Lost-and-Found Entities: A Benchmark for Long-List Entity Extraction

Serhii Shchoholiev[1]* Anton Fedoruk[1]†

[1]Kay.ai, Brooklyn, NY, USA

October 17, 2025

Abstract

Existing datasets for evaluating document extraction using large language models (LLMs) predominantly focus on key-value pair extraction and fail to address the complexities of long list entity extraction. However, real-world business documents such as invoices, insurance claims, purchase orders, and financial statements commonly organize data in table-like structures with repetitive entities of the same type—a critical gap in current benchmarking efforts. We introduce a comprehensive dataset specifically designed to evaluate long list extraction performance, incorporating common challenges observed in production environments, including true duplicate entities, multi-row entries, and various structural inconsistencies. Our dataset construction methodology leverages a corpus of real-world insurance claims gathered through Kay.ai operations, combined with systematically identified problematic patterns in table-like documents. We employ LLMs to generate realistic document layouts, which are subsequently rendered into PDFs and processed through optical character recognition (OCR) to simulate authentic extraction scenarios complete with real-world noise. Additionally, we provide standardized evaluation scripts to facilitate reproducible assessments. We benchmark flagship models from OpenAI, Anthropic, and Google using practical, easily implementable techniques including zero-shot prompting and accumulative generation. Our work addresses a significant gap in document understanding evaluation and provides the research community with essential tools for advancing long list extraction capabilities.

1 Introduction

The introduction should provide context for your research, explain the problem you're addressing, and outline the contributions of your work. Recent studies have explored various approaches to this problem [1], [2].

*Corresponding author: serhii@kay.ai

†anton@kay.ai

1.1 Background and Motivation

Describe the background of the problem and why it's important to solve. As noted by Smith [3], this area has seen significant development in recent years. Online resources [4] provide additional context for understanding the scope of this challenge.

1.2 Research Questions

State your main research questions or hypotheses.

1.3 Contributions

Clearly list the main contributions of this work:

- First contribution
- Second contribution
- Third contribution

1.4 Paper Organization

The rest of this paper is organized as follows: Section 3 concludes the paper and discusses future work.

2 Related Work

3 Conclusion

Summarize your work, restate the main contributions, and suggest directions for future research.

3.1 Summary

Briefly summarize the key points of your paper.

3.2 Future Work

Suggest directions for future research.

Acknowledgments

This work was supported by [funding source]. We thank [people/organizations] for their contributions.

References

[1] F. Author and S. Author, "Title of the article," *Journal Name*, vol. 10, no. 2, pp. 123–145, 2023. DOI: 10.1234/example.

[2] F. Author and S. Author, "Title of the conference paper," in *Proceedings of the Conference Name*, IEEE, 2023, pp. 1–10.

[3] F. Author, *Book Title*, 2nd. Publisher Name, 2022.

[4] F. Author, *Title of web resource*, Accessed: 2023-10-01, 2023. [Online]. Available: https://example.com.