *Production and Operations Management*, 2026, *35* (9), 3255–3271.

__ **and Simrita Singh**, "Using Artificial Intelligence as Gatekeeper or Second Opinion: Designing Patient Pathways for Artificial Intelligence Augmented Healthcare," *Production and Operations Management*, 2026, pp. 1–21. Advance online publication.

**De-Arteaga, Maria, Stefan Feuerriegel, and Maytal Saar-Tsechansky**, "Algorithmic fairness in business analytics: Directions for research and practice," *Production and Operations Management*, 2022, *31* (10), 3749–3770.

**de Caritat Condorcet, Marie Jean Antoine Nicolas**, *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*, Paris: De l'Imprimerie Royale, 1785.

**del Arco, Flor Miriam Plaza, Debora Nozza, and Dirk Hovy**, "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" 2024, pp. 19–30.

**Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**, "Bert: Pre-training of deep bidirectional transformers for language understanding," in "Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)" 2019, pp. 4171–4186.

**Feng, Qi and J George Shanthikumar**, "How Research in Production and Operations Management May Evolve in the Era of Big Data," *Production and Operations Management*, 2018, *27* (9), 1670–1684.

**Fisher, R. A.**, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, 1915, *10* (4), 507–521.

**Fiss, Peer C and Edward J Zajac**, "The symbolic management of strategic change: Sensegiving via framing and decoupling," *Academy of Management Journal*, 2006, *49* (6), 1173–1193.

**Foroughifar, Mohsen, Rozhina Ghanavi, Avi Goldfarb, and Ryan Webb**, "Communicating Uncertainty Can Increase AI Adoption," *Available at SSRN*, 2025.

**Geifman, Yonatan and Ran El-Yaniv**, "Selective classification for deep neural networks," *Advances in neural information processing systems*, 2017, *30.*

**Geiger, R. Stuart, Dominique Cope, Jamie Ip et al.**, ""Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?," *Quantitative Science Studies*, 2021, *2* (3), 795–827.

**Gemma Team, Aishwarya Kamath, Johan Ferret et al.**, "Gemma 3 Technical Report," 2025.

**Geng, Jiahui, Fengyu Cai, Yuxia Wang et al.**, "A Survey of Confidence Estimation and Calibration in Large Language Models," in "Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)" 2024, pp. 6577–6595.

**Grattafiori, Aaron, Abhimanyu Dubey et al.**, "The Llama 3 Herd of Models," 2024.

**Hendrickx, Kilian, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis**, "Machine Learning with a Reject Option: A Survey," *Machine Learning*, 2024, *113* (5), 3073–3110.

**Humphreys, Ashlee**, "Semiotic structure and the legitimation of consumption practices: The case of casino gambling," *Journal of Consumer Research*, 2010, *37* (3), 490–510.

**Imdahl, Christina, William Schmidt, and Kai Hoberg**, "Targeted Automation and Sustaining Human-AI Learning," *Production and Operations Management*, 2026, *35* (4), 1520–1538.

**Ipeirotis, Panagiotis G, Foster Provost, and Jing Wang**, "Quality management on amazon mechanical turk," in "Proceedings of the ACM SIGKDD workshop on human computation" 2010, pp. 64–67.

**Kadavath, Saurav, Tom Conerly, Amanda Askell et al.**, "Language models (mostly) know what they know," *arXiv preprint arXiv:2207.05221*, 2022.

**Khare, Ritu, John D Burger, John S Aberdeen et al.**, "Scaling drug indication curation through crowdsourcing," *Database*, 2015, *2015*, bav016.

**Koleck, Theresa A, Caitlin Dreisbach, Philip E Bourne, and Suzanne Bakken**, "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review," *Journal of the American Medical Informatics Association*, 2019, *26* (4), 364–379.

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

**Kovács, Balázs and Amanda J Sharkey**, "The paradox of publicity: How awards can negatively affect the evaluation of quality," *Administrative Science Quarterly*, 2014, *59* (1), 1–33.

**Lee, Kyuhan and Sudha Ram**, "Intent-Driven Machine Learning for Fake News Detection: A Referential Domain Adaptation Approach," *Production and Operations Management*, 2026, *35* (7), 2890–2909.

**Li, Junyou, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye**, "More agents is all you need," *Transactions on Machine Learning Research*, 2024.

**Li, Peiyao, Noah Castelo, Zsolt Katona, and Miklos Sarvary**, "Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis," *Marketing Science*, 2024, *43* (2), 254–266.

**Li, Xinxin and Lorin M Hitt**, "Self-selection and information role of online product reviews," *Information systems research*, 2008, *19* (4), 456–474.

**Liu, Yizao and Rigoberto A Lopez**, "The impact of social media conversations on consumer brand choices," *Marketing Letters*, 2016, *27*, 1–13.

**Mathur, Saurabh**, "Clickbait-Detector," 2020.

**Mens, Gaël Le, Balázs Kovács, Michael T. Hannan, and Guillem Pros**, "Uncovering the semantics of concepts using GPT-4," *Proceedings of the National Academy of Sciences*, 2023, *120* (49), e2309350120.

**Mithas, Sunil, Zhi-Long Chen, Terence J V Saldanha, and Alysson De Oliveira Silveira**, "How will artificial intelligence and Industry 4.0 emerging technologies transform operations management?," *Production and Operations Management*, 2022, *31* (12), 4475–4487.

**Monin, Philippe, Niels Noorderhaven, Eero Vaara, and David Kroon**, "Giving sense to and making sense of justice in postmerger integration," *Academy of Management Journal*, 2013, *56* (1), 256–284.

**Moore, Don A and Paul J Healy**, "The Trouble With Overconfidence," *Psychological Review*, 2008, *115* (2), 502–517.

**Nair, Devadrita and Arnd Huchzermeier**, "Predictably Unpredictable? How Judgmental and Machine Learning Forecasts Complement Each Other," *Production and Operations Management*, 2024, *33* (5), 1214–1234.

**Ng, Ka Chung, Ping Fan Ke, Mike K P So, and Kar Yan Tam**, "Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach," *Production and Operations Management*, 2023, *32* (7), 2101–2122.

**Ni, Shiyu, Keping Bi, Lulu Yu, and Jiafeng Guo**, "Are Large Language Models More Honest in Their Probabilistic or Verbalized Confidence?," in "Information Retrieval: 30th China Conference (CCIR 2024)," Vol. 15418 of *Lecture Notes in Computer Science* Springer 2024, pp. 124–135.

**OpenAI, Josh Achiam, Steven Adler et al.**, "GPT-4 Technical Report," 2023.

**Ringel, Daniel**, "Creating Synthetic Experts with Generative Artificial Intelligence," Research Paper 4542949, Kenan Institute of Private Enterprise 2023.

**Sahoo, Pranab, Ayush Kumar Singh, Sriparna Saha et al.**, "A systematic survey of prompt engineering in large language models: Techniques and applications," *arXiv preprint arXiv:2402.07927*, 2024.

**Schoenegger, Philipp, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock**, "Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy," *Science Advances*, 2024, *10* (45), eadp1528.

**Sheng, Victor S, Foster Provost, and Panagiotis G Ipeirotis**, "Get another label? improving data quality and data mining using multiple, noisy labelers," in "Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining" 2008, pp. 614–622.

**Snow, Rion, Brendan O'Connor, Dan Jurafsky, and Andrew Y Ng**, "Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks," in "Proceedings of the 2008 conference on empirical methods in natural language processing" 2008, pp. 254–263.

**Sonenshein, Scott**, "We're changing—Or are we? Untangling the role of progressive, regressive, and stability narratives during strategic change implementation," *Academy of Management Journal*, 2010, *53* (3), 477–512.

**Terwiesch, Christian, Lennart Meincke, Karan Girotra et al.**, "Artificial intelligence and its impact on creativity and diversity: An empirical study of large language model-generated product ideas," *Production and Operations Management*, 2026. Advance online publication.

**Tian, Katherine, Eric Mitchell, Allan Zhou et al.**, "Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback," in "Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing" 2023, pp. 5433–5442.

**Vaara, Eero and Janne Tienari**, "On the narrative construction of multinational corporations: An antenarrative analysis of legitimation and resistance in a cross-border merger," *Organization Science*, 2011, *22* (2), 370–390.

**Wang, Junlin, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou**, "Mixture-of-Agents Enhances Large Language Model Capabilities," *arXiv preprint arXiv:2406.04692*, 2024.

**Wang, Xuezhi, Jason Wei, Dale Schuurmans et al.**, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," in "The Eleventh International Conference on Learning Representations" 2023.

**Wei, Jason, Xuezhi Wang, Dale Schuurmans et al.**, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems*, 2022, *35*, 24824–24837.

**Xiong, Miao, Zhiyuan Hu, Xinyang Lu et al.**, "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs," in "International Conference on Learning Representations (ICLR)" 2024.

**Xu, Hongshen, Zichen Zhu, Situo Zhang et al.**, "Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback," in "First Conference on Language Modeling (COLM)" 2024.

**Yang, An, Anfeng Li et al.**, "Qwen3 Technical Report," 2025.

**_ , Baosong Yang et al.**, "Qwen2.5 Technical Report," 2024.

**Zhou, Zhi-Hua**, *Ensemble Methods: Foundations and Algorithms*, 2 ed., Chapman and Hall/CRC, 2025.

# Online Appendix

Additional Results and Prompts

## A Additional Regression Results

### Majority Decision Rule

Tables A1 and A2 reproduce the main accuracy and unresolved-task regressions of Section 5 (Tables 2 and 3) under the two-thirds (67%) majority aggregation rule rather than the unanimity rule used in the main text. The majority rule accepts a classification when at least 67% of the $r$ repeated runs agree, whereas the unanimity rule requires all $r$ to agree. Majority is the more permissive of the two: resolved-task counts rise at the cost of accepting cells in which one or two of the $r$ runs dissented from the consensus. Both aggregation rules are standard in human-labeling protocols—unanimity in jury-style settings, two-thirds majority where moderate disagreement is acceptable—and the choice should leave the qualitative pattern intact, affecting only the magnitudes and the size of the resolved sample.

The qualitative patterns track the unanimity specifications closely. Confidence thresholds remain the dominant accuracy lever: each step in $d$ delivers a coefficient close to its unanimity counterpart, with $d = 100\%$ again the omitted highest-accuracy category. The redundancy coefficient stays effectively zero on the acquisition task and small but precisely estimated on clickbait, matching the unanimity finding. Few-shot prompting continues to shift the efficiency frontier on the acquisition task (positive accuracy coefficient, negative unresolved-task coefficient) and remains a movement along the frontier on clickbait (small negative accuracy coefficient paired with a negative unresolved-task coefficient). The LLM ranking preserves Qwen 3 at the top on both tasks. The headline findings of Section 5 therefore reflect the underlying mechanisms identified in the decision framework (Section 3) rather than artifacts of the unanimity aggregation rule.

**Table A1.** Regression results for the acquisition task under majority agreement.

| | *Dep. Var.: Correlation* | | | *Dep. Var.: Number Unresolved* | | |
|---|---|---|---|---|---|---|
| | (1) N ≥300 | (2) N≥300; redundancy ≤10 | (3) N≥200 | (4) N ≥300 | (5) N≥300; redundancy ≤10 | (6) N≥200 |
| Confidence = 50% | −0.164*** | −0.150*** | −0.162*** | −454.504*** | −444.631*** | −508.390*** |
| | (0.011) | (0.014) | (0.010) | (11.928) | (16.547) | (10.831) |
| Confidence = 75% | −0.174*** | −0.163*** | −0.172*** | −357.711*** | −361.575*** | −409.291*** |
| | (0.011) | (0.014) | (0.010) | (11.937) | (16.304) | (10.750) |
| Confidence = 85% | −0.143*** | −0.136*** | −0.142*** | −241.934*** | −248.210*** | −286.297*** |
| | (0.011) | (0.014) | (0.010) | (11.749) | (16.129) | (10.508) |
| Confidence = 90% | −0.136*** | −0.127*** | −0.134*** | −228.187*** | −231.113*** | −258.740*** |
| | (0.011) | (0.014) | (0.010) | (11.938) | (16.472) | (10.799) |
| Confidence = 95% | −0.094*** | −0.094*** | −0.094*** | −152.508*** | −162.789*** | −174.027*** |
| | (0.011) | (0.014) | (0.010) | (12.178) | (16.602) | (11.027) |
| Redundancy | 0.000 | 0.001 | 0.000 | 3.754*** | 6.717*** | 3.997*** |
| | (0.000) | (0.001) | (0.000) | (0.311) | (1.073) | (0.318) |
| GPT-4 | 0.232*** | 0.211*** | 0.235*** | −130.638*** | −136.875*** | −135.446*** |
| | (0.007) | (0.010) | (0.007) | (9.393) | (14.584) | (9.681) |
| Llama 3.1 | 0.081*** | 0.049*** | 0.085*** | 72.619*** | 54.115*** | 86.428*** |
| | (0.007) | (0.009) | (0.006) | (6.449) | (9.146) | (6.718) |
| Qwen 2.5 | 0.124*** | 0.137*** | 0.128*** | 143.641*** | 127.653*** | 174.062*** |
| | (0.006) | (0.008) | (0.006) | (6.851) | (10.106) | (6.578) |
| Qwen 3 | 0.343*** | 0.345*** | 0.348*** | −6.840 | −56.493*** | 0.151 |
| | (0.004) | (0.005) | (0.004) | (4.986) | (7.776) | (5.412) |
| Temperature = 1.0 | −0.006 | −0.008 | −0.006 | 60.622*** | 47.500*** | 63.567*** |
| | (0.004) | (0.006) | (0.005) | (4.935) | (7.352) | (5.086) |
| Temperature = 2.0 | −0.070*** | −0.055*** | −0.079*** | 156.951*** | 143.932*** | 175.766*** |
| | (0.005) | (0.006) | (0.005) | (6.027) | (8.554) | (6.215) |
| Few-shot | 0.264*** | 0.252*** | 0.265*** | −138.434*** | −132.588*** | −147.334*** |
| | (0.004) | (0.005) | (0.004) | (4.183) | (6.272) | (4.362) |
| Constant | 0.406*** | 0.404*** | 0.405*** | 517.503*** | 530.510*** | 554.553*** |
| | (0.012) | (0.016) | (0.011) | (12.368) | (16.717) | (11.004) |
| Observations | 3,137 | 1,454 | 3,351 | 3,137 | 1,454 | 3,351 |
| $R^2$ | 0.770 | 0.805 | 0.756 | 0.619 | 0.619 | 0.662 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

**Table A2.** Regression results for the clickbait task under majority agreement.

| | *Dep. Var.: Correlation* | | | *Dep. Var.: Number Unresolved* | | |
|---|---|---|---|---|---|---|
| | (1) N ≥300 | (2) N≥300; redundancy ≤10 | (3) N≥200 | (4) N ≥300 | (5) N≥300; redundancy ≤10 | (6) N≥200 |
| Confidence = 50% | −0.225*** | −0.204*** | −0.227*** | −488.515*** | −449.124*** | −527.082*** |
| | (0.009) | (0.013) | (0.008) | (12.157) | (18.044) | (12.056) |
| Confidence = 75% | −0.160*** | −0.145*** | −0.139*** | −359.894*** | −324.887*** | −356.346*** |
| | (0.009) | (0.013) | (0.008) | (13.669) | (19.833) | (13.615) |
| Confidence = 85% | −0.094*** | −0.083*** | −0.080*** | −202.939*** | −171.125*** | −176.371*** |
| | (0.008) | (0.011) | (0.007) | (13.848) | (21.348) | (14.424) |
| Redundancy | 0.002*** | 0.009*** | 0.003*** | 3.065*** | 15.455*** | 5.898*** |
| | (0.000) | (0.001) | (0.000) | (0.456) | (2.001) | (0.565) |
| Qwen 2.5 | 0.211*** | 0.221*** | 0.182*** | −426.813*** | −386.780*** | −455.887*** |
| | (0.012) | (0.015) | (0.011) | (10.022) | (15.013) | (10.610) |
| Qwen 3 | 0.259*** | 0.261*** | 0.243*** | −304.316*** | −261.200*** | −289.404*** |
| | (0.012) | (0.015) | (0.011) | (12.285) | (16.962) | (12.090) |
| Temperature = 1.0 | −0.015*** | −0.004 | −0.007 | 36.653*** | 29.748** | 38.202*** |
| | (0.005) | (0.009) | (0.006) | (8.820) | (14.119) | (9.615) |
| Temperature = 2.0 | 0.032*** | 0.038*** | 0.027*** | 86.991*** | 82.001*** | 101.379*** |
| | (0.007) | (0.009) | (0.006) | (9.230) | (14.265) | (10.331) |
| Few-shot | −0.023*** | −0.026*** | −0.034*** | −31.372*** | −52.512*** | −63.577*** |
| | (0.006) | (0.008) | (0.005) | (8.340) | (11.742) | (8.289) |
| Constant | 0.546*** | 0.485*** | 0.553*** | 862.730*** | 750.755*** | 882.446*** |
| | (0.017) | (0.023) | (0.015) | (17.968) | (25.335) | (18.191) |
| Observations | 1,137 | 517 | 1,287 | 1,137 | 517 | 1,287 |
| $R^2$ | 0.699 | 0.732 | 0.669 | 0.708 | 0.663 | 0.690 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## B Robustness: Fisher Z-Transformed Specifications

We report robustness checks that use Fisher Z-transformed correlations ($Z = \text{arctanh}(r) = \frac{1}{2}\ln\frac{1+r}{1-r}$) as the dependent variable, addressing the bounded support and heteroskedastic variance of the Pearson correlations used in the main regressions (Fisher, 1915). For the clickbait task, Llama 3.1 is excluded because near-zero variance in its responses produces boundary correlations across all confidence thresholds, consistent with the limited classification variation noted in the main analysis. As a result, there are fewer observations in Table A4 as compared to the main clickbait specification, Table 3, and the omitted LLM reference category is Qwen 3 instead of Llama 3.1.

Tables A3 and A4 report the Fisher Z-transformed specifications for the acquisition and clickbait tasks under unanimous agreement. Results confirm the main findings throughout: confidence thresholds are the primary driver of accuracy gains, and the redundancy coefficient remains negligible relative to the confidence-threshold coefficients. The Fisher-Z acquisition specification detects a small positive redundancy effect that is statistically significant but economically negligible in magnitude, without consequence for the paper's guidelines.

**Table A3.** Fisher Z-transformed regression results for the acquisition task under unanimous agreement.

| | (1) N ≥300 | (2) N≥300; redundancy ≤10 | (3) N≥200 |
|---|---|---|---|
| Confidence = 50% | −0.336*** | −0.343*** | −0.382*** |
| | (0.021) | (0.028) | (0.021) |
| Confidence = 75% | −0.358*** | −0.360*** | −0.402*** |
| | (0.021) | (0.028) | (0.021) |
| Confidence = 85% | −0.334*** | −0.327*** | −0.374*** |
| | (0.021) | (0.028) | (0.021) |
| Confidence = 90% | −0.338*** | −0.325*** | −0.370*** |
| | (0.022) | (0.029) | (0.021) |
| Confidence = 95% | −0.247*** | −0.258*** | −0.291*** |
| | (0.022) | (0.029) | (0.022) |
| Redundancy | 0.001** | 0.006*** | 0.002*** |
| | (0.001) | (0.002) | (0.001) |
| GPT-4 | 0.331*** | 0.267*** | 0.319*** |
| | (0.013) | (0.018) | (0.013) |
| Llama 3.1 | 0.016 | 0.004 | 0.034*** |
| | (0.011) | (0.014) | (0.011) |
| Qwen 2.5 | 0.152*** | 0.178*** | 0.176*** |
| | (0.009) | (0.012) | (0.010) |
| Qwen 3 | 0.759*** | 0.716*** | 0.777*** |
| | (0.012) | (0.016) | (0.013) |
| Temperature = 1.0 | −0.047*** | −0.025* | −0.028*** |
| | (0.009) | (0.013) | (0.010) |
| Temperature = 2.0 | −0.091*** | −0.076*** | −0.098*** |
| | (0.012) | (0.014) | (0.012) |
| Few-shot | 0.515*** | 0.463*** | 0.512*** |
| | (0.008) | (0.011) | (0.008) |
| Constant | 0.504*** | 0.511*** | 0.527*** |
| | (0.022) | (0.030) | (0.021) |
| Observations | 3,080 | 1,447 | 3,315 |
| $R^2$ | 0.768 | 0.785 | 0.752 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

**Table A4.** Fisher Z-transformed regression results for the clickbait task under unanimous agreement.

| | (1) N ≥300 | (2) N≥300; redundancy ≤10 | (3) N≥200 |
|---|---|---|---|
| Confidence = 50% | −0.367*** | −0.349*** | −0.357*** |
| | (0.018) | (0.023) | (0.016) |
| Confidence = 75% | −0.299*** | −0.274*** | −0.273*** |
| | (0.018) | (0.024) | (0.016) |
| Confidence = 85% | −0.198*** | −0.176*** | −0.178*** |
| | (0.018) | (0.023) | (0.016) |
| Redundancy | 0.006*** | 0.018*** | 0.007*** |
| | (0.000) | (0.002) | (0.001) |
| Qwen 2.5 | −0.128*** | −0.087*** | −0.147*** |
| | (0.007) | (0.010) | (0.007) |
| Temperature = 1.0 | 0.037*** | 0.020* | 0.037*** |
| | (0.008) | (0.012) | (0.008) |
| Temperature = 2.0 | 0.046*** | 0.023** | 0.036*** |
| | (0.007) | (0.011) | (0.007) |
| Few-shot | −0.050*** | −0.061*** | −0.069*** |
| | (0.006) | (0.009) | (0.006) |
| Constant | 1.091*** | 1.008*** | 1.091*** |
| | (0.020) | (0.025) | (0.017) |
| Observations | 919 | 409 | 1,026 |
| $R^2$ | 0.643 | 0.649 | 0.668 |

Robust standard errors in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## C Accuracy Surface by LLM

Figure A1 visualizes the acquisition-task accuracy surface as a contour plot over the two primary configuration levers—redundancy ($r$) and confidence threshold ($d$)—separately for the four LLMs displayed (Gemma 3n, Llama 3.1, Qwen 2.5, Qwen 3). The plot complements Figure 2 in the main text, which collapses across LLMs to display the accuracy-cost trade-off as a single curve per task; here we hold the LLM fixed and trace the full $(r, d)$ surface model by model.

The dominant visual signature is the orientation of the contour bands, which display accuracy against confidence threshold and degree of redundancy. Looking across all four panels, it is evident that Qwen 3 (as a reasoning model) displays far superior performance to the remaining three LLMs, at all levels of confidence threshold. On the redundancy axis, any accuracy differences are small, confined to $r \leq 3$, and dwarfed by confidence-axis variation—consistent with the near-zero linear redundancy coefficient in Table 2—while further increases in redundancy level merely add cost. Since, as of this writing, Qwen 3—like other reasoning models—consumes substantially more inference compute per run than comparable non-reasoning models, this suggests that any use of redundancy should be capped at small $r$.

Given the distinctly superior performance of Qwen 3, the rest of our discussion evaluates and compares the remaining LLMs. For Gemma 3n and Qwen 2.5 the bands run essentially parallel to the redundancy axis, indicating that accuracy varies almost entirely with the confidence threshold rather than with redundancy. The Llama 3.1 panel shows the most curvature among these models: at permissive confidence thresholds the bands tilt toward the lower-left, so additional redundancy provides modest accuracy recovery when $d$ is loose, before the band orientation flattens to horizontal once $d$ reaches the mid-80s. Even so, confidence remains the larger lever on the Llama 3.1 surface, consistent with its near-zero redundancy coefficient and the much larger confidence-threshold coefficients reported for Llama 3.1 in Table 4.

The panels also expose a sizable across-LLM accuracy gap among the three non-reasoning models. Ignoring Qwen 3, Qwen 2.5 reaches the brightest contour band ($\approx$ 90–95%) in the narrow strip at $d \geq 95\%$ and runs near $\approx$ 65–70% across the rest of its surface; Gemma 3n peaks at $\approx$ 80–85% at strict $d$ over a baseline near $\approx$ 55–60%; Llama 3.1 stays in the purple range across the full surface and does not break past the $\approx$ 55–60% band even at $d = 100\%$. In every panel, the highest-accuracy regions concentrate in the narrow upper strip where $d$ exceeds $\approx$ 95%, with sharper color gradients between $d = 95\%$ and $d = 100\%$ than between lower thresholds—the graphical analogue of the accelerating accuracy gain documented for the final threshold step in Section 5, where the $d = 95\%$ to $d = 100\%$ jump alone yields an 18% gain relative to the mean.

The clickbait task is covered separately by the regression coefficients in Table 3 and the LLM-by-LLM accuracy curves in Figure 3 of the main text, both of which exhibit the same confidence-dominated structure.

Figure A2 plots the companion surface for the share of unresolved tasks, over the same $(r, d)$ grid and the same four LLMs. In contrast to the accuracy surface—where variation runs almost entirely along the confidence axis—the unresolved-task surface responds to both levers: the share of unresolved tasks rises as the confidence threshold tightens and, separately, as redundancy increases. This is the graphical counterpart of the regression evidence in Section 5 that redundancy's principal cost is unresolved tasks rather than forgone accuracy.

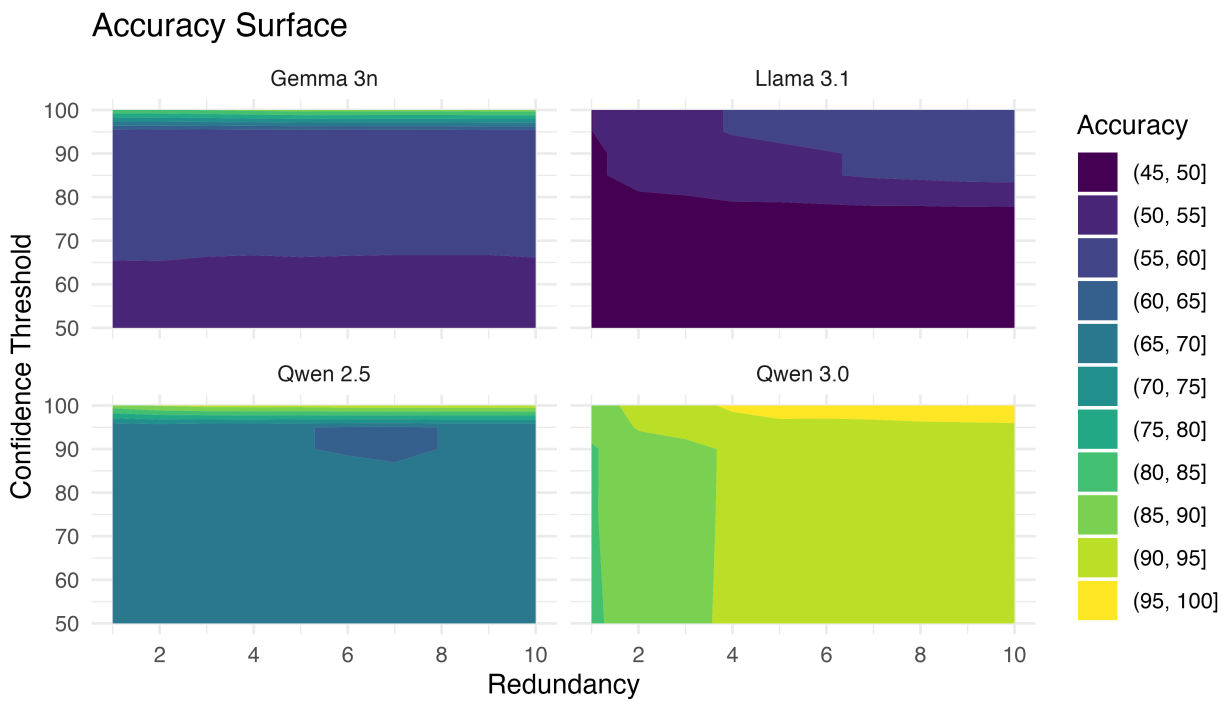


**Figure A1. Confidence, not redundancy, drives accuracy.** Accuracy surface for the acquisition task under unanimous agreement, by LLM (Gemma 3n, Llama 3.1, Qwen 2.5, Qwen 3). Lighter (yellow) regions indicate higher accuracy; darker (purple) regions indicate lower accuracy. In every panel, accuracy variation runs primarily along the confidence-threshold axis rather than the redundancy axis. The clickbait task is covered separately in Table 3 and Figure 3.

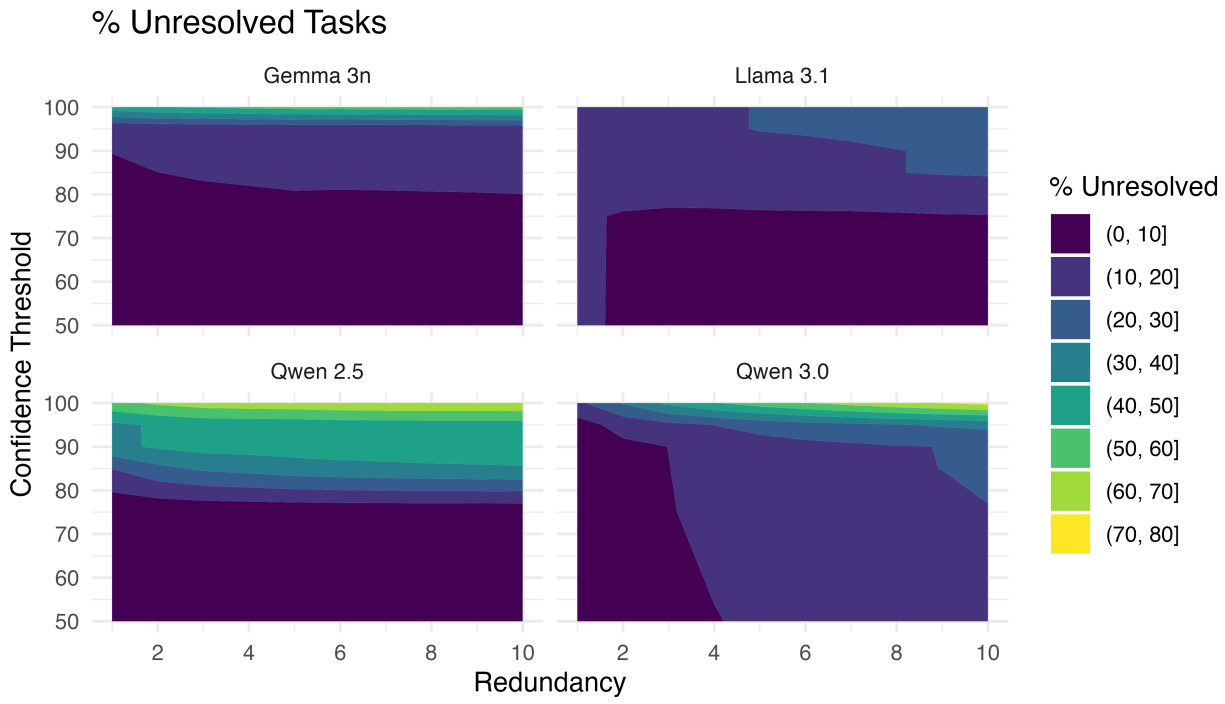


**Figure A2. Confidence and redundancy both drive unresolved tasks.** Percent unresolved tasks for the acquisition task under unanimous agreement, by LLM (Gemma 3n, Llama 3.1, Qwen 2.5, Qwen 3). Lighter (yellow) regions indicate a higher share of unresolved tasks; darker (purple) regions indicate a lower share. The clickbait task is covered separately in Table 3.

## D LLM Execution: Prompts

We used two prompting styles for each classification task, drawn from the prompt-design literature (Brown et al., 2020; Wei et al., 2022; Sahoo et al., 2024): *zero-shot prompting* and *few-shot prompting*. The zero-shot prompt specifies the task, the allowed class labels, the confidence field, and the required output format, with no worked examples and no task-specific definitions. The few-shot prompt keeps the same role, class labels, confidence field, and output format, but adds task-specific definitions, worked examples, and edge-case rules. This design varies the amount of in-context task guidance while holding the requested output structure fixed. None of the worked examples shown in the few-shot prompts below is drawn from the acquisition or clickbait headline samples classified in our experiments; they are included solely to illustrate the task.

The prompts below report the exact prompt text used for both classification tasks—acquisition and clickbait—under the zero-shot and few-shot prompting styles.

**Acquisition Task: Zero-shot Prompt**

**[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]**

**Acquisition Task: Few-shot Prompt**

**[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]**

### Clickbait Task: Zero-shot Prompt

**[SYSTEM]** You are a clickbait detection specialist.

**[USER]** Analyze the headline below and classify it as clickbait (C) or straight news (N).
Clickbait uses sensational language, curiosity gaps, numbered lists, or emotional manipulation to drive clicks.
Straight news reports facts directly without sensationalism.

Output format: Single markdown table row with 4 columns:
| Classification | Confidence % | Hook Phrase | Main Topic |
- Column 1: "C" for clickbait, "N" for news
- Column 2: Confidence percentage (0-100)
- Column 3: Key clickbait phrase (or "NA" if none)
- Column 4: Main subject/entity (or "NA" if unclear)

Headline: **[HEADLINE]**

### Clickbait Task: Few-shot Prompt

**[SYSTEM]** You are a clickbait detection specialist.

**[USER]** Analyze the headline below and classify it as clickbait (C) or straight news (N).
Clickbait uses sensational language, curiosity gaps, numbered lists, or emotional manipulation to drive clicks.
Straight news reports facts directly without sensationalism.

CLICKBAIT PATTERNS:
- Curiosity gaps: "You Won't Believe..."
- Numbered lists: "17 Shocking Facts..."
- Emotional manipulation: "This Will Make You Cry"
- Superlatives: "Most Amazing Ever"
- Direct address: "Here's Why You Need..."

Output format: Single markdown table row with 4 columns:
| Classification | Confidence % | Hook Phrase | Main Topic |
- Column 1: "C" for clickbait, "N" for news
- Column 2: Confidence percentage (0-100)
- Column 3: Key clickbait phrase (or "NA" if none)
- Column 4: Main subject/entity (or "NA" if unclear)

EXAMPLES:
Headline: Scientists Discover New Species in Amazon Rainforest
| N | 85 | NA | Amazon species |

Headline: 24 Times Celebrities Completely Lost It
| C | 90 | 24 Times | Celebrities |

Headline: This Photo Will Change How You See the World Forever
| C | 95 | Will Change How You See | Photo |

Headline: Federal Reserve Raises Interest Rates by 0.25%
| N | 90 | NA | Federal Reserve |

Headline: **[HEADLINE]**