Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Optimierung der Rechnungsextraktion durch Einsatz von Large Language Models: Ansätze und Evaluation bei der aifinyo AG

Florian Pruß
Masterarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 18.09.25

Aufgabenstellung

Ziel der Arbeit ist die Evaluation von Large Language Models (LLMs) zur Extraktion strukturierter Rechnungsdaten aus PDF-Dokumenten. Die Untersuchung erfolgt im Kontext der aifinyo AG, einem Berliner FinTech mit über 13 Jahren Erfahrung in der Rechnungsvorfinanzierung. Das Unternehmen verarbeitet monatlich ~15.000 Rechnungen und benötigt eine automatisierte Erfassung der Felder Rechnungsnummer, -datum und -betrag.

Die aktuelle OCR-Lösung (Gini) stößt bei variablen Rechnungslayouts an Grenzen [1]. Die Arbeit untersucht, ob LLMs eine robustere Alternative bieten.

Konzept

Die Untersuchung nutzt einen zweistufigen Evaluationsansatz: Development-Datensatz mit 3.000 komplexen Rechnungen zur Strategieoptimierung, gefolgt von Blind-Evaluation auf 10.000 repräsentativen Rechnungen von 508 Kreditoren.

Drei LLMs werden getestet (Claude 3 Sonnet, GPT-4.1, Gemma 3 27B) mit verschiedenen Prompting-Strategien: Zero-Shot (ohne Beispiele) [2], Few-Shot (3-5 Beispiele) [2,3] und Chain-of-Thought (schrittweise Analyse) [4]. Evaluation mittels dokumentbasierter Accuracy und McNemar-Tests [5].

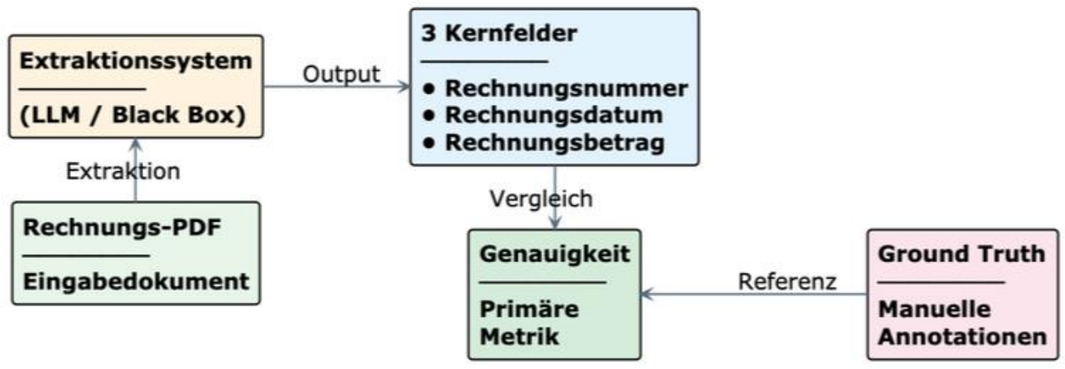

Abb. 1: Blackbox Evaluation

Evaluationsplattform

Zentrales Werkzeug ist eine webbasierte Evaluationsplattform zur systematischen Untersuchung verschiedener LLM-Strategien [6]. Sie verwaltet Dokumente mit Ground-Truth, konfiguriert Prompting-Strategien und führt automatisierte Executions durch. Die Metrikberechnung erfolgt feld- und dokumentbasiert mit statistischen Tests (McNemar [5], Odds Ratios, Wilson-KI [7]). Token-Verbrauch und Laufzeiten werden erfasst.

Die Technische Basis bilden Ruby on Rails, PostgreSQL, einheitliche LLM-Service-API. Textextraktion mittels verschiedener Bibliotheken [8] wie Beispielsweise PDFPlumber (eingebetteter Text) und Tesseract (Bilder).

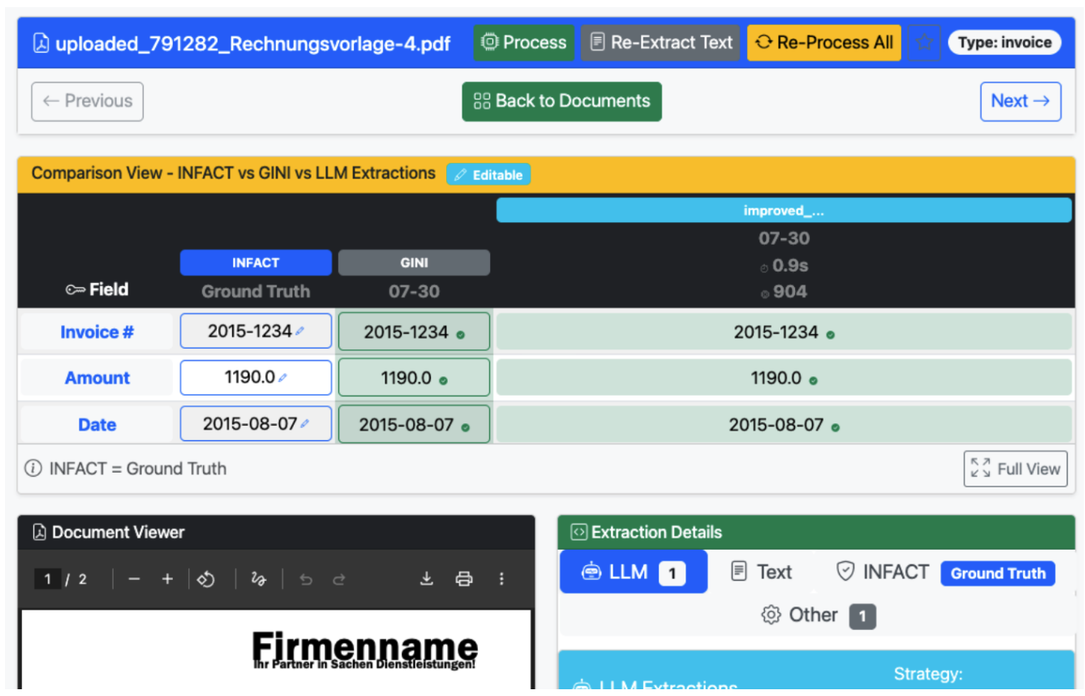

Abb. 2: Evaluationsplattform

Ergebnisse

Die Evaluation auf 10.000 Rechnungen zeigt: Alle LLMs übertreffen die Gini-Baseline deutlich. Claude 3 Sonnet mit Few-Shot CoT erreicht 99,11% vs. Gini mit 87,24%. Dies entspricht +11,87 Prozentpunkten Verbesserung.

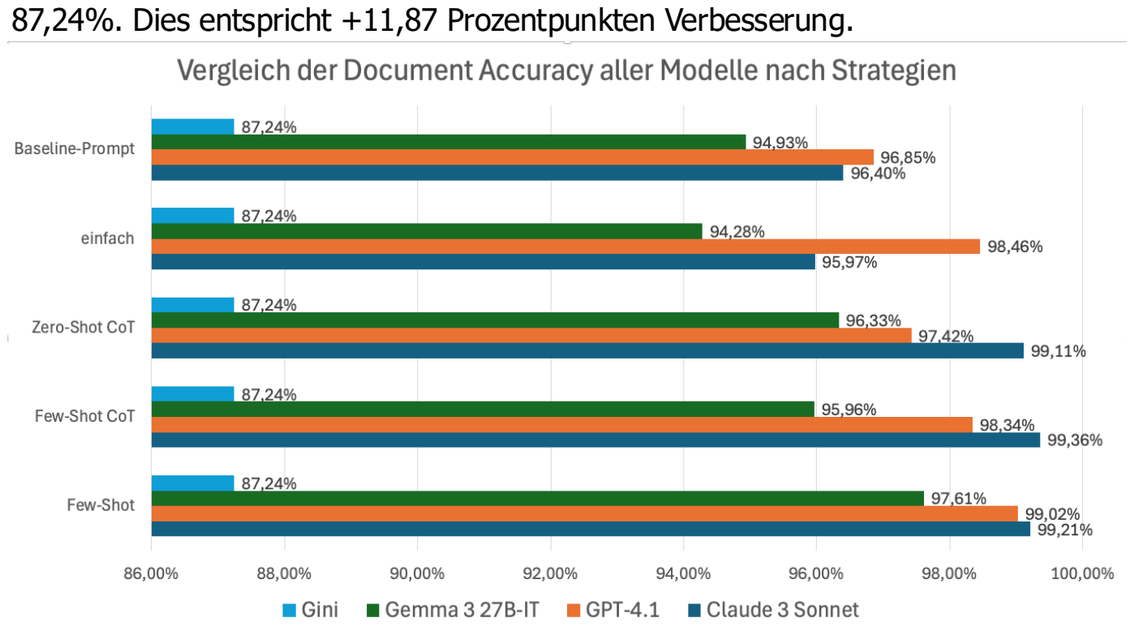

Abb. 3: Strategievergleich

Evaluation der Hypothesen

H_{01} – Layout-Robustheit: Bei 508 Kreditoren zeigen LLMs signifikant bessere Ergebnisse ($p < 0{,}001$). Der McNemar-Test für Claude vs Gini: Odds Ratio 35,6 (95%-KI: 25,5-49,9). Claude korrekt bei 1.247 Fällen wo Gini falsch liegt, umgekehrt nur 35 Fälle. Die Nullhypothese kann somit zurückgewiesen werden.

H_{02} – Prompt-Engineering: Optimierte Strategien steigern Accuracy signifikant.
Claude 3: von 95,97% (Simple) auf 99,11% (+3,14pp).
Gemma 3: von 94,28% (Simple) auf 97,61% (+3,33pp).
GPT-4.1: Few-Shot +0,56pp, aber Zero-Shot-CoT verschlechtert (−1,04pp).
Der McNemar-Test bestätigt statistische Signifikanz. Die Nullhypothese wird für Claude/Gemma zurückgewiesen, für GPT-4.1 teilweise.

H_{03} – Generalisierung: Performance auf Evaluationsdatensatz (n=10.000) entspricht Development (n=3.000). Claude 3 Few-Shot CoT: 99,11% → 99,10%. Es besteht somit keine Überoptimierung und die Nullhypothese kann zurückgewiesen werden.

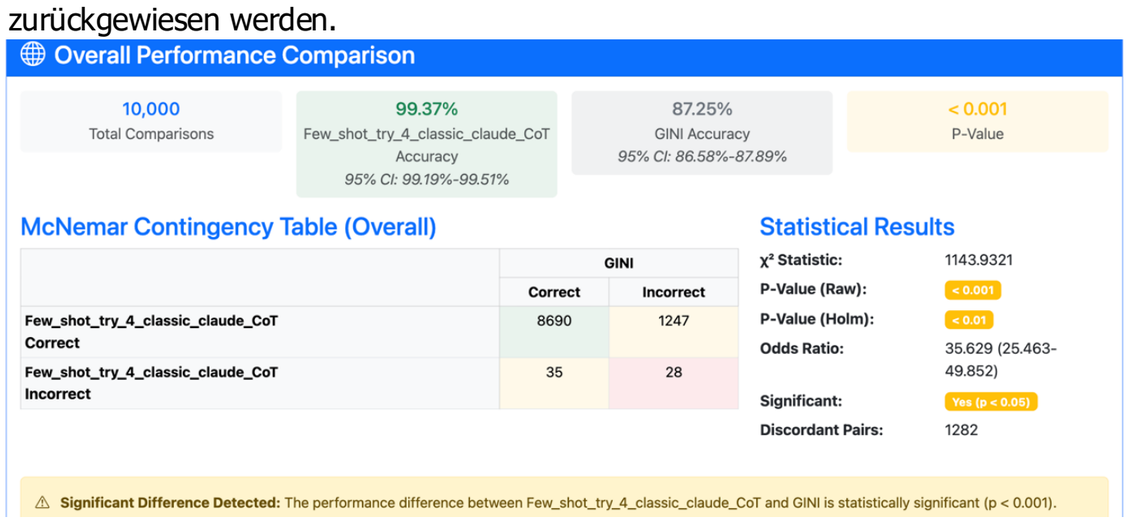

Abb. 4: McNemar-Test

Fazit

LLM-basierte Ansätze verbessern die Rechnungsdatenextraktion signifikant gegenüber OCR. Beste Ergebnisse: Claude 3 Sonnet mit Few-Shot CoT (99,11% auf 10.000 Rechnungen). Alle drei Hypothesen bestätigen: überlegene Layout-Robustheit, messbare Verbesserung durch Prompt-Engineering, zuverlässige Generalisierung.
Empfehlung für aifinyo AG: Pilotierung eines hybriden Systems (LLM-Hauptverarbeitung + OCR-Fallback). Zukünftige Forschung: Multimodale LLMs, Fine-Tuning, Kostenoptimierung.

[1] T. Saout, F. Lardeux, and F. Saubion, "An Overview of Data Extraction From Invoices," *IEEE Access*, vol. 12, pp. 19872-19886, 2024.
[2] T. Brown et al., "Language Models are Few-Shot Learners," *arXiv:2005.14165*, 2020.
[3] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1-35, 2023.
[4] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *arXiv:2201.11903*, 2022.
[5] Q. McNemar, "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages," *Psychometrika*, vol. 12, no. 2, pp. 153-157, 1947.
[6] K. Schmid, S. El-Sharkawy, and C. Kröher, "Improving Software Engineering Research through Experimentation Workbenches," *arXiv:2110.12937*, 2021.
[7] E. B. Wilson, "Probable Inference, the Law of Succession, and Statistical Inference," *J. Am. Stat. Assoc.*, vol. 22, no. 158, p. 209, 1927.
[8] H. Bast and C. Korzen, "A Benchmark and Evaluation for Text Extraction from PDF," in *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 1-10, 2017.

Betreuung: Prof. Dr. Emanuel Kitzelmann • Technische Hochschule Brandenburg • Prof. Dr. Roland Fassauer • CODE University of Applied Sciences