# Class Imbalance and Batch Effects in LLM-Based Screening for Systematic Reviews

**Gilberto Sussumu Hida**[1], **Danilo Monteiro Ribeiro**[1], **Clayton Suguio Hida**[2]

[1]Cesar School – AIBL
Recife - PE - Brazil

[2]UEAP
Macapá - AP - Brazil

gsh@cesar.school, dmr@cesar.school, clayton.hida@ueap.edu.br

***Abstract.*** *This study analyses LLMs in imbalanced binary classification, using study screening in systematic reviews as the application domain. An experiment was conducted in five reviews, comparing individual and batch processing, with and without prevalence metadata. The results indicate a limited influence of the prevalence metadata, with no evidence that it improves performance. In contrast, batch processing produced larger behavioral changes that varied according to the prevalence of the class. The aggregate and item-level analyses did not always coincide. Therefore, batch processing should be evaluated not only in terms of cost, but also in relation to its effects on decision-making behavior.*

## 1. Introduction

Systematic literature reviews (SLRs) have become a central practice in Software Engineering [Kitchenham et al. 2004]. With the growth in publication volume [Bornmann et al. 2021], the title and abstract screening stage has become increasingly costly, requiring the classification of hundreds to thousands of records [Chai et al. 2021]. Large language models (LLMs) have been explored as a support for this task [Hida et al. 2026b, Syriani et al. 2024, Huotala et al. 2025], with promising results in aggregate performance, but still limited understanding of their decision behavior under operational screening characteristics.

Two of these characteristics are methodologically central. The first is class imbalance: in most screening processes, only a fraction of records is included, making traditional metrics misleading and shifting the optimal point toward minimizing false negatives [Hida and Do Nascimento 2025]. The second is the processing mode: LLMs can receive one study per call (individual) or multiple studies in a single call (batch), which reduces cost [Pipal et al. 2026] but can induce implicit comparison between candidates, quota-like effects, or sensitivity to position in batch [Liu et al. 2024].

Recent literature has focused on comparisons between models and prompts, often under individual processing and without isolating the effect of contextual metadata [Madeyski et al. 2025, Syriani et al. 2024]. It remains unclear whether informing screening metadata, such as expected prevalence, changes LLMs' propensity to include or exclude studies, and whether the processing mode has a comparable effect.

This work addresses this gap with a factorial design of $2\times2$: presence or absence of screening metadata in the prompt and processing mode (individual or

batch). We selected five reviews from SESR-Eval [Huotala et al. 2025], with prevalence ranging from $2.9\%$ to $53.0\%$, and evaluated two models: Llama-3.3-70B-Instruct-Turbo [Meta AI 2024] and GPT-5-mini [OpenAI 2025].

The contributions are: (a) evidence that informing prevalence metadata in the prompt produced a limited effect, whereas batch processing has a larger and more variable behavioral effect across reviews; (b) the observation that aggregate and item-level analyses may diverge; and (c) implications for the responsible use of LLMs in screening, showing that how articles are presented to the model matters more than prompt-based calibration.

## 2. Related Work

The application of LLMs to title and abstract screening has been investigated in systematic reviews, including in the Software Engineering domain. Syriani et al. [Syriani et al. 2024] evaluated ChatGPT in this context and observed relevant variation between reviews. Huotala et al. [Huotala et al. 2025] proposed SESR-Eval, a benchmark with 34,528 labeled primary studies from 24 secondary reviews, showing that variation between reviews may exceed variation between models. In a methodological review, Madeyski and Kitchenham [Madeyski et al. 2025] highlight recurring limitations in the evaluation of LLMs for screening, including the underuse of metrics suitable for imbalanced scenarios.

Another line of work investigates how prompt formulation and calibration mechanisms influence LLM performance. Wang et al. [Wang et al. 2024] evaluated LLMs in a zero-shot setting and proposed calibration based on target recall, while Huotala et al. [Huotala et al. 2024] compared zero-shot, one-shot and few-shot configurations. These studies indicate that instructions, examples, and thresholds can alter performance, but they do not directly isolate the effect of contextual screening metadata, such as informed prevalence, on the models' decision propensity.

Batch processing has been treated mainly as an efficiency strategy. Cheng et al. [Cheng et al. 2023] and Pipal et al. [Pipal et al. 2026] showed that batch prompting reduces token, time, and cost while maintaining performance close to individual processing. However, increasing the context may alter the model behavior. Liu et al. [Liu et al. 2024] documented the Lost in the Middle phenomenon, in which performance varies according to the position of information in the context. Fagerberg et al. [Fagerberg et al. 2026] showed that aggregate sensitivity for GPT-5-mini does not degrade up to batch sizes of 200, degrading sharply only at 400. Thus, batch processing should be analyzed as an experimental condition, not only as a cost optimization.

Finally, evaluating LLMs in screening must consider class imbalance, since Include tends to be the minority class and accuracy-based metrics may be misleading. Madeyski and Kitchenham [Madeyski et al. 2025] discuss the divergences between traditional metrics and metrics that are more sensitive to asymmetric cost. Hida and Do Nascimento [Hida and Do Nascimento 2025] discuss robust metrics for imbalanced classification, while Hida et al. [Hida et al. 2026b] show that variability between executions of the same LLM can be comparable to differences between models, recommending more stable agreement metrics and measures aligned with the cost of false negatives, such as the F2-score.

Although the literature advances in benchmarks, prompting, batch processing, and

metrics, three gaps remain: batch processing is commonly treated as an efficiency technique, rather than as a behavioral factor; prompts rarely incorporate contextual screening metadata; and evaluation tends to emphasize aggregate metrics, without examining item-level changes in decisions. This study investigates these gaps through a controlled experimental design.

## 3. Methodology

This controlled experiment evaluated how screening metadata and processing mode influence LLM behavior in a binary classification task involving scientific studies. Each candidate study was classified as *Exclude* or *Include*, based on its title, abstract, and eligibility criteria for the corresponding secondary review. Figure 1 summarizes the overall analysis workflow.

The analysis was guided by four research questions:

- **RQ1**) Do screening metadata change the propensity of LLMs to classify studies as *Include*?
- **RQ2**) Does batch processing modify decisions compared to individual processing?
- **RQ3**) How does batch processing, with or without metadata, affect false negatives and direction of decision changes?
- **RQ4**) Are the observed effects consistent across models from different families?

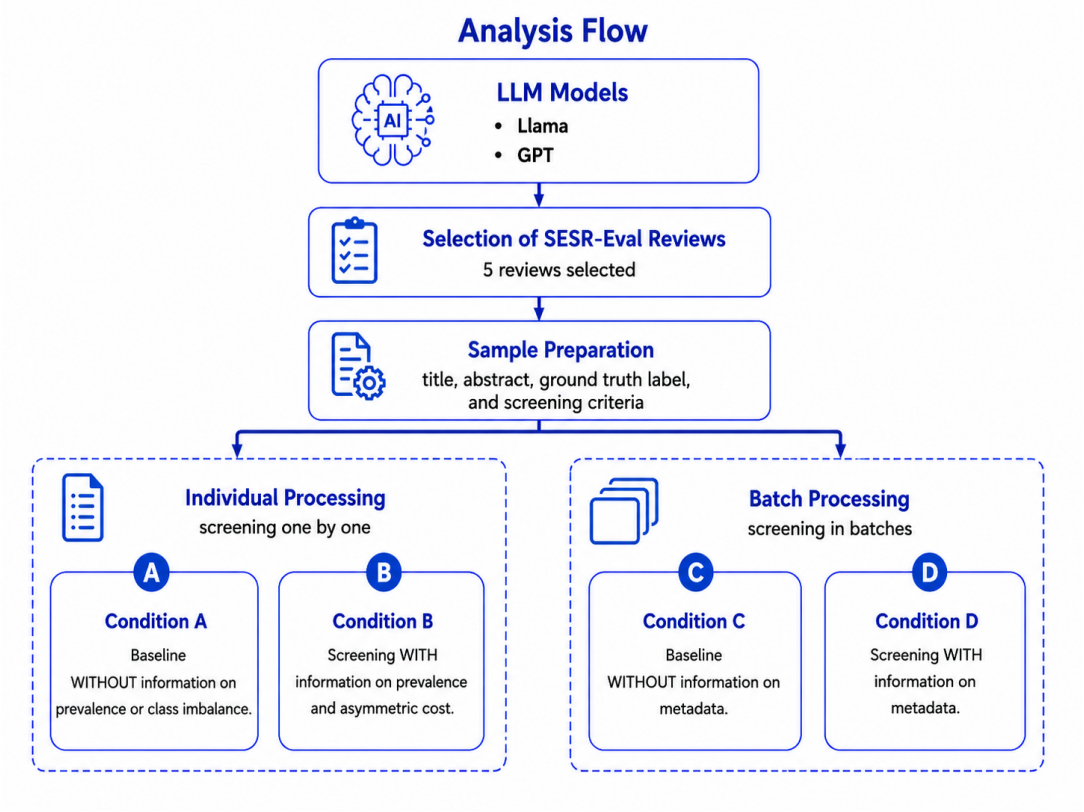



**Figure 1. Workflow of the analysis conducted in the study.**

### 3.1. Experimental design

The experimental design followed a factorial structure of $2\times2$, which combined two factors: the presence or absence of metadata for the selection in the prompt and processing mode, individual or batch. The metadata consisted of a single sentence informing the true (oracle) estimated inclusion rate of the *Include* class in the review, framed as class-imbalance context. The four conditions were: A) individual without metadata; B) individual with metadata; C) batch without metadata; and D) batch with metadata.

### 3.2. Dataset, models, and data preparation

We used SESR-Eval [Huotala et al. 2025], a title-abstract screening dataset built from Software Engineering secondary studies. It was chosen because it provides reference labels and reviews spanning different Include-class prevalence levels, allowing observation of model behavior across degrees of imbalance.

Five secondary reviews were selected to cover different positive-class prevalence values, from highly imbalanced to almost balanced. Reviews lacking one of the two classes, with few records, or with insufficient eligibility criteria were discarded. Table 1 presents the characterization of the reviews used.

**Table 1. Characterization of the selected SESR-Eval reviews.**

| ID | Ref. | N | Inc. | Exc. | Prev. |
|---|---|---|---|---|---|
| rsl_0 | [Kuutila et al. 2020] | 1,812 | 53 | 1,759 | 2.9% |
| rsl_1 | [Rani et al. 2023] | 2,048 | 71 | 1,977 | 3.5% |
| rsl_2 | [Lewowski and Madeyski 2022] | 1,527 | 160 | 1,367 | 10.5% |
| rsl_3 | [Tambon et al. 2022] | 955 | 166 | 789 | 17.4% |
| rsl_4 | [Sharbaf et al. 2023] | 300 | 159 | 141 | 53.0% |

*Note:* N represents the number of studies considered after data preparation for analysis.

The unit of analysis was the candidate study. Each instance contains the study identifier, title, abstract, reference label, and the secondary study to which it belongs. The original label was binarized as 0 for *Exclude* and 1 for *Include*, and the decisions made by the LLMs were treated with the same coding.

Two models were evaluated: Llama-3.3-70B-Instruct-Turbo [Meta AI 2024], accessed via the Together AI API [Together AI 2024], and gpt-5-mini-2025-08-07 [OpenAI 2025], accessed via the OpenAI API. Experiments were executed between March and May 2026. The same elements, conditions, and prompt templates were used for both models. Calls were executed with temperature 0; invalid or unexpected-format responses were monitored, with no invalid decisions in the final set. The comparison between models was treated as a robustness analysis and not as a competitive ranking of performance.

### 3.3. Prompts and scenario execution

The prompts were written in English, following the language of the dataset titles, abstracts, and criteria (available in the artifact package [Hida et al. 2026a]). All conditions followed a common structure that contained the task definition, eligibility criteria, screening instructions, and candidate-study data. The output was requested in JSON format,

with the study identifier and binary decision, in order to reduce parsing ambiguities and enable programmatic validation. In batch conditions (C and D), studies were randomly shuffled with a fixed seed and grouped into batches of 20, with the same batch composition and order reused across conditions C and D; each batch returned one decision per study. This size is well below the batch sizes (400 and above) at which Fagerberg et al. [Fagerberg et al. 2026] observed sharp performance degradation for GPT-5-mini, although it was not itself among the specific batch sizes they tested. The exact prompt templates for all four conditions and both model providers are reproduced in the supplementary material accompanying the artifact package [Hida et al. 2026a].

### 3.4. Metrics, analytical comparisons, and uncertainty

The F2-score was adopted because it assigns greater weight to *recall*, which is appropriate for the screening context, in which false negatives are more critical than false positives [Hida et al. 2026b]. The false negative rate represented the risk of losing relevant studies, while Gwet's AC1 [Vach and Gerke 2023, Hida et al. 2026b] was used as an agreement metric due to its stability in extreme-prevalence scenarios. Because universal interpretation thresholds for agreement coefficients are context-dependent, AC1 values were interpreted primarily in comparative terms across reviews, models, and experimental conditions, rather than as absolute quality categories. This choice is particularly relevant in imbalanced screening settings, where agreement coefficients may be affected by prevalence and marginal distributions.

In addition to aggregate metrics, item-level indicators were calculated. The Decision Flip Rate (DFR) was defined as the proportion of studies whose decision changed between two conditions, inspired by decision flip and prediction churn metrics [Laban et al. 2023]. The *net shift* was defined as the net difference between the changes from *Exclude* to *Include* and from *Include* to *Exclude*, indicating whether a condition made the model more inclusive or more restrictive.

The comparisons A↔B and C↔D assessed the effect of metadata in the individual and batch modes, respectively; A↔C and B↔D assessed the effect of batch processing without and with metadata. McNemar's exact test and article-level improvement or worsening analysis were also used. Paired bootstrap [Bestgen 2022] confidence intervals were calculated for differences in F2-score and DFR, preserving the correspondence between studies in the compared conditions. These procedures assume independence; batch decisions sharing a prompt may be correlated, so uncertainty in batch comparisons may be underestimated.

### 3.5. Reproducibility

The study artifacts, including data tables, Jupyter notebooks used to prepare the data, process the LLM outputs, and reconstruct the analyses, tables, and figures, and a supplementary document with the exact prompt templates for all four conditions, are available in a Zenodo record (DOI: 10.5281/zenodo.20414963).

## 4. Results and Discussion

This section examines how batch processing and screening metadata affect the decision behavior of LLMs. The analysis combines three sources of evidence: agreement with the reference label, changes at the item-level, and the aggregate impact on performance.

### 4.1. Overview of agreement

Figure 2 presents an overview of the agreement between LLM decisions and the reference label, using Gwet's AC1. The most evident pattern is that GPT-5-mini presented a higher and more stable agreement than Llama-3.3-70B-Instruct-Turbo in almost all reviews and conditions. In Llama, batch conditions reduced the agreement in reviews such as rsl_0 and rsl_3, suggesting that the mode of presentation of studies can alter the agreement between LLM decisions and the reference label. In GPT-5-mini, the variation between conditions was smaller, although the nearly balanced review (rsl_4) also presented the lowest AC1 values.

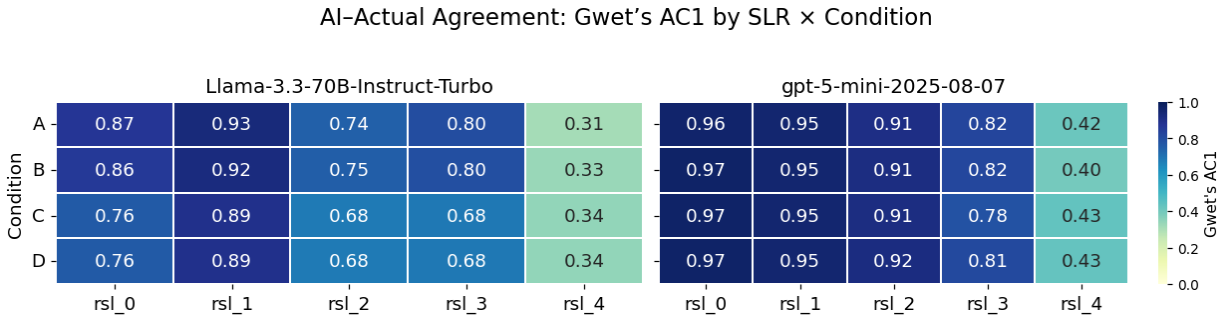



**Figure 2. Gwet's AC1 by systematic review, experimental condition, and model.**

This initial reading is important for two reasons. First, agreement varies more between reviews than between some conditions within the same review, indicating that the nature of the secondary study and the prevalence of the Include class influence the results. Second, given that the rsl_i labels were ordered so that rsl_0 has the lowest prevalence proportion (<3%) and rsl_4 the highest (53%), the lower agreement in rsl_4 shows that more balanced reviews are not necessarily easier for LLMs; in contrast, they may require finer distinctions between relevant and irrelevant studies.

### 4.2. RQ1: Do screening metadata change the propensity of LLMs to classify studies as *Include*?

RQ1 examines whether the inclusion of screening metadata (informed prevalence) changes LLM decisions in two contexts: individual processing (A↔B) and batch processing (C↔D). Figure 3 presents the DFR for all pairwise comparisons, by model.

In individual processing (A↔B), the metadata produced a small effect in both models. In Llama, DFR ranged from $0.9\%$ (`rsl_1`) to $2.3\%$ (`rsl_4`), and the corresponding $\Delta$F2-scores remained close to zero, with 95% CIs crossing zero in all five reviews (Figure 4). In GPT-5-mini, the DFRs ranged from $0.5\%$ to $6.3\%$, with a greater amplitude in `rsl_4`. Although McNemar's test indicated statistical evidence in some combinations ( $p < 0.05$ in three of the five SLRs for Llama), the magnitude suggests a small effect. Because models were given the true prevalence rather than a pilot estimate, this result reflects the effect of oracle prevalence metadata, not necessarily the effect of the imperfect estimates available in practice, with no uniform directional pattern of change.

The most notable result occurred in C↔D for Llama: DFR was 0.000 in the five SLRs, preserving 6,642 decisions. This suggests that prevalence metadata did not alter

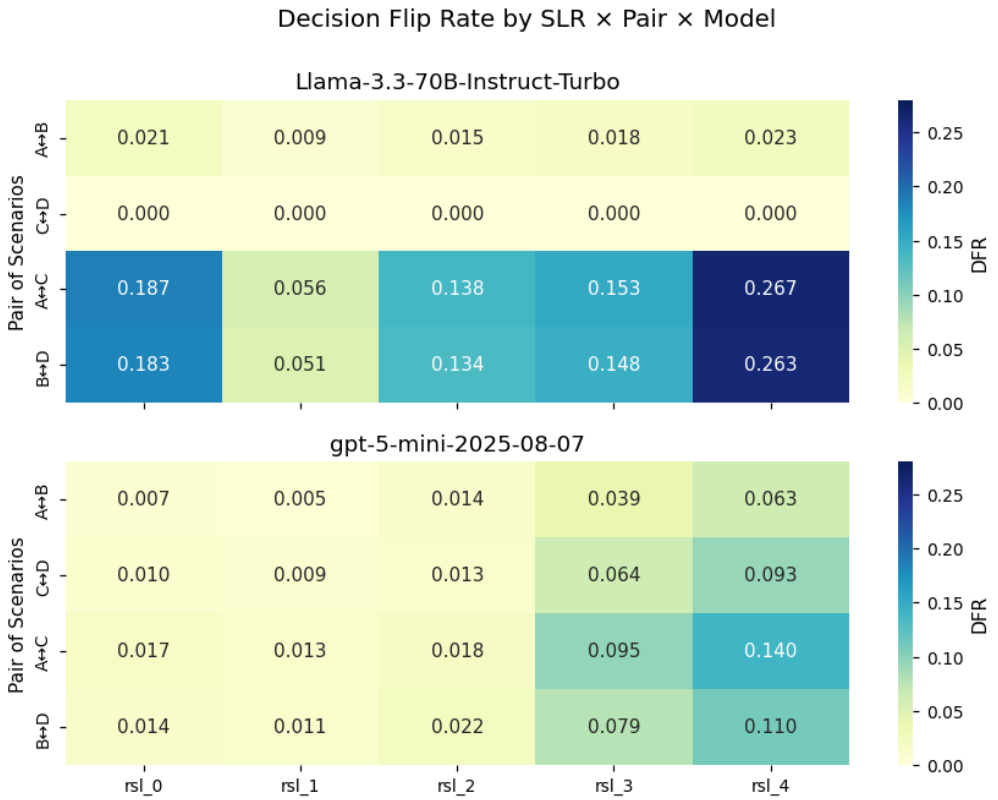



**Figure 3. Decision Flip Rate by condition pair, SLR, and model.**

decisions when inserted into prompts with multiple studies. The finding is consistent with the hypothesis that, in long contexts, the distribution of attention between items may attenuate contextual information [Liu et al. 2024]. In addition, the absence of changes indicates low stochastic noise between C and D, which share the same batch prompt structure; this does not rule out variation in A↔C and B↔D, whose prompts differ structurally, and which would require repeated calls to estimate directly.

In GPT-5-mini, C↔D produced DFR between $0.9\%$ and $9.3\%$, with small $\Delta$F2 values (between $-0.042$ and $+0.028$) and 95% CIs crossing zero in all SLRs. Taken together, metadata, in the tested formulation, does not constitute a stable calibration mechanism: null effect in C↔D for Llama and modest and variable effect in GPT-5-mini.

### 4.3. RQ2: Does batch processing modify decisions compared to individual processing?

RQ2 examines whether batch processing, in isolation, changes LLM decisions. The comparison A↔C isolates this effect, since neither condition includes metadata, while B↔D verifies whether the pattern is maintained when metadata are present. In contrast to RQ1, batch processing was associated with substantially larger changes. In Llama, DFR in A↔C ranged from $5.6\%$ (rsl_1) to $26.7\%$ (rsl_4), with significant McNemar results ($p < 0.001$) in all reviews. The B↔D produced a very similar pattern, with DFR from $5.1\%$ to $26.3\%$, which is consistent with C↔D = 0 in this model. In GPT-5-mini, the effect was smaller, but still systematic, with DFRs between $1.3\%$ and $14.0\%$ in A↔C and significant McNemar results in four of the five SLRs.

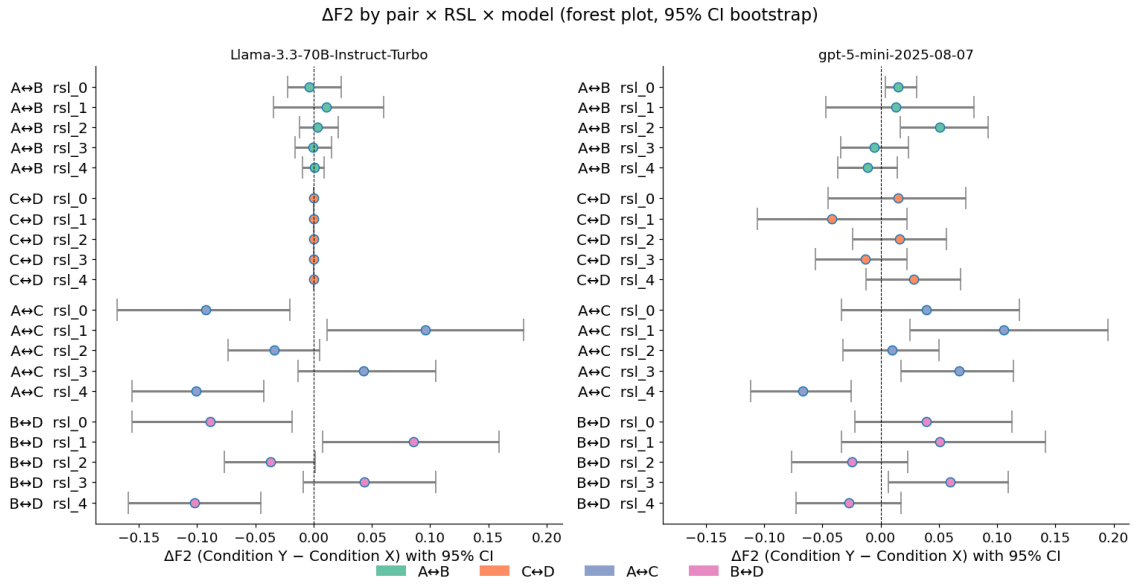



**Figure 4. ΔF2 by condition pair, SLR, and model (95% CI, *bootstrap*). Points to the right of the dashed line indicate F2 gain; points to the left indicate loss.**

The effect of batch processing on performance reveals a relevant contrast between the decision change and the aggregate gain. For Llama in A↔C, the sign and magnitude of ΔF2 vary between SLRs (Figure 4). There were gains in rsl_1 ($+0.096$; CI: $+0.012$, $+0.180$) and rsl_3 ($+0.043$), a value close to zero in rsl_2 ($-0.034$), and losses in rsl_0 ($-0.092$) and rsl_4 ($-0.101$). The case of rsl_1 is instructive: batch processing increased F2 from 0.37 to 0.46, but 101 studies changed from *Exclude* to *Include*, of which only 29 were correct; at the item level, 29 studies improved and 85 worsened. Similar or opposite patterns in rsl_3 and rsl_4 reinforce that interpreting ΔF2 in isolation may hide relevant redistributions of error at the article level.

These pieces of evidence suggest a practical recommendation. Across the five SLRs evaluated, the median ΔF2 of Llama in A↔C is close to zero ($-0.034$), and the bootstrap CIs cross zero in two of the five reviews, indicating that, in this set of reviews, batch processing did not show a uniform loss of aggregate performance. Considering that batch processing can substantially reduce inference costs [Pipal et al. 2026, Cheng et al. 2023], it may be an operational option for LLM-assisted screening. However, given the variability of the effect between reviews observed in these results, with ΔF2 ranging from $-0.101$ to $+0.096$ between SLRs, the adoption of batch processing in real pipelines should be accompanied by review-specific validation, ideally with an initial sample executed in individual mode as a reference for comparison.

### 4.4. RQ3: How does batch processing, with or without metadata, affect false negatives and direction of decision changes?

RQ3 investigates how batch processing, with or without metadata, affects false negatives and the net direction of decision changes. Since C↔D was already discussed in RQ1, the answer comes mainly from the comparison of B↔D. In Llama, B↔D reproduces A↔C almost exactly, because C↔D = 0 and A↔B is close to zero; therefore, the metadata in Condition B was insufficient to compensate for the batch effect. In GPT-5-mini, B↔D

presented DFR between $1.1\%$ and $11.0\%$. The sign of $\Delta$F2 remained the same as that of A↔C in three SLRs and was reversed in two. In terms of false negatives, reductions were observed in some combinations, such as rsl_1 and rsl_3 in both models, and increases in others, such as rsl_4 in both models. Thus, the combination of batch processing and metadata did not produce a consistent reduction in false negatives: the impact varies by review and by model, and in some cases coexists with an increased risk of losing relevant studies.

The analysis of *the net shift*, that is, the net direction of changes, in A↔C reveals a cross-cutting pattern. When SLRs are ordered by increasing prevalence (Table 2), the four reviews with prevalence $\leq 20\%$ present a positive *net shift*, between $+3.5$ and $+10.5$ percentage points. In these cases, batch processing made Llama more inclusive. In the single almost balanced SLR available (prevalence of $53\%$), *the net shift* was strongly negative (-25.3 points), an observed pattern rather than an established mechanism, given the reliance on a single review at this prevalence level. In GPT-5-mini, the same directional pattern appears in rsl_4, but there is no directional dominance.

**Table 2. Net shift of the batch effect without metadata (A↔C), by prevalence.**

| RSL | Prev. | Llama | GPT-5-mini |
|---|---|---|---|
| rsl_0 | 2.9% | +0.078 | -0.007 |
| rsl_1 | 3.5% | +0.043 | +0.006 |
| rsl_2 | 10.5% | +0.035 | -0.003 |
| rsl_3 | 17.4% | +0.105 | +0.060 |
| rsl_4 | 53.0% | -0.253 | -0.110 |

This pattern is compatible with a possible tendency toward implicit normalization of decisions within the batch. When processing 20 articles together, the model can approximate its decisions to an intermediate proportion of inclusions, although the experimental design does not allow direct identification of the causal mechanism. This behavior tends to benefit reviews with few true positives because it shifts more decisions from *Exclude* to *Include*, but it may penalize balanced reviews by excessively reducing the inclusion rate. Condition D, even with explicit metadata, did not correct this behavior in the balanced review in either model. The finding is compatible with implicit comparison effects in batch prompting [Pipal et al. 2026] and with sensitivity to positioning in long contexts [Liu et al. 2024].

### 4.5. RQ4: Are the observed effects consistent across models from different families?

RQ4 evaluates whether the effects are sustained across models of different families. Partial consistency is observed between the models. The qualitative pattern is repeated in both cases: the effect of metadata is small, the effect of batch processing is substantially larger, and the direction of changes depends on review prevalence. However, the magnitude of the effects differs. Llama presented DFRs in A↔C approximately twice as large as those of GPT-5-mini, with an average of $0.160$ versus $0.057$, in addition to the particular result of C↔D = 0. In absolute terms, GPT-5-mini showed more conservative behavior, with a predicted inclusion rate below the true prevalence in most SLRs with prevalence below 20%, except in the lowest-prevalence review (rsl_0) and, under batch processing, in rsl_3. This pattern limits *recall* and is reflected in an absolute F2 lower than Llama's

in these reviews. In rsl_2, for example, this operational difference was substantial: Llama in individual mode reached a recall of $0.925$ and F2 of $0.705$, while GPT-5-mini reached a recall of $0.369$ and F2 of $0.409$. Complete precision, recall, F2, FNR, and predicted inclusion rate values for all conditions and reviews are reported in Supplementary Material, Part II [Hida et al. 2026a].

This contrast corroborates the finding of Huotala et al. [Huotala et al. 2025] that variation between reviews may be greater than variation between models, but adds an operational layer. The choice of the most suitable model for assisted screening tends to depend on the profile of the target review. Even when the qualitative pattern converges across models, the magnitude of the differences may still be sufficient to affect practical processing-mode decisions.

## 5. Conclusion

This study investigated how screening metadata and batch processing influence LLM decisions in title and abstract selection. In a factorial design of $2\times2$ with five SESR-Eval reviews and two models, textual metadata had a limited effect, while batch processing produced larger changes.

Batch processing produced the largest effect, with substantially larger DFRs in the A$\leftrightarrow$C and B$\leftrightarrow$D comparisons. This effect varied according to the review and model profile: in more imbalanced reviews, batch processing tended to make Llama more inclusive, whereas in the single almost balanced review evaluated, it produced more restrictive behavior (an observed pattern, not an established mechanism). Furthermore, aggregate metrics and item-level analyses did not always converge, indicating that gains in the F2-score may coexist with worsening in individual decisions. Thus, batch processing should not be treated only as a cost-reduction strategy; its adoption in LLM-assisted screening should also consider possible changes in decision behavior.

From a methodological perspective, the results reinforce that LLM evaluation in screening should not depend only on global metrics. The combination of aggregate metrics, DFR, *net shift*, and paired article-level analysis proved necessary to characterize how decisions change between experimental conditions.

This study has limitations. The results were obtained from five Software Engineering reviews, with two models, a specific prompt formulation, and a single batch arrangement and execution per condition. In addition, the binary *inclusion/exclusion* decision simplifies real screening practice, which can involve uncertainty, discussion among reviewers, and additional contextual criteria.

In future work, we intend to expand the analysis to other datasets, domains (e.g., medical, social science), and model families, as well as to test different batch sizes, position effects within batches, and adaptive strategies that choose between individual and batch processing according to the review profile, complemented by cluster bootstrap resampling for uncertainty estimation and systematic instrumentation of token usage, monetary cost, and execution time.

## Acknowledgments

The authors thank Cesar School for the institutional and financial support provided through its internal research program, which supports the AIBL research group, to which

the first two authors are affiliated.

## Generative AI use statement

ChatGPT and Claude were used to support the linguistic revision, textual clarity, editorial refinement, and visual preparation of Figure 1. All scientific content, analyses, interpretations, and conclusions were prepared and reviewed by the authors.

## References


Bestgen, Y. (2022). Please, don't forget the difference and the confidence interval when seeking for the state-of-the-art status. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5956–5962.

Bornmann, L., Haunschild, R., and Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224.

Chai, K. E., Lines, R. L., Gucciardi, D. F., and Ng, L. (2021). Research screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic reviews*, 10(1):93.

Cheng, Z., Kasai, J., and Yu, T. (2023). Batch prompting: Efficient inference with large language model apis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 792–810.

Fagerberg, P., Sallander, O., Patil, K. V., Berg, A., Nyman, A., Borg, N., and Lindén, T. (2026). Batch size effects on mid-2025 state-of-the-art large language model performance in automated title and abstract screening. *Cochrane Evidence Synthesis and Methods*, 4(3):e70082.

Hida, G. S. and Do Nascimento, A. C. A. (2025). Overview of machine learning in class imbalance scenarios: Trends, challenges, and approaches. *Expert Systems with Applications*, page 129592.

Hida, G. S., Ribeiro, D. M., and Hida, C. S. (2026a). Research Artifacts: Class Imbalance and Batch Effects in LLM-Based Screening for Systematic Reviews. Zenodo record. DOI: 10.5281/zenodo.20414963.

Hida, G. S., Ribeiro, D. M., and Yahata, E. (2026b). Beyond accuracy: Llm variability in evidence screening for software engineering slrs. *arXiv preprint arXiv:2604.27006*.

Huotala, A., Kuutila, M., and Mäntylä, M. (2025). Sesr-eval: Dataset for evaluating llms in the title-abstract screening of systematic reviews. In *2025 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 01–12. IEEE.

Huotala, A., Kuutila, M., Ralph, P., and Mäntylä, M. (2024). The promise and challenges of using llms to accelerate the screening process of systematic reviews. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering*, pages 262–271.

Kitchenham, B. et al. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Kuutila, M., Mäntylä, M., Farooq, U., and Claes, M. (2020). Time pressure in software engineering: A systematic review. *Information and Software Technology*, 121:106257.

Laban, P., Murakhovs' ka, L., Xiong, C., and Wu, C.-S. (2023). Are you sure? challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint arXiv:2311.08596*.

Lewowski, T. and Madeyski, L. (2022). How far are we from reproducible research on code smell detection? a systematic literature review. *Information and Software Technology*, 144:106783.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Madeyski, L., Kitchenham, B., and Shepperd, M. (2025). Llm4screenlit: Recommendations on assessing the performance of large language models for screening literature in systematic reviews. *arXiv preprint arXiv:2511.12635*.

Meta AI (2024). Llama 3.3 70B Instruct. Model card. Hugging Face. Released on December 6, 2024. Accessed May 2026.

OpenAI (2025). GPT-5 mini (gpt-5-mini-2025-08-07). OpenAI API documentation. Snapshot: gpt-5-mini-2025-08-07. Accessed May 2026.

Pipal, C., Vogel, E.-M., Wack, M., and Esser, F. (2026). Researchers waste 80% of llm annotation costs by classifying one text at a time. *arXiv preprint arXiv:2604.03684*.

Rani, P., Blasi, A., Stulova, N., Panichella, S., Gorla, A., and Nierstrasz, O. (2023). A decade of code comment quality assessment: A systematic literature review. *Journal of Systems and Software*, 195:111515.

Sharbaf, M., Zamani, B., and Sunyé, G. (2023). Conflict management techniques for model merging: a systematic mapping review. *Software and Systems Modeling*, 22(3):1031–1079.

Syriani, E., David, I., and Kumar, G. (2024). Screening articles for systematic reviews with chatgpt. *Journal of Computer Languages*, 80:101287.

Tambon, F., Laberge, G., An, L., Nikanjam, A., Mindom, P. S. N., Pequignot, Y., Khomh, F., Antoniol, G., Merlo, E., and Laviolette, F. (2022). How to certify machine learning based safety-critical systems? a systematic literature review. *Automated Software Engineering*, 29(2):38.

Together AI (2024). Together AI inference api. API documentation. Accessed May 2026.

Vach, W. and Gerke, O. (2023). Gwet's ac1 is not a substitute for cohen's kappa–a comparison of basic properties. *MethodsX*, 10:102212.

Wang, S., Scells, H., Zhuang, S., Potthast, M., Koopman, B., and Zuccon, G. (2024). Zero-shot generative large language models for systematic review screening automation. In *European Conference on Information Retrieval*, pages 403–420. Springer.