When Ink Meets Pixel: Intent-Aware OCR with VLMs for Human-Corrected Financial Documents

Anonymous ACL submission

Abstract

Human annotations on documents–such as arrows, carets, strikethroughs, and insertions–express corrective intent yet existing OCR and document understanding systems treat these marks as visual noise rather than semantic edit operations. We introduce *intent-aware OCR*, the task of generating corrected transcriptions by jointly reasoning over original document content and visual intent cues (handwritten or digital) that convey edit intent. To support this setting, we present `DocIntent`, a multimodal dataset of 1,600 financial documents with over 5,700 annotated visual intent cues, each paired with layout-aware, intent-resolved OCR targets. We further introduce `IntentBench`, a 216-document benchmark designed to evaluate model robustness under diverse real-world annotation conditions. Leveraging `DocIntent`, we fine-tune `DocIntentOCR`, a suite of 3B and 7B parameter vision-language models, under direct intent supervision on corrected transcriptions. On `IntentBench`, the fine-tuned 3B and 7B models achieve F1-scores of 0.81 and 0.83, respectively – substantially outperforming commerical closed-source (e.g., GPT-4.1 (0.626)), and open-source models (e.g., LLaMA-3.2-Vision-90B (0.50)) demonstrating that intent-aware supervision enables compact domain-specific models to outperform substantially larger general-purpose systems on this task. This work establishes intent-aware OCR as a novel human-in-the-loop document understanding task and provides foundational resources (dataset, benchmark, models) to advance research in interactive document processing.

1 Introduction

When an accountant draws a strikethrough across an incorrect line item on an invoice, or a compliance officer inserts a corrected figure with a caret symbol, these marks constitute a decision – a deliberate instruction to modify the document. Across finance, legal, and auditing workflows, printed and digital documents routinely pass through human hands, accumulating annotations that encode corrective intent: deletions, replacements, insertions, and verifications. These visual cues – strikethroughs, arrows, carets, and circled values – form a rich, informal language through which reviewers communicate how document content should be changed. The scale of this problem is significant: of the estimated 550 billion invoices produced globally each year (Koch, 2019), fewer than one-third achieve straight-through processing without human intervention (Bartolini, 2024), leaving the majority subject to manual review workflows where such annotations accumulate.

Optical Character Recognition (OCR) systems, from traditional engines such as Tesseract (Smith, 2007) to modern Vision-Language Models (VLMs) like LayoutLM (Xu et al., 2020), Donut (Kim et al., 2022), DeepSeek-OCR (Wei et al., 2025) are designed to extract text while being invariant to visual noise. The result is a fundamental disconnect: OCR output reflects the original, unannotated document, not the human *intended* version, because visual cues are treated as noise. In practice, this forces organizations to rely on additional manual correction or brittle rule-based post-processing pipelines that operate outside the visual document space (Barrachina et al., 2009).

We argue that this limitation reflects a missing abstraction: existing OCR formulations lack a mechanism for modeling *human intent*. Annotations (handwritten or digital) are not arbitrary marks—they are visual edit commands: strikethroughs indicate *deletions*, carets indicate *insertions*, and arrows indicate *moves*. We formalize this as *intent-aware OCR*, a document analysis task in which transcription is conditioned on both printed content and visual intent cues, interpreted as semantic edit operations. Depending on scope, these operations may span tokens, fields, table cells, rows, or entire columns.

To enable systematic study of this task, we introduce `DocIntent`, a multimodal dataset of 1,600 synthetic financial documents augmented with over 5,700 visual intent cues, each paired with a layout-aware, intent-resolved OCR target. For realistic evaluation, we present `IntentBench`, a 216-document benchmark of real-world annotated financial documents. Leveraging these resources, we fine-tune `DocIntentOCR`—a family of 3B and 7B models—under direct intent supervision, showing that these compact, domain-specific models achieve up to 0.83 F1, substantially outperforming significantly larger closed- and open-source models, including GPT-4.1 (0.626) and LLaMA-3.2-Vision-90B (0.50).

In summary, this work makes the following contributions:

- We formalize *intent-aware OCR*, a task that generates corrected transcriptions by interpreting visual feedback cues (e.g., arrows, strikethroughs, insertions) as edit operations.
- We introduce `DocIntent` (1,600 documents, 5,700+ intent cues) and `IntentBench` (216 real-world documents), the first dataset and benchmark for intent-aware OCR.
- We fine-tune `DocIntentOCR` under direct supervision, showing that intent-aware fine-tuning enables compact VLMs to outperform much larger models on this task.

By treating handwritten annotations as executable edits rather than noise, this work opens a new direction in document understanding – one where OCR systems become participants in, rather than bystanders to, human review workflows.

2 Related Work

OCR research has progressed from rule-based engines (Smith, 2007) to layout-aware pretrained models (Xu et al., 2020; Huang et al., 2022), end-to-end architectures (Kim et al., 2022; Blecher et al., 2023), and unified models such as GOT-OCR (Wei et al., 2024) and DeepSeek-OCR (Wei et al., 2025). General-purpose VLMs (Bai et al., 2025) and document-specialized architectures (Liu et al., 2024) achieve strong results on benchmarks like DocVQA (Mathew et al., 2021). These systems optimize for text fidelity—handwritten annotations are transcribed verbatim or suppressed as noise, never interpreted as edit operations. Offline handwriting recognition (Marti and Bunke, 2002; Gervais et al., 2024) and strikethrough-removal systems (Heil et al., 2021) similarly treat annotations as signal to recognize or remove.

Human-in-the-loop paradigms such as active learning (Settles, 2009) and interactive machine translation (Barrachina et al., 2009) incorporate user feedback through iterative refinement *after* a model produces output. Intent-aware OCR inverts this: corrective intent is embedded directly in the input as visual annotations, and the model must recover it from a single static image without iterative clarification.

3 Intent-Aware OCR: Task Definition

This section formalizes intent-aware OCR as a document analysis task, defines the space of visual intent cues and their induced edit operations, and specifies the evaluation protocol.

3.1 Formulation

Let I denote a document image containing printed text, layout structure, and zero or more handwritten or digital annotations. A conventional OCR system maps I to a transcription T that faithfully reproduces all printed content. Intent-aware OCR instead maps I to an *intent-resolved transcription* O that reflects modifications specified by visual annotations. Importantly, the formulation is agnostic to the serialization format of O: it applies equally whether the output is a layout-aware transcription that preserves two-dimensional document structure, a linear reading-order rendering, or any other text representation. The edit operations defined below operate over logical document regions (tokens, cells, rows, columns, tables) and are therefore independent of how the final transcription is linearized.

We model this as follows. Let T denote the base transcription of the printed content, and let $\mathcal{H} = \{h_1, h_2, \ldots, h_n\}$ denote the set of visual intent cues present in the image. Each cue h_i is associated with:

1. a **cue type** c_i drawn from the taxonomy in Table 3 (Appendix A),
2. a **structural scope** s_i specifying the granularity of the edit (token, field, cell, row, column, or table), and
3. an optional **replacement value** v_i (for replacement and insertion operations).

Each cue h_i induces a deterministic edit operation $e_i = (c_i, s_i, v_i)$ over the base transcription. The

intent-resolved output is defined as the sequential application of all edit operations:

$$O = e_n \circ e_{n-1} \circ \cdots \circ e_1(T)$$

In our formulation, each edit e_i targets a spatially distinct region of the base transcription (a specific token span, cell, row, column, or table). When edit scopes are pairwise disjoint [1], the composition is commutative and the ordering is immaterial. In principle, overlapping scopes can arise (e.g., a row-level deletion and a cell-level replacement within the same row); we treat such cases as annotation ambiguities outside the scope of this taxonomy, and they do not occur in `DocIntent` or `IntentBench`.

Critically, the model does not observe T or $\mathcal{H}$ directly–it must infer both from the raw image I. The task therefore requires the model to simultaneously: (1) recognize the printed document content, (2) detect and classify visual annotations, (3) determine the structural scope of each annotation, and (4) apply the induced edits to produce the final output.

3.2 Edit Operation Taxonomy

We define three high-level categories of intent-induced transformations:

- **Deletion:** existing text or structured content is marked for removal (e.g., a strikethrough with no replacement value).
- **Replacement:** existing content is invalidated and substituted with a corrected value (e.g., a strikethrough accompanied by a handwritten correction).
- **Insertion:** new content is introduced at a specified position without removing existing text (e.g., a caret with handwritten text).

Each category may operate at varying structural granularities—from individual tokens to entire rows, columns, or tables. A key property is that the same visual cue maps to different operations depending on context: a strikethrough alone induces deletion, whereas a strikethrough with a nearby handwritten value induces replacement. Table 3 (Appendix A) enumerates all cue types, scopes, and expected OCR behaviors.

3.3 Illustrative Example

Figure 1 illustrates the task on a sample from `DocIntent`. A human reviewer has: (1) struck through the invoice number `#437346` and drawn an arrow pointing to the handwritten replacement `246952` (arrow-indicated replacement at the token level), (2) placed a caret below the word "drop" in the item description with the handwritten text "gold" (insertion at the token level), and (3) scribbled out the phone number in the billing section (deletion at the token/field level). Standard OCR transcribes the document verbatim, ignoring these annotations. Intent-aware OCR interprets each annotation as an edit operation and produces a corrected transcription—outputting `246952` in place of the original invoice number, expanding the description to "Gold m.o.p orbit drop *gold* earrings," and omitting the redacted phone number.

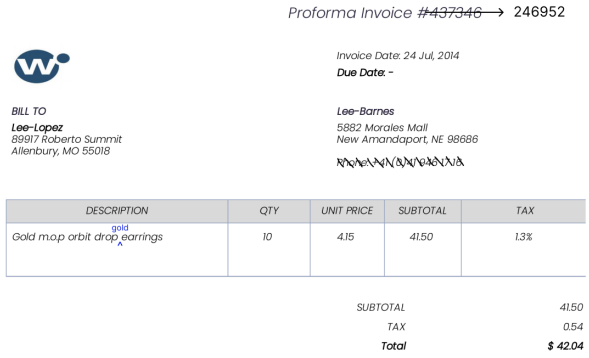

Figure 1: A sample annotated invoice from `DocIntent`. Three intent cues are present.

3.4 Evaluation Protocol

We evaluate by comparing predicted intent-resolved transcriptions against ground-truth outputs using token-level F1-score, measuring precision and recall over whether deletions, replacements, and insertions are correctly applied. All evaluation is performed on `IntentBench`. In

4 Datasets

We introduce two complementary resources: `DocIntent`, a large-scale synthetic dataset for training, and `IntentBench`, a real-world benchmark for evaluation. Both are constructed around the visual intent cue taxonomy (Appendix A). Table 1 summarizes key characteristics.

4.1 `DocIntent`: Training Dataset

`DocIntent` is a multimodal dataset of 1,659 financial documents—invoices, statements, and structured forms—augmented with 5,641 visual intent

[1] This holds for all samples in our proposed datasets.

Table 1: Overview of DocIntent and IntentBench.

Statistic	DocIntent	IntentBench
# Documents	1,659	216
Total intent cues	5,641	758
Avg. cues per doc	3.40	3.03
Avg. tokens per doc	157.51	175.79
Cue types covered	10	9
Digital (%)	74.8	88.4
Scanned (%)	25.2	11.6
# Document types	33	18

Table 2: IntentBench OCR Performance (F1).

Model	F1
GPT-4.1	0.6261
Claude Sonnet 4.6	0.5974
Mistral-Large-3-675B-Instruct-2512	0.6077
Llama-3.2-90B-Vision-Instruct	0.4982
Qwen3-VL-32B-Instruct	0.5414
Qwen2.5-VL-7B-Instruct	0.6485
Qwen2.5-VL-3B-Instruct	0.6563
DocIntentOCR-3B	**0.8102**
DocIntentOCR-7B	**0.8328**

cues spanning all taxonomy categories, with 74.8% digitally rendered and 25.2% physically annotated and scanned. Table 1 summarizes key statistics; per-cue-type breakdowns are in Appendix B. See Appendix C for full annotation protocol details.

Supervision Signal. Each document in DocIntent is paired with a layout-aware intent-resolved transcription that applies all visual cues while preserving two-dimensional document structure (table boundaries, row/column associations, reading order).

4.2 IntentBench: Evaluation Benchmark

IntentBench is a held-out benchmark constructed independently of DocIntent and excluded entirely from training. It features diverse financial documents with substantial variation in handwriting styles, annotation density, and layouts, testing generalization beyond the training distribution. Ground-truth annotations are intent-resolved, layout-aware transcriptions. See Table 1 for summary statistics and Appendix B for detailed breakdowns.

Inter-Annotator Agreement. Two annotators independently annotated all 216 IntentBench documents. Agreement, measured by normalized Levenshtein ratio (Levenshtein, 1966), yielded a mean similarity of 0.77 ($\sigma = 0.12$), indicating moderate-to-strong consistency. Full IAA statistics are in Appendix D.

5 Experimentation

We fine-tune 3B and 7B variants of Qwen2.5-VL-Instruct models on DocIntent using an 80/20 train/validation split. The fine-tuned models are denoted DocIntentOCR-3B and DocIntentOCR-7B, and evaluated on IntentBench. Both models were fine-tuned using LoRA (r=16, α=32) on language model layers with the last vision encoder block unfrozen, optimized with AdamW (1×10^{-4}/1×10^{-5}), for 7 and 5 epochs respectively on two NVIDIA A100 40GB GPUs. To fairly evaluate zero-shot models, we provide prompt that explicitly defines annotation semantics (E.1); as well as the hyperparameter configurations (E.2).

6 Results

Table 2 reports IntentBench F1 scores. The results reveal a clear pattern: intent-aware fine-tuning on DocIntent is the dominant factor, not model scale—zero-shot models, including GPT-4.1 (0.626), struggle because they have no mechanism to interpret visual correction marks as semantic edit operations. DocIntentOCR-3B (0.810) and DocIntentOCR-7B (0.833) both outperform all baselines, demonstrating that a compact, locally deployable model trained on intent-aware supervision is a practical and privacy-friendly alternative to large frontier APIs.

7 Conclusion and Future Work

We present intent-aware OCR, a novel document understanding task that interprets human annotations-strikethroughs, arrows, carets, and insertions as semantic edit operations. To support this task, we introduce DocIntent, a dataset of 1,600+ annotated financial documents, and IntentBench, a 216-document real-world benchmark. Fine-tuned on DocIntent, DocIntentOCR in both 3B and 7B configurations substantially outperform GPT-4.1 and other frontier models in the zero-shot setting, demonstrating that intent-aware supervision is the critical factor rather than model scale. Future work includes validating the formulation on legal and medical documents, expanding the taxonomy to handle compositional annotations, and scaling training data via synthetic image generation.

Limitations

Our dataset and evaluation are restricted to financial documents, and generalization to other domains remains unvalidated. The taxonomy does not cover all annotation styles such as circling, underlining, and marginal comments, and the predominantly digital training data (74.8%) may limit robustness to diverse real-world handwriting. Inter-annotator agreement of 0.77 indicates that spatially ambiguous annotations admit subjective interpretations. Finally, the structural scopes in our taxonomy (token, field, cell, row, column, table) are designed primarily for tabular layouts; documents with irregular or free-form layouts may require alternative scope definitions.

Ethics Statement

All documents in `DocIntent` and `IntentBench` are synthetically generated or sourced from publicly available templates; no real financial records or personally identifiable information were used. Human annotators were recruited voluntarily, informed of the task purpose, and compensated fairly for their time. Annotation instructions were designed to avoid subjective or harmful content.

Portions of this paper were drafted with the assistance of large language model tools (Claude, OpenAI). All scientific content, experimental results, and conclusions were produced and verified by the authors.

References

Shuai Bai, Keqin Chen, Xuejing Liu, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Sergio Barrachina, Oliver Bender, and 1 others. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Andrew Bartolini. 2024. The state of ePayables 2024: Money never sleeps. Technical report, Ardent Partners.

Lukas Blecher, Guilherme Correia, and Klemens Müller. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Philippe Gervais, Asya Fadeeva, and Andrii Maksai. 2024. MathWriting: A dataset for handwritten mathematical expression recognition. *arXiv preprint arXiv:2404.10690*.

Raphaela Heil, Ekta Vats, and Anders Hast. 2021. Strikethrough removal from handwritten words using CycleGANs. In *Proceedings of ICDAR*, pages 572–586.

Zhe Huang, Jianlong Wu, Yibo Chen, and 1 others. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Geewook Kim, Teakgyun Kim, Jinyoung Kim, and Byung-Seok Roh. 2022. Donut: Document understanding transformer without OCR. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 414–430.

Bruno Koch. 2019. The E-Invoicing journey 2019–2025. Technical report, Billentis.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Yuliang Liu, Biao Yang, Qiang Liu, and 1 others. 2024. TextMonkey: An OCR-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Urs-Viktor Marti and Horst Bunke. 2002. The IAM database: An english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

Minesh Mathew, Dimosthenis Karatzas, and 1 others. 2021. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Burr Settles. 2009. Active learning literature survey. Technical report, University of Wisconsin–Madison.

Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633.

Haoran Wei, Chenglong Liu, Jinyue Chen, and 1 others. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Haoran Wei, Yaofeng Sun, Yukun Li, and 1 others. 2025. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1192–1200.

A Intent Cue Taxonomy

Table 3 enumerates the full taxonomy. While the ten cue types are grounded in financial documents, the three edit operation categories (deletion, replacement, insertion) are domain-agnostic primitives, and the visual cues that express them (strikethroughs, carets, arrows) are conventional marks used across professional domains. Extending coverage to legal redlining, medical records, or engineering change orders may require additional cue types and scope definitions; empirical validation in non-financial domains remains an important direction for future work.

B Dataset Statistics

Tables 4 and 5 report the per-cue-type distribution of visual intent cues in `DocIntent` and `IntentBench`, respectively. Replacement cues dominate both splits, with strikethrough-with-replacement and arrow-indicated replacement being the most frequent types.

C Annotation Protocol

Base documents are generated from a library of custom-built HTML financial templates spanning 33 document types—including invoices, statements, and structured forms—populated with synthetically generated content (names, line items, amounts) designed to be mathematically consistent (e.g., correct subtotals, taxes, and grand totals). Templates are rendered to images using a headless browser pipeline, producing clean, high-resolution document images with precise control over layout, table structure, and spatial organization—properties essential for deterministic placement of intent cues at specific structural granularities.

Handwritten annotations were produced by five annotators with academic backgrounds in commerce and prior exposure to accounting and business documentation. Annotators were given document-specific correction instructions specifying which content to delete, replace, or insert, and were asked to apply corrections using natural handwriting without access to model predictions or automated suggestions. Each annotated document was independently reviewed by a second annotator to verify that all intended cues were correctly applied and visually unambiguous. Annotators were fairly compensated, and the study adhered to established ethical standards.

D Inter-Annotator Agreement

To assess the reliability of `IntentBench` reference annotations, two annotators independently produced intent-resolved transcriptions for all 216 documents; agreement was quantified using the normalized Levenshtein ratio (Levenshtein, 1966). Table 6 reports the resulting similarity statistics, with a mean score of 0.77 indicating moderate-to-strong consistency across annotators.

E Experimental Configurations

E.1 Prompts

To evaluate `DocIntentOCR-3B` and `DocIntentOCR-7B`, we use the same instruction as employed during training: *"You are an OCR assistant. Analyze the document image and transcribe all visible text, interpreting any annotations as edit instructions. Produce the final corrected transcription."*

For all other non-finetuned models and closed-source models, we use a slightly expanded prompt to explicitly define annotation semantics: *"You are an OCR assistant for annotated financial documents. Interpret annotations as edit instructions (e.g., strikethroughs for deletions or replacements, arrows for replacements, carets for insertions, etc.). Output the final corrected transcription, preserving layout."*

E.2 Inference Hyperparameters

All models were evaluated with temperature 0.1, top-p 0.90, and a maximum output length of 8,000 tokens. Zero-shot baselines were accessed via the OpenAI API (GPT-4.1), Anthropic API (Claude Sonnet 4.6), and AWS Bedrock (Llama-3.2-90B and Mistral-Large-3-675B).

F Output comparison of DocIntentOCR-7B with GPT-4.1

Table 3: Taxonomy of visual intent cues for intent-aware OCR. Each cue induces a deterministic edit over the base transcription at a specific structural granularity.

Category	Cue Type	Scope	Expected OCR Behavior
Deletion	Strikethrough	Token / Field	Marked content is omitted from the output.
	Row Strikethrough	Table Row	All cells in the struck-through row are removed.
	Column Strikethrough	Table Column	All values in the struck-through column are removed.
	Full-Table Strikethrough	Table	The entire table is omitted from the output.
Replacement	Strikethrough + Handwritten Value	Token / Field	Struck-through text is replaced by the nearby handwritten value.
	Arrow-Indicated Replacement	Token / Field	Struck-through text is replaced by the value the arrow points to.
	Cell Correction	Table Cell	The cell value is replaced by the handwritten correction.
	Row Correction	Table Row	Each cell in the row is substituted with its per-cell replacement.
	Column Correction	Table Column	Each cell in the column is substituted with its replacement value.
Insertion	Caret Insertion	Token / Field	New handwritten text is inserted at the caret position without deleting existing content.

Table 4: Distribution of visual intent cues in `DocIntent`. Each row reports the total number of cue instances, the number of documents containing at least one instance, and the average frequency per document across all 1,659 documents.

Category	Cue Type	Total	Docs	Avg./Doc
Deletion	Strikethrough w/o Replacement	986	833	0.59
	Table Row Deletion	189	169	0.11
	Table Column Deletion	333	303	0.20
	Full Table Deletion	133	130	0.08
Replacement	Strikethrough w/ Replacement	1,880	1,233	1.13
	Arrow-Indicated Replacement	1,214	1,010	0.73
	Table Cell Replacement	463	315	0.28
	Table Row Replacement	16	12	0.01
	Table Column Replacement	45	33	0.03
Insertion	Caret Insertion	382	331	0.23
Total		**5,641**	—	**3.40**

Table 5: Distribution of visual intent cues in `IntentBench`. Each row reports the total number of cue instances, the number of documents containing at least one instance, and the average frequency per document.

Category	Cue Type	Total	Docs	Avg./Doc
Deletion	Strikethrough w/o Replacement	160	143	0.74
	Table Row Deletion	1	1	0.00
	Table Column Deletion	30	30	0.14
	Full Table Deletion	12	12	0.06
Replacement	Strikethrough w/ Replacement	209	185	0.97
	Arrow-Indicated Replacement	192	180	0.89
	Table Cell Replacement	28	23	0.13
	Table Column Replacement	1	1	0.00
Insertion	Caret Insertion	22	21	0.10
Total		**655**	—	**3.03**

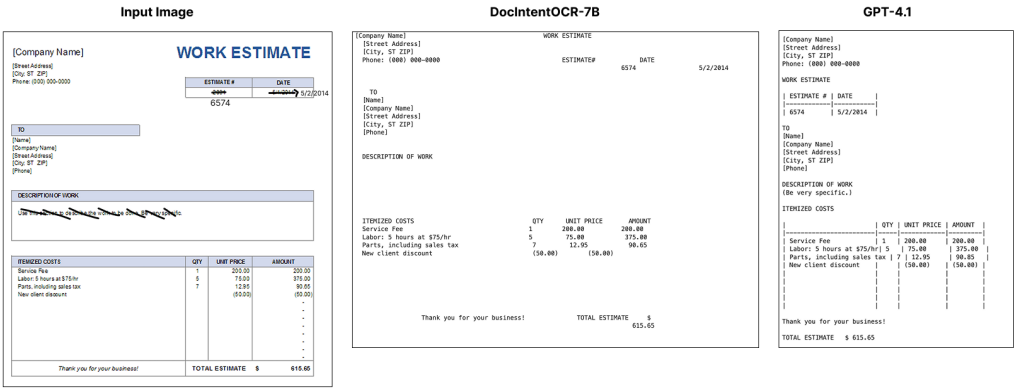

Figure 2: Comparison of outputs from DocIntentOCR-7B and GPT-4.1 on annotated financial documents.

Table 6: Inter-Annotator Agreement (IAA) on `IntentBench` annotations using fuzzy string similarity (normalized Levenshtein ratio).

Metric	**Value**
Number of annotation pairs	216
Mean similarity	0.77
Minimum similarity	0.31
Maximum similarity	0.99
Standard deviation	0.12