Automatically Extracting Scientific Metrics with LLMs: A Case Study of ImageNet Papers

Abstract

In this paper, we introduce a large-scale dataset of papers annotated with their reported Top-1 accuracy on the ImageNet test set, and compare existing and new automatic metric extraction methods, along with a detailed qualitative error analysis. Our study highlights common reporting challenges—such as ambiguous dataset references, table-only metrics, and missing Top-1 values—that drive extraction errors. We curate and release a dataset of 200 manually annotated ImageNet classification papers, larger than prior work, and evaluate our pipeline against both existing approaches and ablated baselines.

1 Introduction

Scientific performance metrics – such as top-1 accuracy on ImageNet – play a central role in benchmarking model progress, establishing state-of-the-art (SOTA) claims, and guiding research trends [Bornmann et al., 2021, Barry et al., 2022]. However, metrics are often inconsistently reported across research papers: embedded in tables, mentioned only in captions or abstracts, expressed as error rates, or omitted entirely. As a result, extracting similar metrics at scale remains a largely manual and error-prone process.

In this paper, we present a case study in the automated extraction of top-1 accuracy for image classification reported on ImageNet papers. We construct and release a manually-annotated dataset of 200 ImageNet classification papers—substantially larger than SCILEAD [Şahinuç et al., 2024], which annotated 43 papers across various NLP tasks. By releasing our annotated dataset at `https://anonymous.4open.science/r/imagent-leaderboard-sample-25`, we aim to establish a reproducible benchmark for LLM-assisted scientific metric extraction and support broader efforts in automating literature understanding in machine learning research.

2 Annotated dataset

In this section, we describe the construction of our dataset, which is publicly accessible at `https://anonymous.4open.science/r/imagent-leaderboard-sample-25`.

2.1 Paper collection

Our dataset originates from an automated collection of publication entries through the PaperWithCode platform, focusing on computer vision papers that report results on the ImageNet dataset for the image classification task. We implemented a simple "try/catch" script to repeatedly query PaperWithCode API for PDFs or arXiv links, automatically skipping entries that lacked valid URLs or produced parsing errors. The step yielded an initial pool of candidate papers, all of which claimed to present top-1 accuracy on ImageNet or its widely recognized variants such as Tiny-ImageNet [Le and Yang, 2015]. We programmatically retrieved ImageNet image-classification papers using the `paper_dataset_list`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 LLM Evaluation Workshop.

endpoint of the PaperWithCode API, which returns results in its default (unspecified) order.[1] We first selected 12 papers to tune our prompts and then curated another 100 papers to build our development set. For the validation set, we retrieved all papers (focused on the Image Classification task and ImageNet) from PaperWithCodes and selected the top 100 entries, sorted by descending publication date, with preference given to those published in journals afterwards.

2.2 Label-verification protocol

From the successfully retrieved documents, we curated a corpus of papers on image classification using the ImageNet dataset. We performed a manual examination of the performance metric reported in each individual paper. Specifically, we identified references to "top-1 accuracy," and pinpointed the corresponding numerical values. We manually annotated each paper with explicit labels such as *(Dataset: ImageNet, Metric: Top-1 Accuracy)*.

2.3 Dataset statistics and alignment on ImageNet

Most papers in our corpus report Top-1 Accuracy on the ImageNet dataset. However, many of them were evaluated on *variants* or *subsets* of the ImageNet dataset, such as Tiny-ImageNet [Le and Yang, 2015], ImageNet-100 and etc. For instance, it is a commonplace for evaluations for ILSVRC-2012 or ILSVRC-2015, each of which differ slightly in the number of classes or distribution of images.

Table 2 summarizes the distribution of ground-truth top-1 accuracy presence in both the development and validation sets. In both subsets, only about one-quarter of the papers explicitly report a top-1 accuracy value. This skew reflects a common trend in the literature, where performance metrics—particularly for ImageNet—are often reported only on validation sets, or embedded within complex tables or figures, making automated extraction more challenging. In Fig. 3, we examine the reported metrics from each paper—often referencing *only* a validation subset or using a multi-crop evaluation strategy—and align them to a consistent schema. We attempt to unify various reporting practices; however, directly comparing reported performance metrics across the literature remains challenging.

2.4 Comparison with SCILEAD [Şahinuç et al., 2024]

While SCILEAD [Şahinuç et al., 2024] introduces a broad leaderboard spanning multiple NLP tasks and metrics, our dataset is uniquely focused on Image Classification on the ImageNet dataset. We curate a collection of 200 papers that explicitly report top-1 accuracy for the ImageNet classification task, with manually verified ground-truth annotations. We observe 26 papers with metric presence in the development set and 27 in the validation set (demonstrated in Table 2). Moreover, the presence or absence of metrics in our dataset is not artificially balanced or stratified. A detailed qualitative analysis of papers with and without reported metrics is provided in Section 3.2.

3 Experimental results and discussion

3.1 Experimental results

We benchmarked our pipelines—*VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY*, against the existing SCILEAD baseline in our full annotated validation and development set described in Section 2. Figure 1 presents results on the validation set, and Figure 2 shows results on the development set, reporting presence–correctness confusion matrices for the evaluated extraction systems. Full definitions, additional baselines, and regression metrics are provided in Appendix E.

3.2 Qualitative analysis

We illustrate our results with a qualitative analysis of easy and difficult cases, highlighting the challenges from the evaluated extraction systems. We extend our qualitative analysis in Appendix F.

[1] The API documentation does not state the sorting criterion; see `https://paperswithcode-client.readthedocs.io/en/latest/api/client.html`.

figs/100val_gpt_error_matrix.jpeg figs/100val_extract-and-verify_error_matrix.jpeg figs/100val_error_matrix.jpeg figs/100val_scilead_error_matrix.jpeg

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

Figure 1: Presence–correctness confusion matrices on the validation set in comparison with extraction quality across four methods: (a) *VOTE-ENSEMBLE (Ours*)*, (b) *EXTRACT-AND-VERIFY (Ours*)*, (c) *EXTRACT-only (Ours*)*, and (d) SCILEAD. Each matrix summarizes extraction outcomes conditioned on whether a Top-1 accuracy value is present in the ground-truth (*rows*) and whether the system produced an extraction (*columns*). Green cells indicate correct extractions, red cells indicate incorrect or hallucinated values, and blank cells indicate non-extraction when ground truth is absent.

figs/100dev_gpt_error_matrix.jpeg figs/100dev_extract-and-verify_error_matrix.jpeg figs/100dev_scilead_error_matrix.png

(a) GPT Vote Ensemble (Ours*) (b) Extract-and-Verify (Ours*) (c) SCILEAD

Figure 2: Error matrices comparing extraction quality across methods on the development set.

3.2.1 Straightforward ("easy") examples

- **Abstract, 1807.11626v3** (Fig. 5): "*On the ImageNet classification task, our **MnasNet** achieves 75.2% top-1 accuracy . . .*" *—A single numeric value appears in the abstract; the metric name and dataset are explicit.*

3.2.2 Difficult ("challenging") Examples

- **Alternate dataset variant** (Fig. 7): "*V-CapNet reaches 99.83% validation accuracy on the **Natural Images** dataset.*" *—Dataset differs from canonical ImageNet; the numeric value should be excluded from extraction.*
- **Table-only Top-1 value** (Fig. 10): ImageNet top-1 accuracy (76.82%) is given only in a table, with no reference in the main text. *—Requires table reading; plain text search may be error-prone.*
- **Multiple candidate Top-1 values** (Fig. 11): A table in 1909.13863v1 lists four *Case* rows with different top-1 accuracies (55.5–57.1%). *—Require an investigation on which value is the main result or flag ambiguity.*
- **Top-1 accuracy omitted, only Top-5 present** (Fig. 12): Sentence in 1807.10119v3 reports "top-5 accuracy drops slightly . . ." while never stating Top-1. *—Require returning "missing" rather than hallucinate a Top-1 value.*
- **Test vs. Validation ambiguity** (Fig. 17): "*Classification performance comparison on ImageNet (single crop, single model)... "* *—Mentions classification on ImageNet without specifying test/validation split, creating ambiguity on the split.*
- **Top-1 metrics embedded in large metric tables** (Fig. 9, Fig. 22–31, with additional examples in Fig. 32–35, detailed in Section F.1): The top-1 accuracy for ImageNet appears within tables that contain dozens of entries, dense formatting, and mixed dataset contexts. These settings present challenges for automated extraction systems, due to ambiguous column headers, multi-dataset rows, or inconsistent labeling. We highlight ten such cases:
  - **Ambiguous metric header** (Fig. 22): A table contains values such as 85.6 and 89.5 without specifying the evaluation split (validation/test) or whether the numbers correspond to top-1 accuracy. Systems returned mismatched values that failed to align with the ground truth (87.2).
  - **Cross-dataset overload** (Fig. 23): The table mixes results for ImageNet-1K, CIFAR-100, and TinyImageNet. The extracted value (82.0) does not correspond to the correct ImageNet Top-1 metric (85.9), highlighting difficulties in aligning rows with the intended benchmark.

—Extractors must jointly reason over model–metric–dataset alignment, disambiguate unlabeled columns, and avoid numeric heuristics. Without semantic parsing or table structure awareness, these large tables often lead to hallucinations or near-miss errors.

Key take-aways Easy cases share three traits: a single accuracy value, explicit dataset naming, and standard wording. Failures occur when authors (i) report on validation instead of test set, (ii) report on ImageNet variants or sampled subsets, (iii) place numbers only in supplementary material, or (iv) use alternate metrics such as error rates or Top-5 accuracy. These observations guided our rule-based post-processing (for error-to-accuracy conversion) and the heuristics that flag ambiguous dataset references.

3.2.3 Cross-system top-1 accuracy extraction comparison on selected failure cases

We analyze additional representative failure cases across systems, with extended examples provided below.

Top-1 metrics reported on validation set only (Fig. 21): A table presents classification and localization error rates (Top-1, Top-5) on ILSVRC-15 *validation set*. *—Top-1 accuracy must be inferred by subtracting from 100%; Top-1 scores on test set is never given.*

ImageNet was used in pretraining in the paper

- **Non-ImageNet dataset with pretrained model** (Fig. 20): "*Caltech-256: 84.7 (ImageNet-CLS), 76.7 (OpenImages)*" *—Although the model is pretrained on ImageNet, evaluation is done on Caltech-256; such results should not be extracted as ImageNet scores.*

Top-1 metrics embedded in large metric tables (Fig. 26, Fig. 27 and Fig. 28) Top-1 accuracy for ImageNet frequently appears within large, multi-column tables featuring dense formatting and mixed dataset benchmarks. The complex layouts pose challenges for automated extraction systems, particularly when headers are ambiguous or dataset-metric alignment is unclear. Below, we present three additional annotated examples.

- **Missing label context** (Fig. 26): Extractors failed to identify that 80.99 was the correct ImageNet Top-1 accuracy. Instead, they returned 77.85 due to ambiguity in model-type alignment and lack of direct sentence reference.
- **Column overload in architecture benchmarking** (Fig. 27): Dense layout with models, FLOPs, and multiple accuracy metrics makes correct alignment difficult. Ground truth (79.96) was not captured by any system.
- **Grouped ViT-family entries with sparse labels** (Fig. 28): The ground-truth Top-1 of 79.6 is lost among model configurations. Systems instead hallucinated higher values (e.g., 82.0, 84.2) drawn from unrelated rows.

Paper	Ground-truth	VOTE-ENSEMBLE	EXTRACT-AND-VERIFY	SCILEAD	EXTRACT-only
`1610.02391v4.pdf` (Fig. 21)	69.62	–	70.58	70.58	–
`2505.14124v1.pdf` (Fig. 26)	80.99	–	77.85 x	–	77.85 x
`2412.15077v1.pdf` (Fig. 27)	79.96	–	–	–	68.28 x
`2411.15241v1.pdf` (Fig. 28)	79.6	84.22 x	–	81.9 x	82.0 x

Table 1: Cross-system Top-1 Accuracy extraction comparison on selected failure cases. Each system either outputs an incorrect value (denoted by x) or abstains ("–"). These examples are highlighted in the qualitative analysis (Sec. 3.2).

4 Conclusion

We presented a case study of extracting scientific metrics from the scientific literature using a historically significant metric: top-1 accuracy reported on the ImageNet test set. We release a large annotated dataset (substantially larger than previous annotated datasets of 43 papers) and present detailed qualitative error analysis.

We experimented with prior work as well as several methods based on recent ideas such as self-critique. Expanding the labelled dataset would allow for enough statistical power to compare different methods.

References

Erin S Barry, Jerusalem Merkebu, and Lara Varpio. Understanding state-of-the-art literature reviews. *Journal of graduate medical education*, 14(6):659–662, 2022.

Lutz Bornmann, Robin Haunschild, and Ruediger Mutz. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases, 2021. URL https://arxiv.org/abs/2012.07675.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug, 2023. URL https://arxiv.org/abs/2304.05128.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, and Iryna Gurevych. Efficient performance tracking: Leveraging large language models for automated construction of scientific leaderboards. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7963–7977, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.453. URL https://aclanthology.org/2024.emnlp-main.453/.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks, 2024. URL https://arxiv.org/abs/2402.08115.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification, 2023. URL https://arxiv.org/abs/2212.09561.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

img/validation_set_GT_adjustment.png

Figure 3: ground-truth table for a selected set of 12+ papers, recording Top-1/Top-5 values from the test set, validation set, or a specific multi-crop procedure (e.g., 10-crop validation from Krizhevsky et al., 2017). "–" and space denotes that no explicit metric was identified for that field.

A Related work

Our work builds on and integrates a sequence of ideas, including diverse prompting techniques and the self-critique paradigm. Prior work such as SCILEAD [Şahinuç et al., 2024] has advanced the construction of scientific leaderboards through an automated extraction pipeline. Our contribution lies in applying these techniques—beginning with structured prompts incorporating verification and ensemble ideas—for an automated extraction of scientific metrics from research papers.

In few-shot prompting, also referred to as "in-context learning," [Brown et al., 2020], an LLM is provided with a prompt consisting of multiple examples of the target task, each in the form of input-output pairs.

Query self-refinement is argued to enhance the initial outputs of large language models (LLMs) [Madaan et al., 2023]. Self-refinement is inspired by the way humans revise their written text. This method consists of an iteratively feedback-driven process that improves the initial responses generated by LLMs.

The Least-to-Most prompting framework was introduced by Zhou et al. [2022], which decomposes reasoning tasks into structured subproblems to improve performance. Each subproblem is solved in sequence with subsequent steps conditioned on previous answers. The least-to-most prompting enables LLMs to generalize for problems with more difficulties than those present in the prompts.

SCILEAD introduced an LLM-based method for the automatic constructions of scientific leaderboards [Şahinuç et al., 2024]. SCILEAD contributes a manually curated dataset of leaderboards drawn from 43 scientific papers and proposes an extraction schema based on task–dataset–metric (TDM) triples, where each triple represents an extraction task, the associated dataset, and the reported evaluation metric. The work exhaustively annotates individual papers by labeling all unique TDM combinations along with their respective top-reported results.

Several recent studies have explored the concept of direct self-critique by LLMs [Stechly et al., 2024, Weng et al., 2023, Chen et al., 2023]. Inspired by human cognitive processes, these approaches leverage the intuition that verifying or critiquing an answer is typically easier or fundamentally different from generating it from scratch. Hence, the strategies demonstrate potential to improve the overall quality of outputs [Stechly et al., 2024]. In a standard self-critique pipeline, an LLM first generates an answer and subsequently receives its own response as input, along with explicit prompting instructions to critique, refine, or revise the original answer. The self-critique loop iterates and then stops at a predefined stopping criterion.

Our work builds upon these ideas by incorporating explicit verification steps, enabling cross-checking and refinement of extracted performance metrics.

Metric Presence	Development Set	Validation Set
Ground-truth Present	26	27
Ground-truth Absent	74	73
Total	100	100

Table 2: Distribution of ground-truth Top-1 Accuracy presence across the 100-paper development and validation sets. Many papers do not explicitly report Top-1 Accuracy, often deferring such results to supplemental materials or validation splits.

B Extraction experiments

We describe our extraction experiments for *EXTRACT-AND-VERIFY* in detail. We first introduce the *VOTE-ENSEMBLE* style of prompting for aggregating predictions across paper sections, and then describe how *EXTRACT-AND-VERIFY* builds upon it with an additional verification step.

Let n denote the number of text segments derived from each paper (e.g., abstract, results, conclusion), and k the number of prompt attempts per section for diverse extraction outputs. The value of n is computed dynamically for each paper using a lightweight chunking function, which groups every three consecutive pages into a single segment. We set $k = 5$ in all experiments.

***VOTE-ENSEMBLE* prompting** We begin by dividing each paper into n segments—typically corresponding to sections such as the abstract, experimental results, and conclusion. For each segment, we apply an extraction prompt that instructs the LLM to extract a top-1 accuracy reported on ImageNet and its corresponding source sentence from literature. This extraction process is repeated k times per segment to capture various responses. The resulting k sentence–accuracy pairs for i-th segment are then aggregated using our *VOTE-ENSEMBLE* strategy, which selects the pair most frequently occurring as the final output. The prompt used in this step is shown in Section C.1.

***EXTRACT-AND-VERIFY* prompting** We introduced a verification phase during the extraction process of our *EXTRACT-AND-VERIFY*. For each section, the LLM re-evaluates each extracted sentence–accuracy pair in the context of the original PDF page to verify whether the sentence appears in the text and whether the extracted value explicitly corresponds to a top-1 accuracy on ImageNet. The verification is applied across the k extractions generated during the ensemble step. The prompt template used for verification is demonstrated in Section C.2.

To isolate the effect of ensembling and verification, we introduce *EXTRACT-only*—a minimal baseline where extraction is performed with a single prompt iteration ($k = 1$) based on VOTE-ENSEMBLE. It enables the independent assessment of each system component.

C Prompt examples

C.1 *VOTE-ENSEMBLE prompt demonstration*

Prompt Input

```
Find the accuracy value associated with most common sentences from the list of
sentences and accuracies. Only output the accuracy value.

Example 1:
'Sentence: "ImageNet1K Top-1 Accuracy ViT 88.5 89.1 88.6 92.4": 92.4', 'Sentence:
"SSv2 Top-1 Accuracy ViViT 65.4 68.6 80.1 85.4 ImageNet1K Top-1 Accuracy ViT 88.5 89.1
88.6 92.4 Sun RGBD Top-1 Accuracy Simple3D-former 57.3 62.4 71.4 74.6"Accuracy: 92.4',
'Sentence: "ImageNet1K Top-1 Accuracy ViT 88.5 89.1 88.6 92.4": 92.4', 'Sentence:
"SSv2 Top-1 Accuracy ViViT 65.4 68.6 80.1 85.4 ImageNet1K Top-1 Accuracy ViT 88.5
89.1 88.6 92.4 Sun RGBD Top-1 Accuracy Simple3D-former 57.3 62.4 71.4 74.6": 74.6',
Expected Output: 92.4

Example 2:
'Sentence: "SSv2 Top-1 Accuracy ViViT 65.4 68.6 80.1 85.4 ImageNet1K Top-1 Accuracy
ViT 88.5 89.1 88.6 92.4 Sun RGBD Top-1 Accuracy Simple3D-former 57.3 62.4 71.4
74.6": 74.6', 'Sentence: "ImageNet1K Top-1 Accuracy ViT 88.5 89.1 88.6 82.4": 82.4',
'Sentence: "SSv2 Top-1 Accuracy ViViT 65.4 68.6 80.1 85.4 ImageNet1K Top-1 Accuracy
```

ViT 88.5 89.1 88.6 82.4 Sun RGBD Top-1 Accuracy Simple3D-former 57.3 62.4 71.4 74.6": 82.4', Expected Output: 82.4

Example 3:
'Sentence: "It's not mentioned top-1 accuracy on ImageNet": NA', 'Sentence: "-": NA', 'Sentence: "It's not mentioned top-1 accuracy on ImageNet": NA', 'Sentence: "-": NA', Expected Output: NA

Example 4:
'Sentence: "It's not mentioned.": NA', 'Sentence: "-": NA', 'Sentence: "It's mentioned top-5 accuracy on ImageNet": NA, 'Sentence: "Cocoa 23.3 21.2": 23.3', Expected Output: NA

Now extract the accuracy value associated with most common sentences, sentences$_a nd_a ccuracies$

Expected Output:

C.2 *EXTRACT-AND-VERIFY* prompt demonstration

Prompt Input

INPUTS
• page: full text of one PDF page, page
• pair: "Sentence:<sentence> : <accuracy>", a_sentence_and_accuracy

TASK
Check whether the sentence appears in the page (with minor formatting variations), and whether the accuracy is a valid Top-1 ImageNet value.

- If both sentence and accuracy are correct and found in the page, return the pair unchanged. - If the accuracy is incorrect but a correct one exists, return the corrected pair. - If no Top-1 ImageNet accuracy is found, return an empty string.

IMPORTANT
• Only match Top-1 accuracy on ImageNet (not Top-5, CIFAR, COCO, etc.)
• Output exactly one line, or nothing at all.

OUTPUT
Sentence:<sentence> : <accuracy>

D Error analysis metrics summarization

D.1 Regression metrics definitions

We present the regression metrics definitions in Table 3.

D.2 Presence–correctness matrix analysis

D.2.1 12-paper set

Figure 4 presents the error analysis in the form of confusion matrices along two axes:

1. whether the paper reports a Top-1 Accuracy value on ImageNet (*row*)
2. whether the system extracts such a value (*column*)

The counts are computed over the 12 manually inspected papers listed in Table 7.

Table 3: Evaluation metrics and error categorization for extraction tasks, with matrix axis definitions and term abbreviations.

Category	Metric Description
Presence–Correctness Matrix Axes	
Row: Ground-truth Presence	Whether the paper reports a Top-1 Accuracy value on ImageNet. Two possible states: **Ground-truth Present** – the metric is reported in the pape. **Ground-truth Absent** – the metric is not reported.
Column: System Extraction	Whether the system extracts a Top-1 Accuracy value. Three possible outputs: **Correctly Extracted/Extract Match(EM)** – value is correctly extracted and matches the ground-truth. **Incorrectly Extracted** – value is extracted but incorrect (e.g., hallucinated or wrong metric). **Extracted Absent** – system does not extract any value.
Classification-style Metrics	
#Extract Match (EM)	Exact Match (EM) is defined as a correct extraction in which the extracted top-1 accuracy value exactly equals the ground-truth value recorded in the paper. EMs are only counted when the ground-truth value is present, and the extracted value aligns numerically with it (e.g., 76.82 matches 76.82 exactly).
#Incorrect Extractions	the number of incorrect extractions when ground-truth is present.
Regression-style Metrics	
	Let $\hat{y}_i$ denotes extracted values, y_i denotes ground-truth values and n denotes the total number of papers where both ground-truth and extractions are present:
MAE	**Mean Absolute Error** between extracted and ground-truth numeric values: $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$.
RMSE	**Root Mean Square Error**: $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$.

figs/12dev_gpt_error_matrix.png figs/12dev_extract-and-verify_error_matrix.png figs/12dev_scilead_error_matrix.png

(a) GPT Vote Ensemble (Ours*) (b) Extract-and-Verify (Ours*) (c) SCILEAD

Figure 4: Error matrices comparing extraction quality across methods.

E Full experiment results

E.1 Runtime discussion & API rate limits

We report the runtime and cost to process one PDF on a single 8-core CPU with a 100 Mbps network link to the OpenAI endpoint. The user tier 1 imposes `90 000` tokens min^{-1} and `3 500` requests min^{-1}. Our pipeline uses a single `gpt-4o` call per sections (consisting of a few pages), averaging ~1 650 input tokens and 120 output tokens. We apply a dynamic safety delay time based on the characteristics of each paper between calls of OpenAI requests. As a result, the average processing time per document is approximately 15 minutes—making the pipeline more computationally intensive than initially expected. We evaluate the performance of our proposed extraction pipelines, *VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY*, against the existing SCILEAD baseline in our full annotated validation and development set described in Section 2. We then report on regression-based metrics—quantifying how close the extracted values are numerically. The details of metric definitions can be found in Section D.

System	#Extract Match ↑	#Incorrect Extractions ↓	MAE ↓	RMSE ↓
VOTE-ENSEMBLE (Ours)*	3	12	19.633	35.672
EXTRACT-AND-VERIFY (Ours)*	6	8	3.033	4.500
EXTRACT-only (Ours)*	4	14	**2.831**	**4.103**
SCILEAD	**7**	8	5.026	11.027

Table 4: Regression metrics comparing *VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY*, *EXTRACT-only*, and SCILEAD on the 100-paper validation set. *EXTRACT-only* represents a minimal baseline where extraction is performed using a single section without voting.

E.2 100 Validation set with *EXTRACT-only* baseline

We compare our proposed methods, *VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY*, *EXTRACT-only* against SCILEAD on our 100-paper validation set. Figure 1 shows the presence–correctness confusion matrices for four extraction systems evaluated in the validation set.

Among the four systems, SCILEAD achieved the highest number of exact matches (7), followed by *EXTRACT-AND-VERIFY* (6), *EXTRACT-only* (4), and *VOTE-ENSEMBLE* (3). Notably, *EXTRACT-AND-VERIFY* exhibited lower regression errors compared to SCILEAD and *VOTE-ENSEMBLE*, despite recovering slightly fewer exact matches. As reported in Table 4, both *EXTRACT-AND-VERIFY* and *EXTRACT-only* achieved the lowest Mean Absolute Error (MAE) of 3.211 and RMSE of 4.550. In contrast, *VOTE-ENSEMBLE* reported high numeric errors among its 12 incorrect extractions, yielding an MAE of 19.633 and RMSE of 35.672. SCILEAD showed moderate regression performance (MAE: 5.026, RMSE: 11.027) but had a higher count of incorrect extractions (8). While these results suggest performance differences across systems, we caution that the observed variations may not be statistically significant due to the limited size of the validation set. To better understand the impact of prompt aggregation and verification, we include an additional *EXTRACT-only* baseline with details in Section E.4.

E.3 Development set results

We assess our proposed methods, *VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY* against SCILEAD on our 100-paper development set. Figure 2 presents presence–correctness confusion matrices comparing *VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY*, and SCILEAD on our development set. To quantify performance, we compute regression metrics over incorrect extractions—cases where both the ground-truth and predictions are present, but the extracted values are incorrect. Table 5 reports the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for these incorrect extractions.

Both of our methods, *VOTE-ENSEMBLE* and *EXTRACT-AND-VERIFY*, correctly extracted 19 top-1 accuracy values on ImageNet (73.1%), outperforming SCILEAD, which achieved 4 exact matches [2] (15.4%). While both of our methods reported 5 incorrect extractions, *EXTRACT-AND-VERIFY* showed higher regression error in terms of regression errors among them. As shown in Table 5, *EXTRACT-AND-VERIFY* achieved a higher Mean Absolute Error (MAE) of 1.632 and RMSE of 6.029, compared to 0.642 MAE and 2.112 RMSE for *VOTE-ENSEMBLE*. SCILEAD presented higher error (MAE: 14.482, RMSE: 28.800). The results are based on a limited development set and may not be statistically significant.

E.4 *EXTRACT-only* baseline result

To better understand the impact of prompt aggregation and verification, we include an additional *EXTRACT-only* baseline. The method corresponds to setting $k = 1$ in *VOTE-ENSEMBLE*. While *EXTRACT-only* underperformed SCILEAD in terms of exact match (4 vs. 7), it achieved lower regression error, with an MAE of 2.831 and RMSE of 4.103. Again, we reiterate that our findings are based on a small sample and suggest they be interpreted with caution.

F Extensions to Qualitative Analysis

F.1 More Cross-System Top-1 Extraction Comparison on Selected Failure Cases

We analyze additional representative failure cases across systems, with extended examples provided below.

[2] Exact Match (EM) refers to a correct extraction where the extracted value exactly matches the ground-truth value present in the paper.

System	#Extract Match ↑	#Incorrect Extractions ↓	MAE ↓	RMSE ↓
VOTE-ENSEMBLE (Ours)*	19	5	0.642	2.112
EXTRACT-AND-VERIFY (Ours)*	19	5	1.632	6.029
SCILEAD	4	8	14.482	28.800

Table 5: Regression metrics comparing *VOTE-ENSEMBLE*, *EXTRACT-AND-VERIFY* and SCILEAD on development set.

Top-1 metrics reported on validation set only (Fig. 15, Fig. 16, Fig. 21):

- **Top-1 scores on validation set only** (Fig. 15): "*We acquire better classification results on complex validation set . . . "* *—Top-1 scores on validation reported; Top-1 scores on test set is never given.*
- **Top-1 scores on validation set only** (Fig. 16): "*. . . top-1 and top-5 error rates on the ImageNet* ***validation set*** *. . . "* with Top-1 = 22.15% for DenseNet-264. *—Top-1 must be derived from the error rate (100% - 22.15%), and validation set must be interpreted correctly.*
- **Top-1 scores on validation set only** (Fig. 21): A table presents classification and localization error rates (Top-1, Top-5) on ILSVRC-15 *validation set*. *—Top-1 must be inferred by subtracting from 100%; Top-1 scores on test set is never given.*

ImageNet was pretrained in the paper (Fig. 18, Fig. 19, Fig. 20): *—Requires filtering out such examples as false positives despite mentioning ImageNet.*

- **Non-ImageNet dataset with pretrained model** (Fig. 18): "*We obtained an accuracy of 91.66% and 78.01% for the CALTECH 101 and neuromorphic CALTECH 101 datasets respectively.*" *—Reports accuracy while using ImageNet-pretrained networks, but on non-ImageNet datasets (CALTECH 101); must be excluded.*
- **Table-only accuracy, but not ImageNet** (Fig. 19): Table lists results like "85.3" under DA–ADAGE Incremental for MNIST-M and SVHN. *—Metrics shown in table format but pertain to other domains (MNIST-M, SVHN); dataset mismatch with ImageNet.*
- **ImageNet-pretrained model on different datasets** (Fig. 20): "*Caltech-256: 84.7 (ImageNet-CLS), 76.7 (OpenImages)*" *—Although the model is pretrained on ImageNet, evaluation is done on Caltech-256; such results should not be extracted as ImageNet scores.*

Top-1 metrics embedded in large metric tables (continued) (Fig. 9, Fig. 22–31, and Fig. 32–35): Top-1 accuracy for ImageNet frequently appears within large, multi-column tables featuring dense formatting and mixed dataset benchmarks. The complex layouts pose challenges for automated extraction systems, particularly when headers are ambiguous or dataset-metric alignment is unclear. Below, we present four additional annotated examples.

- **Misaligned extraction due to unlabelled columns** (Fig. 25): Although the table includes Top-1 classification accuracy (81.02), extractors reported nearby but incorrect values (e.g., 76.1 or 79.5), likely due to metric misinterpretation.
- **Under-specified table with missing axis labels** (Fig. 30): The ImageNet Top-1 accuracy (76.2) is never directly labeled. Extractors returned 72.2, reflecting structural ambiguity in the source format.
- **Multiple candidate metrics in same table** (Fig. 31): The Top-1 accuracy (78.1) is buried among other results. Extractors selected 70.0, suggesting overreliance on numerical proximity rather than structured alignment.
- (Fig. 32): "*Comparison of Top-1 accuracy across various methods on the ImageNet dataset...*"

  Top-1 scores are shown across architectures like RN50 and ViT-B/16.
- (Fig. 33): "*ImageNet: 74.1*" appears in a large benchmark spanning multiple datasets.

  Despite being explicitly labeled, the metric is embedded among heterogeneous datasets, making it unclear. Systems may incorrectly associate the wrong metric with the wrong dataset.
- (Fig. 34): "*Image classification results (Acc, %) on ImageNet.*"

  The table mixes logit-based and feature-based methods across two settings. Extractors must infer correct Top-1 values from rows and columns with inconsistent grouping and abbreviations. Without semantic understanding of headers, incorrect matches frequently occur.
- (Fig. 35): "*ImageNet: 62.78*" appears in a row alongside CUB200, EuroSAT, and other datasets.

Paper	Ground-truth	VOTE-ENSEMBLE	EXTRACT-AND-VERIFY	SCILEAD	EXTRACT-only
2501.10640v2.pdf (Fig. 22)	87.2	85.6 x	–	83.9 x	89.5 x
2501.07783v1.pdf (Fig. 23)	85.9	82.0 x	82.1 x	–	82.0 x
2505.14062v1.pdf (Fig. 24)	83.0	–	–	67.5 x	–
2410.08407v2.pdf (Fig. 25)	81.02	–	76.1 x	79.51 x	76.1 x
2412.02366v3.pdf (Fig. 29)	78.73	73.3 x	–	65.8 x	77.23 x
2410.10773v1.pdf (Fig. 31)	78.1	–	70.0 x	–	–
2504.08710v1.pdf (Fig. 30)	76.2	–	72.2 x	–	–
2412.20110v3.pdf (Fig. 32)	74.23	82.0 x	66.17 x	36.88 x	66.17 x
2503.12206v2.pdf (Fig. 33)	74.1	–	83.44 x	–	–
2412.08139v1.pdf (Fig. 34)	73.69	82.2 x	–	72.49 x	71.35 x
2412.11917v3.pdf (Fig. 35)	62.78	71.89 x	–	–	63.31 x

Table 6: Cross-system Top-1 Accuracy extraction comparison on selected failure cases. Each system either outputs an incorrect value (denoted by x) or abstains ("–"). These examples are highlighted in the qualitative analysis (Sec. 3.2 and Sec. F.1).

G Limitations

While introducing a larger annotated dataset than prior work, several limitations remain in our study. First, despite being larger than previous work, our dataset may still lack sufficient statistical power to demonstrate significant performance differences across various extraction systems. Moreover, our use of the PapersWithCode API for sampling—given its default ordering—may introduce sampling bias.

While ImageNet has historically served as a foundational benchmark dataset, it may not fully capture the diversity of datasets or evaluation metrics in computer vision or the broader computer science community. Therefore, our findings may not generalize to other tasks, domains, or scientific fields with different reporting conventions and metric structures.

H Ethics statement

Our work introduces a large-scale dataset of papers annotated with their reported top-1 accuracy on the ImageNet test set. All PDF papers in our study are publicly available on arXiv, which permits fair use and supports responsible, reproducible, and transparent scientific research practices. All annotation work was performed by the authors.

Our released dataset contains entries consisting of arXiv identifiers and their corresponding labeled Top-1 accuracy values on ImageNet. The intended use of our released dataset is strictly for academic research and analysis. It is not designed for, nor licensed to support, commercial or production use, in accordance with the original access conditions of the data sources.

File Name	GT Top-1 Accuracy	*VOTE-ENSEMBLE* (Ours*)	Extract Verify (Ours*)	SCILEAD
1909.13863v1.pdf	57.1	57.1	57.1	**NA**
1807.11164v1.pdf	77.2	77.2	**75.4**	**18.56**
omniVec_2023.pdf	92.4	92.4	92.4	**NA**
1803.00942v3.pdf	NA	NA	NA	NA
1703.09844v5.pdf	75	75	75	NA
1807.11626v3.pdf	75.2	**66**	75.2	**76.7**
1807.11459v1.pdf	NA	NA	NA	NA
1807.11254v2.pdf	77.86	**NA**	**NA**	**-0.82**
1807.10108v5.pdf	NA	NA	NA	NA
1512.03385v1.pdf	NA	NA	NA	NA
1909.11155v1.pdf	76.82	76.82	76.82	76.82
1807.10119v3.pdf	NA	NA	NA	NA

Table 7: Ground-truth versus extractor outputs on the 12-paper set. Cells in **red bold** indicate a disagreement with the ground-truth.

I Screenshots from ArXiX Papers with Extract Sections Highlighted

img/1807.11626v3.pdf_easy_abstract.png

Figure 5: **Easy example.** Top-1 accuracy (75.2 %) is stated plainly in the abstract of *1807.11626v3.pdf*.

img/1703.09844v5.pdf_easy.png

Figure 6: **Easy example.** A single sentence in the main text of *1703.09844v5.pdf* reports top-1 accuracy (75 %).

img/1807.10108v5.pdf_challeng_ex.png

Figure 7: **Challenging example.** Top-1 accuracy is reported only on a validation split of an ImageNet variant in *1807.10108v5.pdf*.

img/1512.03385v1.pdf_chanlleng.png

Figure 8: **Challenging example.** The original ResNet paper (*1512.03385.pdf*) only reported on ImageNet validation error rates.

img/omnivec_error.png

Figure 9: **Challenging example.** In *omniVec_2023.pdf*, the ImageNet Top-1 value (92.4 %) appears as one cell in a table containing multiple datasets.

img/1909.11155v1.pdf_error.png

Figure 10: **Challenging example.** *1909.11155v1.pdf* gives top-1 accuracy only within a table; the metric is absent from the surrounding text.

img/1909.13863v1_challenge.png

Figure 11: **Challenging example.** *1909.13863v1.pdf* gives top-1 accuracy only within a table.

img/1807.10119v3_challenging.png

Figure 12: **Challenging example.** *1807.10119v3.pdf* omits top-1 accuracy, reporting only Top-5 (80 %).

figs/1412.6598v2_challenge.png

Figure 13: **Challenging example.** *1412.6598v2.pdf* reports multiple references to ImageNet and performance, but no clear top-1 accuracy value.

figs/1502.03167v3_challenge_top5gt_wrong_number_reported.png

Figure 14: **Challenging example.** *1502.03167v3.pdf* reports only Top-5 validation error (4.9%); no Top-1 value.

figs/1506.04701v3.pdf_gt_on_valset.png

Figure 15: **Challenging example.** *1506.04701v3.pdf* with validation results; top-1 accuracy on test set not stated.

figs/1608.06993v5.pdf_gt_on_valset.png

Figure 16: **Challenging example.** *1608.06993v5.pdf* where Top-1 error rate on ImageNet validation set (e.g., 22.15%) needs conversion to accuracy and split is ambiguous.

figs/1610.02357v3.pdf_gt_on_valset.png

Figure 17: **Challenging example.** *1610.02357v3.pdf* mentions top-1 accuracy to ImagetNet but split is ambiguous.

img/1807.00578v1.pdf_wrong_dataset_imagenet_pretrained.png

Figure 18: Accuracy reported (91.66% and 78.01%) is for CALTECH datasets, not ImageNet.

img/1902.04992v1.pdf_wrong_dataset_imagenet_pretrained.png

Figure 19: Accuracy values (e.g., 85.3) are shown in table format, but target domains are not ImageNet.

img/1904.05871v1.pdf_wrong_dataset_imagenet_pretained.png

Figure 20: Pre-trained models are ImageNet-based, but classification is done on other datasets (e.g., Caltech-256).

img/1610.02391v4.pdf_gt_on_valset.png

Figure 21: Classification and localization error rates (%) on ILSVRC-15 validation set from 1610.02391v4.pdf. The table reports Top-1 classification error for models like VGG-16 and AlexNet. Top-1 metrics on test set is not stated.

img/2501.10640v2.pdf_large_metric_table.png

Figure 22: Metric table from 2501.10640v2.pdf. No explicit Top-1 label or split is provided. Extracted value (85.6) does not match the ground truth (87.2).

img/2501.07783v1.pdf_large_metric_table.png

Figure 23: Large benchmark comparison in 2501.07783v1.pdf with top-1 accuracy buried among multiple datasets. The extracted value of 82.0 does not match the ground truth (85.9).

img/2505.14062v1.pdf_large_metrics_table.png

Figure 24: Ambiguous accuracy reporting in 2505.14062v1.pdf. Top-1 accuracy for ImageNet is co-listed with CIFAR/Tiny-ImageNet rows. Systems failed to extract a valid value.

img/2410.08407v2.pdf_large_metrics_table.png

Figure 25: In 2410.08407v2.pdf, Top-1 accuracy (81.02) appears in a table with closely related numbers. Extractors returned 76.1 or 79.51, misaligned with the correct value.

img/2505.14124v1.pdf_large_metrics_table.png

Figure 26: Example from 2505.14124v1.pdf where 80.99 is reported, but the Top-1 metric is embedded among unlabeled entries. Systems extracted 77.85, a nearby but incorrect value.

img/2412.15077v1_large_metric_table.png

Figure 27: Figure from 2412.15077v1.pdf, where top-1 accuracy (79.96) appears in a multi-column architecture table. No extractor returned the correct value.

img/2411.15241v1_large_metric_table.png

Figure 28: In 2411.15241v1.pdf, 79.6 is reported in a dense comparison table with ViT variants. Systems hallucinated or extracted higher values (e.g., 82.0, 84.2).

img/2412.02366v3_large_metric_table.png

Figure 29: Metric table from 2412.02366v3.pdf, where 78.73 is buried among comparisons across variants. Extractors returned mismatched outputs like 73.3 or 65.8.

img/2504.08710v1_large_metric_table.png

Figure 30: Top-1 accuracy of 76.2 from 2504.08710v1.pdf is presented in a large table with no metric labels. Extracted values (e.g., 72.2) reflect misalignment.

img/2410.10773v1_multi_candidate_metric.png

Figure 31: In 2410.10773v1.pdf, top-1 accuracy (78.1) appears with multiple candidate rows and no clear indicator. Systems returned incorrect values such as 70.0.

img/2412.20110v3_large_metrics_table.png

Figure 32: Large table showing top-1 accuracies across various architectures on ImageNet, but split (validation/test) is not explicitly stated.

img/2503.12206v2.pdf_large_metrics_table.png

Figure 33: Top-1 accuracy of 74.1 appears in a multi-dataset benchmark.

img/2412.08139v1.pdf_large_metrics_table.png

Figure 34: Table 4 compares classification accuracy (%) across methods and settings on ImageNet.

img/2412.11917v3.pdf_large_metric_table.png

Figure 35: Top-1 accuracy reported on ImageNet appears in a wide comparison table (e.g., 62.78).