Perception, Layout, and Validation: Calibrated Confidence for Reliable Straight-Through Processing of Financial Documents

Yichao Jin, Yushuo Wang, Yuxuan Han, Kwan Ching Yee Sonia, Weiyang Song, Chiu Jin-Chun Kent, Wong Chong Hwee, Wong Tiong Kiat, Kenneth Zhu Ke, Jingyuan Zhao
{JinYichao, YushuoWang, YuxuanHan, SoniaKwan, WeiyangSong, KentChiu, ChongHweeWong, WongTK, KennethZhu, JingyuanZhao}@ocbc.com
OCBC, Singapore

Abstract

Straight-through processing (STP) on extracted key-value fields from financial documents without human review requires a calibrated probability together with a bounded guarantee on the residual error of the auto-approved tier. The emergence of modern Vision Language Models (VLMs) provides an out-of-the-box capability for extracting the key-values, but their verbalized confidence signals are unreliable and weakly track field correctness.

This paper introduces a decomposed confidence layer along three interpretable channels, including perception, layout, and validation. Together with a final conformal risk control, the score can be used for reliable STP of financial documents. The method is validated on three public datasets covering real invoices, synthetic invoices, and ad-buy forms, using two different VLM families (Qwen3.6-27B and Gemini-3.1-Flash-Lite). Our decomposed score consistently improves the separation of correct from incorrect extractions, substantially raising the AUROC from 0.54–0.74 for VLM verbalized signals to 0.90-0.99 with contributions from all three designed channels. Crucially for industrial deployment, this enables usable STP. The native VLM confidence signals could clear only 0.1%-7.0% of fields under risk control at a target error of $\leq 10\%$. In contrast, the proposed method auto-approves 49-72% of fields while holding the empirical error of the accepted tier at or below the target.

1 Introduction

Financial back-office operations, such as accounts payable and client onboarding, have to process large volumes of semi-structured documents (e.g., invoices and trade orders) under tight time constraints. Currently, in most financial institutions, manual verification of the extracted fields still dominates the per-document cost. Therefore, *Straight-through processing* (STP), which auto-approves the extractions with minimal human review, becomes the key driver of productivity and scalability.

In a financial setting, STP carries a stringent bar. In particular, a single wrongly auto-approved value (e.g., a transposed amount or a misread account number) is a critical financial and compliance error. Safe STP thus requires two things beyond a point extraction. First, a per-field probability that is *calibrated*. Second, a *distribution-free bound* on the residual error of the auto-approved tier. Together, these hold the error rate among accepted fields below an operator-chosen budget.

Modern vision-language models (VLMs) have largely solved the extraction step. Prompted with a page image and a field schema, they emit values directly and usually accurately, without task-specific training. What they lack is a reliable confidence signal. A VLM *can* verbalize a confidence score, but these are often saturated and over-confident. On our data, verbalized confidence separates correct from incorrect extractions with an AUROC of only 0.54–0.74 (i.e., at times worse than chance). Token log-probabilities, where exposed, sit over generated sub-word tokens and track field correctness only weakly. Even sampling-based self-consistency (Wang et al., 2022) is blind to *systematic* errors, where the model returns the same wrong value on every sample. Consequently, a VLM-only STP system is caught between auto-approving almost nothing and auto-approving errors.

We introduce a decomposed confidence layer for VLM key-value extraction, which scores each extracted field along three interpretable channels. The **Perception** channel checks if the evidence is legible from region-local image quality, VLM-OCR agreement, and value-string cues. The **Layout** channel asks if the value sits where this field normally appears. We retrieve historical layouts similar to the current page, and score the field's anchor-relative position under a retrieval-conditioned spatial prior. A value in an unusual location would be

The Proceedings of Workshop on Document Intelligence and Understanding, pages 232–243
October 24, 2026 ©2026 Association for Computational Linguistics

flagged even when read perfectly. Lastly, the **Validation** channel applies type, checksum, and rules, and checks cross-model agreement against a second independent VLM. A gradient-boosted model fuses the three channels into a single probability, with per-field reason codes. We wrap this score in a Learn-Then-Test conformal procedure (Angelopoulos et al., 2025), which certifies a distribution-free bound on the error rate of the auto-approved tier. Finally, when two VLMs disagree, we select the value our layer scores higher. This resolves the disagreement rather than rejecting it, raising STP throughput at the same guaranteed error.

We validate the method on three public datasets. They span real invoices from DocILE (Šimsa et al., 2023), synthetic invoices from FATURA (Limam et al., 2023), and advertising order forms from VRDU (Wang et al., 2023), using two VLM families, Qwen3.6-27B (Qwen Team, 2026) and Gemini-3.1-Flash-Lite (DeepMind, 2026). Our main contributions include

- A decomposed and interpretable confidence layer consisting of perception, layout, and validation channels. The fused conformal score reliably reflects the field correctness with a distribution-free error guarantee. Experiments show a substantial AUROC lift from 0.54–0.74 (VLM baseline) to 0.90–0.99.

- A cross-model verification-and-selection mechanism, that resolves disagreements between two VLMs, further lifts the guaranteed STP coverage on top of fusion of the three channels. The full layer auto-approves 49-72% of fields at a $\leq$ 10% target error, versus 0.1-7.0% from native signals.

- Per-channel and per-feature SHAP attributions that make each accept and reject decision auditable, as required for model-risk and governance review in a regulated setting.

2 Related Work

Document key-value extraction. Early field extractors coupled OCR with layout-aware encoders, such as the LayoutLM family (Xu et al., 2020, 2021; Huang et al., 2022) and OCR-free models like Donut (Kim et al., 2022). These require task-specific fine-tuning and predict grounded spans. Modern VLMs instead read a page image and emit field values directly, without training. However, the extracted values are often ungrounded without reliable boxes and confidence scores.

Our proposed layer is agnostic to the extractor, scoring the confidence of already-extracted values and recovering grounding from OCR spans.

Confidence and calibration for extraction. Confidence for extraction and generation has relied on token log-probabilities, verbalized confidence (Xiong et al., 2024), and sampling-based self-consistency (Wang et al., 2022). For VLM field extraction these signals are weak. The closest line fuses heterogeneous signals (i.e., log-probabilities, OCR quality, image quality, and coarse spatial features) into a single learned score and reports an aggregate AUROC (Kumar, 2026).

Our work differs in three ways. First, we add a retrieval-conditioned layout prior as a first-class signal, rather than coarse spatial features. Second, we keep the score interpretable through a per-channel decomposition with reason codes. Third, we target a distribution-free straight-through-processing guarantee, not only a ranking metric.

Conformal risk control and selective prediction. Selective classification abstains on low-confidence inputs (Geifman and El-Yaniv, 2017). Split conformal prediction turns any score into distribution-free coverage (Vovk et al., 2005). Learn-Then-Test (Angelopoulos et al., 2025) and conformal risk control (Angelopoulos et al., 2024) extend this to control general risks, including the selective error of an accepted set through multiple-hypothesis testing.

We use Learn-Then-Test to certify a distribution-free bound on the error rate of the auto-approved tier, class-conditional by field type. Distinctively, our retrieval score is both an input to the confidence model and the variable that indexes the calibration regime, linking template familiarity to the risk-control guarantee.

3 Problem Statement and Methodology

Figure 1 shows the end-to-end confidence layer. The input is a pdf document, where a VLM extracts field values directly with verbalized confidence score. In parallel, OCR yields word spans, and a retrieval step fetches historical layouts similar to the page. Then three channels examine the corresponding features to score each extracted value. In particular, the **perception** channel checks if the evidence is legible. The **layout** channel checks whether the value sits where the field normally ap-

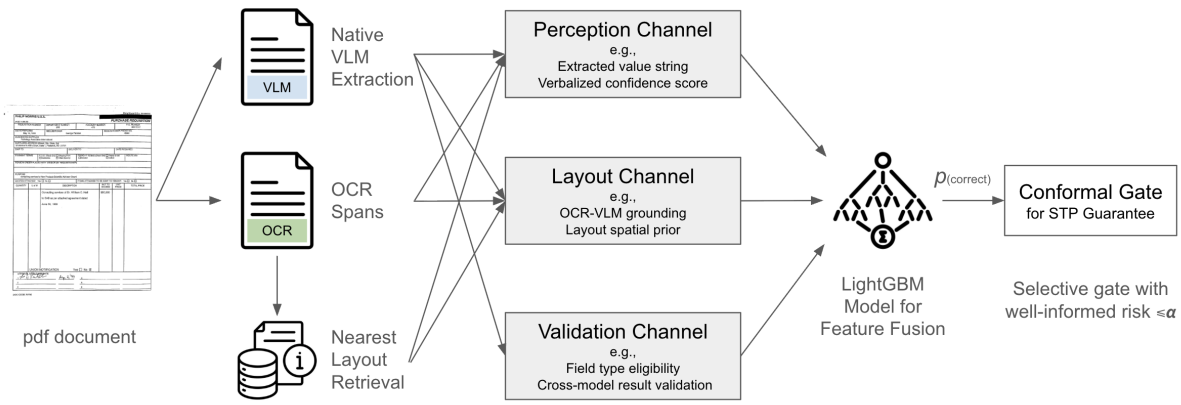


Figure 1: End-to-end confidence layer. The VLM extracts ungrounded field values from the page image. OCR spans recover grounding, and nearest-layout retrieval supplies similar historical templates. Three interpretable channels (i.e., perception, layout, and validation) score each extracted value. A gradient-boosted model (LightGBM) fuses them into a calibrated probability $p(\text{correct})$, which a conformal selective gate turns into an auto-approval decision at a guaranteed selective risk $\leq \alpha$.

pears. The **validation** channel applies content rules and cross-model agreement. Subsequently, a LightGBM model fuses the three channels into a single probability p_f. Finally, a conformal selective gate turns p_f into an auto-approval decision with a controlled error rate. We detail each component below.

3.1 Grounding recovery

A VLM reads a document image and a field schema, and returns a value v_f and its verbalized confidence score s_f for each field type f. Our layer assigns each extracted field (f, v_f, s_f) a calibrated probability p_f that the value is correct, using the page and its recovered OCR grounding.

We first recover grounding from OCR word spans, where each span is a candidate c with printed text $\text{text}(c)$ and a bounding box. The candidate set for field f collects the spans whose canonical text matches the extracted value

$$C_f = \{\, c : \text{canon}_f(\text{text}(c)) = \text{canon}_f(v_f) \,\}$$

where v_f is the VLM value and canon_f is a per-type canonicalizer which handles numeric tolerance, ISO-8601 dates, and normalized strings. $|C_f|$=0 means the value is not on the page (a strong error signal) and $|C_f|$>1 means the value is ambiguous. The details of canonicalizer design are illustrated in Appendix A.

3.2 Perception channel

The perception channel asks whether the evidence is legible and read correctly with 11 features. From the value string we take length, token count, digit ratio, and confusable-glyph mass (e.g., $0/O$, $1/l$, $5/S$). At the candidate box, we take region-local image quality (i.e., Laplacian variance, contrast, ink density, and glyph height) and the edit distance between v_f and the OCR text there. We also include the VLM's verbalized confidence.

3.3 Layout channel

The layout channel contributes 15 features and is illustrated end to end in Figure 2. It asks whether v_f sits where field f normally appears from documents with similar layout.

Anchor-relative geometry. We encode a candidate's position relative to a nearby label anchor rather than in absolute page coordinates. The anchor is chosen geometrically, not semantically. Among all OCR words on the page (excluding the value's own tokens), we take the nearest label-like word under a fixed preference order. A word on the same text line and to the left of the value (within $0.6h_v$ vertically, where h_v is the value height) is preferred, scored by its horizontal distance. So the closest same-row-left word wins. This captures the common "`Label: value`" layout. If not found, we fall back to a word directly above and roughly column-aligned (within $3h_v$ horizontally). This handles column-header layouts. The matching uses no keyword list or field-type dictionary, so it is robust to unseen templates. Given the anchor at center $center(a)$ with height h_a, the position feature is the offset normalized by anchor height as

$$z = \frac{\text{center}(v) - \text{center}(a)}{h_a}$$

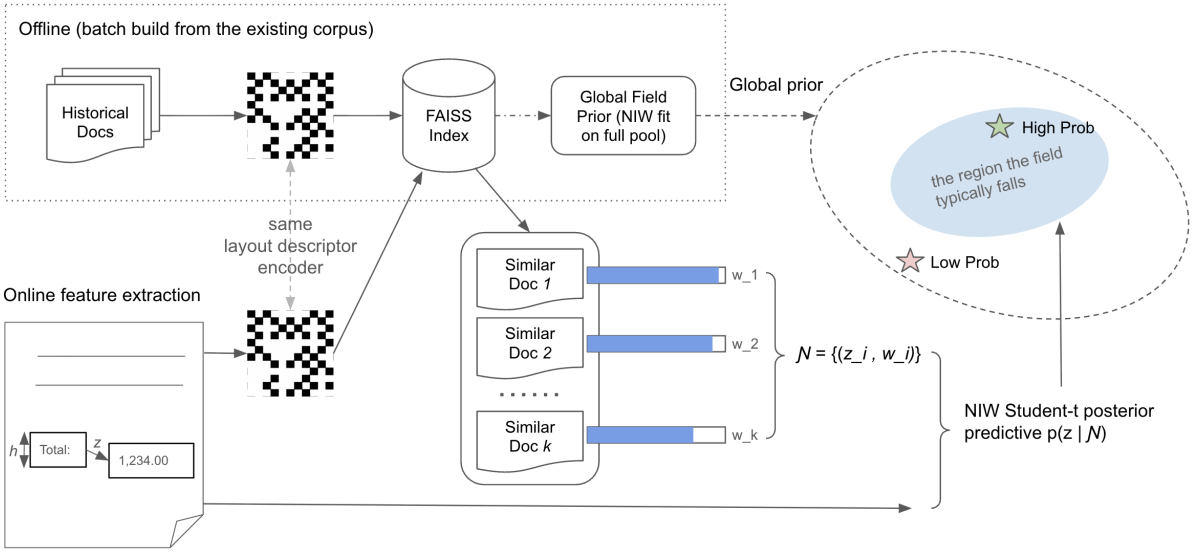


Figure 2: Layout-channel drill-down. **Offline** (top): each historical document is encoded by the layout-descriptor encoder ϕ into a FAISS index, and the field positions across the corpus are pooled to fit a global Normal–Inverse–Wishart (NIW) prior. **Online** (per field): (1) grounding gives the anchor-relative position z of the value (offset from its label anchor, normalized by anchor height h); (2) the query page is encoded by the *same* descriptor ϕ and retrieves the k nearest historical layouts, each contributing its field position and a similarity weight, $\mathcal{N} = \{(z_i, w_i)\}$; (3) The global prior updated with the similarity-weighted neighbours $\mathcal{N}$ yields a Student-t posterior predictive $p(z \mid \mathcal{N})$. The shaded blue region is where the field typically lands for this template (tight), while the dashed ellipse is the global prior it falls back to when no neighbours are found (wide). A candidate whose position lies in the blue region scores high, whereas one in the tail (Low Prob) is penalized even when read correctly.

z is invariant to translation and to scale across templates and page sizes. When no anchor qualifies, z falls back to the offset from the page centre.

Retrieval. We encode each page as a binary token-mask descriptor (details in Appendix D) and index the historical documents with FAISS (Douze et al., 2025). To inference a new page we retrieve the k nearest layouts with similarity weights $\{w_i\}_{i=1}^{k}$. We summarize familiarity by the top similarity $s_{\text{match}} = \max_i w_i$ and a soft effective neighbour count $k_{\text{eff}} = \left(\sum_i w_i\right)^2 / \sum_i w_i^2$.

Retrieval-conditioned spatial prior. We model z for each field type with a Normal–Inverse–Wishart (NIW) (Murphy, 2007) prior on its mean and covariance, fit on the global field-conditional pool. We update the prior with the retrieved neighbours' positions, weighted by similarity, and marginalize to a Student-t posterior predictive $p(z \mid \mathcal{N})$. A single shrinkage hyperparameter κ_0 governs how fast neighbours override the global prior. The behaviour is self-regularizing in familiarity. In particular, at $k_{\text{eff}}{=}0$ the posterior collapses to the global field prior (cold start); small k_{eff} yields heavy tails and appropriately low confidence; large k_{eff} concentrates around the template-specific location.

Non-circular scoring. A value can appear at several places on the page, so C_f may contain multiple positions z_c. We score the field's placement with the layout score s_L, the log-average posterior density over these occurrences as

$$s_L = \log \frac{1}{|C_f|} \sum_{c \in C_f} p(z_c \mid \mathcal{N}).$$

Averaging rather than taking the maximum avoids a selection bias. Selecting the maximum alone would make a value look correct merely because one of its occurrences happens to land in a high-density region. Instead, the spread across occurrences is surfaced separately as ambiguity signals, derived from the occurrence distribution $\pi_c \propto p(z_c \mid \mathcal{N})$ and its entropy H_f. We expose three instances of s_L into the feature set, including the retrieval-conditioned posterior `s_l_marg`, the global prior with retrieval off `s_l_cold` whose gap to the primary exposes the retrieval lift, and under an absolute-position prior `s_l_abs` as a weak baseline.

3.4 Validation channel

The validation channel applies content rules and contributes 8 features. It checks type and format validity, ISO-4217 currency and date validity, checksums where a field admits one (e.g. IBAN mod-

Method	DocILE		FATURA		VRDU	
	Qwen	Gemini	Qwen	Gemini	Qwen	Gemini
Self-consistency	0.728	0.660	0.605	0.544	0.629	0.560
VLM baseline	0.736	0.702	0.614	0.597	0.597	0.550
+ P/L/V fusion	0.911	0.908	0.990	0.996	0.921	0.934
+ cross-model	**0.915**		**0.997**		**0.928**	

Table 1: AUROC for predicting extraction correctness across four approaches, on three datasets and two VLM families. The cross-model row fuses both VLMs into one output (one value per dataset).

97, ISIN), and arithmetic residuals across amount fields (e.g., net + tax = gross). It also records cross-model agreement on whether a second, independent VLM returns the same canonical value.

3.5 Feature fusion model

A gradient-boosted model fuses the three channels into a single probability p_f. We use LightGBM over the concatenated per-channel features, learning correlated interactions. The model is tabular and yields SHAP attributions, so each accept/reject decision carries per-feature reason codes for informative model-risk review. The full feature list can be found in Appendix B.

3.6 Conformal selective gate

A calibrated probability is not yet a decision. The gate accepts a field iff $p_f \geq \lambda$, where λ is chosen on a held-out calibration set so that the *selective risk* (error rate among accepted fields) stays at or below an operator-chosen budget α. Learn-Then-Test (Angelopoulos et al., 2025) makes this threshold search valid, giving a distribution-free guarantee at confidence $1-\delta$. As a result, the auto-approved tier carries error at or below α, and everything below the threshold is routed to human review.

4 Experiments

4.1 Experimental setup

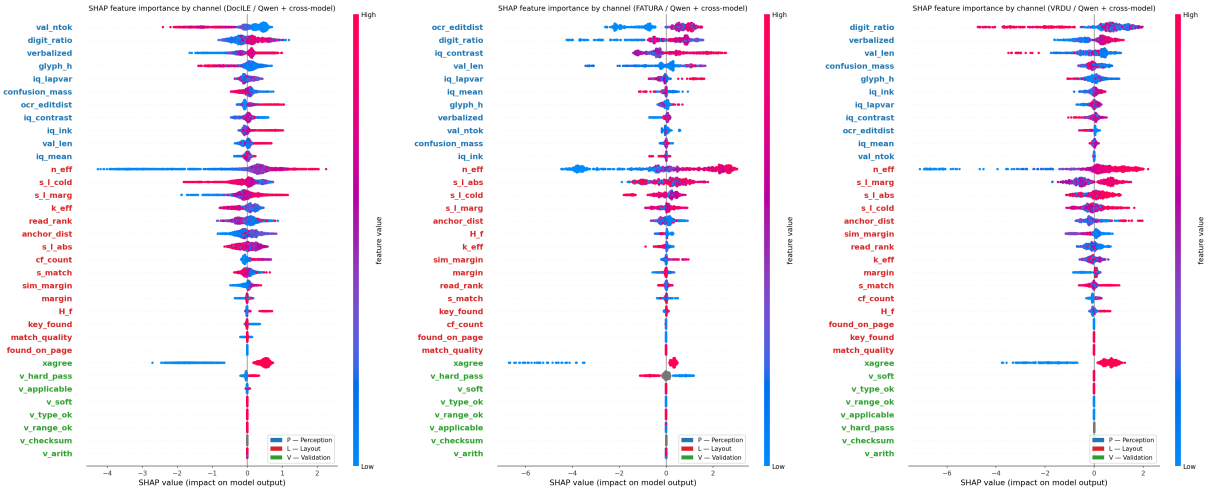
Datasets. We evaluate three public document collections spanning real invoices DocILE (Šimsa et al., 2023), synthetic invoices FATURA (Limam et al., 2023), and advertising order forms VRDU (Wang et al., 2023). Each contributes a subset of fields unambiguously under a shared taxonomy (19, 14, and 9 types respectively). The offline layout index is built for each dataset from the training split and is disjoint from the evaluation documents. The details are given in Appendix C.

VLM extractors. To test that the confidence layer is model-agnostic, we run every experiment with two independent VLM families, Qwen3.6-27B and Gemini-3.1-Flash-Lite. A value and its verbalized confidence for each field are returned from a single ungrounded prompt over the page image. We query each document K_q=5 times to estimate self-consistency and take the canonical value as the extraction. The full prompts can be found in Appendix E, and the detailed fusion model configurations can be found in Appendix F.

Baselines. We compare against the confidence signals from the VLM alone (i) *self-consistency*: the fraction of the $K_q = 5$ samples agreeing with the modal value; and (ii) a *VLM baseline*: the native verbalized confidence score. We report our full end-to-end pipeline based on feature engineering from three channels and the fusion layer.

4.2 Correctness discrimination

Table 1 reports AUROC for predicting whether an extracted value is correct. As baselines, the VLM's own confidence signals are weak: self-consistency and the VLM baseline reach only 0.54–0.74 across the six dataset–model settings. Notably, self-consistency is not stronger than the single-call VLM baseline despite costing K_q calls. This is because sampling agreement detects only stochastic errors, where resampling changes the answer, but is blind to systematic ones, where the same wrong value is returned every time. In practice, many VLM extraction errors are systematic, so high agreement reflects determinism, not correctness. In contrast, adding the P/L/V channels together with the fusion layer considerably raises AUROC to 0.91–0.99, which holds across all dataset and VLM settings, confirming the pipeline is largely model-agnostic. Moreover, cross-model agreement adds a further increment when both VLMs are available. The lift is largest on FATURA at 0.997, whose clean synthetic templates make layout and OCR cues highly reliable, and is consistent on the other two real-document collections at 0.91–0.93.



(a) DocILE (b) FATURA (c) VRDU

Figure 3: SHAP feature importance by channel (Qwen extractor + cross-model) on the three datasets. Labels are colored by channel (P blue, L red, V green). The channel emphasis shifts slightly with document type: FATURA is perception/OCR-driven, VRDU is layout-driven, DocILE is in between.

4.3 What the fusion model learns

Figure 3 shows SHAP attributions for the fused model, grouped and colored by channel. Three patterns hold across all datasets. First, the retrieval prior is the strongest signal. The effective-neighbour count `n_eff` has the widest impact in every panel, with a long low-value tail that correctly lowers confidence at cold start. And the retrieval-conditioned spatial scores (`s_l_marg`, `s_l_cold`) and anchor distance (`anchor_dist`) form a consistent supporting cluster. Second, cross-model agreement (`xagree`) is a clean and high-impact validator that agreement raises confidence and disagreement lowers it. Third, within perception the value-string and OCR cues do most of the work using token count (`val_ntok`), digit ratio (`digit_ratio`), glyph height (`glyph_h`), and VLM–OCR edit distance (`ocr_editdist`). While the VLM's own verbalized confidence ranks only mid-pack, consistent with its weakness as a standalone signal.

Aggregated by channel, layout, and perception dominate (49% and 39% of overall attribution), with validation contributing the rest. The channel emphasis shifts with document type. FATURA is perception/OCR-driven because of its clean and synthetic templates, VRDU is layout-driven, and DocILE is in between. The top signals are stable across all datasets and VLM families.

The interpretability of the layer is a deployment prerequisite in a regulated financial setting. An automated approval that cannot be explained is, in practice, often not acceptable. Model-risk management and audit functions require that every straight-through decision be traceable to a justifiable cause, where a monolithic neural confidence score is insufficient. Because our fusion model is tabular and its inputs are organized along the interpretable perception, layout, and validation channels, every accept/reject decision decomposes via SHAP into signed per-feature contributions that read as human-legible. A field would be routed to review because its value has little retrieval support (e.g., an unfamiliar template), or the two VLMs disagree, or it sits an unusual distance from its label anchor, or the VLM reading diverges from the OCR text. These inspections are exactly the artifacts a model validation or regulator review asks for, allowing a human reviewer to either trust or override the gate on a case-by-case basis. The confidence layer is therefore not merely a score but an *auditable control* in a compliance-bound environment.

4.4 Straight-through-processing coverage

Table 2 shows the deployment-facing result on the percentage of fields that can be auto-approved while holding the accepted-tier error at or below a target. This is the operative metric in the real regulatory setting where straight-through processing is precision-first. Specifically, a wrongly auto-approved value is a financial or compliance error, whereas a correct value sent to review is only a small efficiency cost. Consequently, the question is not average accuracy but how much can be auto-approved at a guaranteed precision.

Dataset	Method	$\leq$5% err	$\leq$10% err	$\leq$20% err
DocILE	Self-consistency (Qwen/Gemini)	0.1 / 0.0	0.2 / 0.0	0.5 / 0.3
	VLM baseline (Qwen/Gemini)	0.3 / 0.1	1.6 / 0.1	12.2 / 0.6
	+ P/L/V fusion (Qwen/Gemini)	20.7 / 21.4	39.3 / 42.6	58.0 / 61.1
	+ cross-model	**28.1**	**49.4**	**66.5**
FATURA	Self-consistency (Qwen/Gemini)	0.1 / 0.2	0.1 / 0.2	0.2 / 0.5
	VLM baseline (Qwen/Gemini)	0.1 / 1.4	0.1 / 7.0	0.4 / 18.1
	+ P/L/V fusion (Qwen/Gemini)	62.4 / 66.6	66.3 / 70.5	74.8 / 79.5
	+ cross-model	**68.4**	**72.3**	**81.6**
VRDU	Self-consistency (Qwen/Gemini)	0.2 / 0.2	0.4 / 0.3	1.4 / 0.9
	VLM baseline (Qwen/Gemini)	1.5 / 0.1	2.9 / 0.1	10.4 / 0.2
	+ P/L/V fusion (Qwen/Gemini)	25.7 / 35.9	60.5 / 66.9	78.7 / 79.7
	+ cross-model	**41.3**	**70.8**	**83.8**

Table 2: Straight-through-processing coverage, in terms of the percentage of fields that can be auto-approved while holding the accepted-tier error at or below the target. Native VLM signals clear almost no fields, while the channel fusion and cross-model selection unlock up to 84% when $\leq 20\%$ error.

In particular, at a 10% error target ($\geq 90\%$ precision) self-consistency and the VLM baseline auto-approve 0.1–7.0% of fields. Whereas the P/L/V fusion auto-approves 39–71% at the same target, which is a one to two orders of magnitude increase, driven collaboratively by all three channels (Section 4.3). Moreover, cross-model selection adds a premium tier (49–72%, up to 84% at a 20% target), most visibly on VRDU, where resolving disagreements rather than rejecting them lifts coverage from 61–67% to 71%. The gains hold across both VLM families. Even DocILE, the real and heterogeneous invoices, reaches 49% auto-approval at $\geq 90\%$ precision from a near-zero baseline.

Appendix G gives qualitative case studies to make the STP concrete. In each case a VLM returns a wrong value with high verbalized confidence (≥ 0.90), yet our pipeline scores it low and cross-model selection recovers the correct value.

5 Conclusion

We presented a decomposed confidence layer for straight-through processing of VLM-extracted fields from financial documents. It scores each field along three interpretable channels (perception, layout, and validation) with a retrieval-conditioned layout prior. A LightGBM model fuses the channels into a calibrated probability with per-field reason codes and a Learn-Then-Test conformal gate turns it into an auto-approval decision with a distribution-free error bound. Across three public datasets and two VLM families, the layer raises correctness-discrimination AUROC from 0.54-0.74 for native VLM confidence to 0.90–0.99, and auto-approves 49–72% of fields at a 10% error target where native signals clear almost none. The gains hold across both VLMs showing its model-agnostic. A cross-model selection step resolves disagreements from two VLMs to further lift guaranteed throughput.

The limitations are mild rather than fundamental in target banking settings. First, the layout channel is weakest at cold start. For a genuinely novel template with no similar layouts in the index, the retrieval-conditioned prior collapses to the global field prior and adds little signal. However, in the real operation, the document stream is dominated by recurring templates (e.g., the same forms and counter-parties processed repeatedly). A truly new template is rare, and it degrades gracefully to other channels and, under the conformal gate, is routed to human review rather than mis-approved. Each such document then enters the index, so a new template becomes familiar after only a few occurrences. Second, fitting the fusion model and calibrating the gate require field-level correctness labels. However, financial institutions usually have large volumes of approved historical documents, providing enough supervision. Moreover, mandatory sample-checking of processed items (including auto-approved STP cases) yields a continual stream of audited field-level labels for periodic refresh. Finally, the rule validators add little marginal signal on the reported dataset, where errors are well-formed and no check-summable identifiers appear. Their value should grow on documents such as trade orders containing IBAN / ISIN / BIC.

References

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. 2024. Theoretical foundations of con-

formal prediction. *arXiv preprint arXiv:2411.11824*.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. 2025. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662.

Google DeepMind. 2026. Gemini 3.1 flash-lite: Model card.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 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, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Nitesh Kumar. 2026. Beyond logprobs: A multi-signal confidence engine for llm-based document field extraction. *arXiv preprint arXiv:2606.24420*.

Mahmoud Limam, Marwa Dhiaf, and Yousri Kessentini. 2023. Fatura: A multi-layout invoice image dataset for document analysis and understanding. *arXiv preprint arXiv:2311.11856*.

Kevin P Murphy. 2007. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27b dense model.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and 1 others. 2023. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Springer.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2021. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 2579–2591.

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 international conference on knowledge discovery & data mining*, pages 1192–1200.

A Canonicalization Rules

The canonicalizer canon_f defines both candidate matching and the ground-truth correctness label in section 3. An extraction $\hat{v}$ is correct iff it matches the gold value $v^\star$ under the per-category comparison below. We fix the ruleset at version `canon-v2` and treat it as a versioned artifact.

Field categorization. Each field type is routed to one category below by substring rules on its name, tested in this order: date -> number -> identification. All others default to string.

Dates. We avoid full date parsing. We extract every run of digits and form the set of numeric components. A two-digit value $v \leq 99$ contributes both itself and its century expansion ($2000{+}v$ if $v < 50$, else $1900{+}v$). Month abbreviations (`JAN-DEC`) contribute their month index. Two dates match if their component sets overlap *and* either one set is a subset of the other or they share at least two components. This order-agnostic rule is robust across the many formats in the data such as `3/4/2020`, `06/07/99`, `DEC31/01`, `JUL20/02`.

Numbers. We strip percent signs and any character outside {`0-9`, `.`, `,`, `-`}. The decimal separator is taken to be the rightmost `.` or `,` followed by one or two digits, all other separators are removed. The remainder is parsed as a signed float. Two numbers

Feature	Description
Perception Channel — 11 features	
verbalized	*self-report.* VLM verbalized confidence s_f, rescaled to $[0, 1]$.
iq_lapvar	*pixels.* Laplacian variance of the value crop (sharpness; low = blurry).
iq_contrast	*pixels.* Pixel contrast of the value crop.
iq_ink	*pixels.* Ink density (dark-pixel fraction) of the crop.
iq_mean	*pixels.* Mean gray level of the crop.
glyph_h	*pixels.* OCR text-box height in pixels (small = tiny/low-resolution print).
ocr_editdist	*OCR.* $1-$ string similarity of v_f to the OCR text at the location (VLM–OCR disagreement).
val_len	*string.* Character length of the value string.
val_ntok	*string.* Whitespace-token count of the value string.
confusion_mass	*string.* Fraction of chars in the confusable-glyph set ($0/O$, $1/l/I$, $5/S$, $8/B$, $./,$).
digit_ratio	*string.* Fraction of value characters that are digits.
Layout Channel — 15 features	
found_on_page	*grounding.* 1 if the value is grounded on the page ($\|C_f\| > 0$), else 0.
cf_count	*grounding.* Number of candidate occurrences $\|C_f\|$.
key_found	*grounding.* 1 if a label anchor sits beside the value.
match_quality	*grounding.* Token coverage of the value by its OCR span (strings). 1 for exact-canon id/date/number.
s_l_marg	*spatial.* Marginal (log-mean-exp over occurrences) spatial log-likelihood — the core score s_L.
s_l_cold	*spatial.* Cold-start spatial score: global field prior with retrieval off.
s_l_abs	*spatial.* Spatial score from absolute page position (weak baseline).
anchor_dist	*geometry.* Distance to the nearest label anchor z.
read_rank	*geometry.* Vertical reading-order position of the value.
s_match	*retrieval.* Top retrieval similarity to a historical layout (template familiarity).
sim_margin	*retrieval.* Top-1 minus top-2 retrieval similarity.
k_eff	*retrieval.* Soft effective neighbour count from the retrieval weights.
n_eff	*retrieval.* Effective neighbours contributing to the spatial prior.
H_f	*ambiguity.* Entropy of the occurrence distribution π_c (slot ambiguity).
margin	*ambiguity.* Top-two margin of π_c (large = unambiguous).
Validation Channel — 8 features	
v_type_ok	*rule.* 1 if the value has valid type/format for the field.
v_range_ok	*rule.* 1 if the value lies in a plausible range.
v_soft	*rule.* Soft aggregate v_type_ok $\times$ v_range_ok (always defined).
v_checksum	*rule.* Checksum result where the field admits one (IBAN, ISIN); NaN otherwise.
v_arith	*rule.* Arithmetic consistency (net + tax = gross; items vs. total); NaN if inapplicable.
v_applicable	*rule.* 1 if any hard validator (checksum or arithmetic) applies.
v_hard_pass	*rule.* 1 if all applicable hard validators pass; NaN if none apply.
xagree	*cross-model.* 1 if a second, independent VLM returns the same canonical value, else 0.

Table 3: Feature dictionary for the fusion model, grouped by channel. Each entry leads with its sub-type (in italics). NaN rule features indicate the validator does not apply to that field (e.g. no check-summable identifier on an invoice).

match under a strict *absolute* tolerance of half a cent, $|a - b| \leq 0.005$.

Identifiers and Strings. We strip accents (Unicode NFKD), lowercase, and remove all non-alphanumeric characters. Two identifiers match on exact equality of this normalized form.

B Feature Dictionary

The fusion model consumes 34 features across the three channels. Table 3 lists all the features and the meanings. Features are computed at the best-scoring occurrence (Section 3). When the value is not grounded ($|C_f|$=0), layout and pixel features fall back to fixed defaults and only the value-string and self-report features remain informative.

C Dataset Statistics

Table 4 reports the statistics for each dataset and VLM. The *offline layout index* is built once per dataset from its training split (*Indexed Layouts*), and is disjoint from the *evaluation* documents (*Eval*). The reporting unit is a field extraction (one row per (document, field)). Row counts differ slightly between the two VLMs on the same documents because grounding recovery depends on what each model extracted. DocILE uses its official validation split with template-familiarity buckets from train-cluster recurrence. FATURA and VRDU use 200-document evaluation samples.

The three datasets are evaluated under a shared field taxonomy, with each dataset contributing the subset of field types it annotates unambiguously.

Dataset	Indexed Layouts	Eval Doc	Field Types	VLM	Fields	Correct	Grounded
DocILE	5,180	500	19	Qwen3.6-27B	5,141	51.7%	85.4%
				Gemini-3.1-Flash-Lite	5,077	53.8%	86.8%
FATURA	731	200	14	Qwen3.6-27B	1,850	60.2%	95.0%
				Gemini-3.1-Flash-Lite	1,767	63.9%	98.4%
VRDU	417	200	9	Qwen3.6-27B	1,698	63.3%	97.9%
				Gemini-3.1-Flash-Lite	1,666	63.7%	98.0%

Table 4: Dataset statistics. *Indexed Layouts* shows the unique number of training documents in the offline index. *Eval Doc* is the number of evaluation documents. *Types* is the number of distinct field types. *Fields* is the number of labeled field extractions. *Correct* is the fraction whose canonical value matches gold. *Grounded* is the fraction recovered on the page ($|C_f| > 0$). Note that the two VLMs extract slightly different fields.

DocILE (Šimsa et al., 2023) uses the full set of 19 single-value field types, including

- identifiers (document_id, order_id, customer_id)
- dates (date_issue, date_due)
- monetary amounts (amount_total_gross, amount_total_net, amount_due, currency_code_amount_due, amount_paid, amount_total_tax)
- vendor fields (vendor_name, vendor_address, vendor_tax_id, vendor_email)
- customer fields (customer_billing_name, customer_billing_address)
- payment fields (payment_terms, payment_reference).

FATURA (Limam et al., 2023) contains synthetic invoices annotating 24 classes. We map them into 14 with an unambiguous single-value analogue onto the shared taxonomy, by collapsing rate-specific tax classes and duplicate invoice-number classes. Layout regions (e.g., "table", "logo") and free-text blocks (e.g., "note", "payment_details") have no clean field analogue and thus are excluded.

VRDU (Wang et al., 2023) contains real advertising order forms from a different domain. So its 9 fields are mapped by role onto the shared taxonomy. In particular, the "contract number" to "document_id", the "product/estimate number" to "order_id", the "campaign flight start and end" to "date_issue" and "date_due", the "gross total" to "amount_total_gross", the "TV station" to "vendor_name/address", the advertiser to "customer_billing_name", and the "agency" to "customer_other_name".

D Layout Descriptor

The layout descriptor ϕ encodes a page's coarse text-occupancy geometry, independent of the text content. We overlay a $G \times G$ grid (G=32) on the page in normalized coordinates and mark every cell that any OCR token box overlaps, giving a binary occupancy mask; the mask is flattened to a G^2=1024-dimensional vector and L2-normalized. Two pages are compared by the inner product of their descriptors, which equals cosine similarity after normalization, so the FAISS inner-product index returns the layouts whose text falls in the same places. One descriptor is computed per document from all its OCR words (across pages).

The descriptor is deliberately coarse and content-blind. It fingerprints *where* fields sit on the page, not what they say. This is exactly the invariance retrieval needs. Two invoices from the same vendor template map to nearby descriptors regardless of their amounts, dates, or names, so the spatial prior for a field can borrow strength across all documents that share its layout. A cell is set whenever a token box overlaps it, with an inclusive end cell, so a text line thinner than one grid cell still marks its cell rather than being dropped by an exclusive slice. Listing 1 gives the details of the implementation.

Listing 1: Layout descriptor ϕ is a binary $G \times G$ OCR-token occupancy mask, flattened and L2-normalized. Word boxes are in normalized page coordinates.

```
def token_mask_descriptor(words, grid=32):
    """Binary grid x grid OCR-token occupancy
     mask, L2-normalized."""
    m = np.zeros((grid, grid), np.float32)
    for w in words:
        l, t, r, b = w.bbox
        l = min(max(l, 0.0), 1.0)
        r = min(max(r, 0.0), 1.0)
        t = min(max(t, 0.0), 1.0)
        b = min(max(b, 0.0), 1.0)
        x0 = min(int(l * grid), grid - 1)
        x1 = min(int(r * grid), grid - 1)
        y0 = min(int(t * grid), grid - 1)
        y1 = min(int(b * grid), grid - 1)
        # cells the token overlaps
        m[y0:y1 + 1, x0:x1 + 1] = 1.0

    # flatten to grid * grid
```

```
    v = m.reshape(-1)
    n = np.linalg.norm(v)
    if n > 0:
        v = v / n
    return v.astype(np.float32)
```

At query time, a descriptor is issued against the index to retrieve the K_{NN}=50 most similar layouts. Their similarities become the retrieval weights that condition the spatial prior. The same encoder is used offline to build the index and online to form the query, so familiarity is measured consistently.

E VLM Extraction Prompt

Each document is extracted in a single user message: the page image followed by the following instruction. The model returns strict JSON with a value and verbalized confidence per present field, omitting absent ones. The same template is used for both VLMs. Only a per-family "thinking" toggle differs (Qwen has `enable_thinking=false` and Gemini has `reasoning_effort=none`). The fields listed below are the DocILE schema. FATURA and VRDU use the same template with their own mapped fields. The amount rules are appended to resolve the common net/gross/due confusion.

Listing 2: The VLM extraction prompt with DocILE field schema as the example.

```
Extract the following fields from this business
   document image.
Return STRICT JSON only, no prose. For each
  field that is PRESENT on the page, give an
  object {"value": <exact text as printed>, "
  confidence": <integer 0-100>}. OMIT fields
  that are absent.
Fields:
  - document_id: the invoice/document number
  - order_id: the purchase order number
  - date_issue: the issue/invoice date
  - date_due: the payment due date
  - amount_total_gross: grand total INCLUDING
  tax (equals net + tax)
  - amount_total_net: subtotal BEFORE tax (net
  of tax)
  - amount_total_tax: the TAX/VAT amount only (
  not a total)
  - amount_due: amount actually payable now (
  often equals gross)
  - amount_paid: amount already paid
  - currency_code_amount_due: currency symbol
  or ISO code
  - vendor_name: the seller/vendor company name
  - vendor_address: the full vendor postal
  address
  - vendor_tax_id: vendor tax / VAT id
  - vendor_email: vendor email address
  - customer_billing_name: the buyer/customer
  billing name
  - customer_billing_address: the full customer
   billing address
  - customer_id: the customer account id
  - payment_terms: payment terms (e.g. NET 30)
  - payment_reference: payment reference /
  variable symbol
Respond as: {"document_id": {"value": "...", "
  confidence": 90}, ...}

Amount rules: net = subtotal before tax; tax =
  the tax amount only; gross = grand total
  including tax (gross = net + tax); due =
  amount payable now. If the page shows a
  single total, use it for BOTH gross and due.
  Report each amount exactly as printed.
```

Hyperparameter	Value
objective	binary log-loss
boosting	gbdt
n_estimators	200
learning_rate	0.08
num_leaves	15
reg_lambda L2 regularization	1.0
min_child_samples	20
random_state	0

Table 5: LightGBM fusion-model configuration.

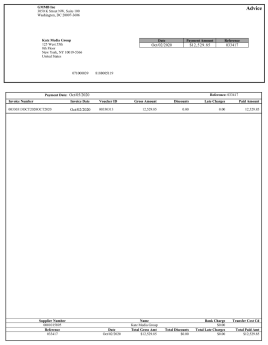
Sampling. Each document is queried K_q=5 times. The first sample is greedy (temperature 0) and fixes the reported value, the remaining four are sampled (temperature 0.7) to estimate self-consistency. The reported value is the modal canonical value across samples, the verbalized confidence s_f is the mean self-reported confidence, and the sampling-agreement fraction is the self-consistency baseline. The deployed operating point is the single greedy call. The additional samples are used only to compute the self-consistency baseline.

F Fusion model configuration

The fusion model is a gradient-boosted decision tree (`LightGBM.LGBMClassifier`) with the binary log-loss objective. Table 5 lists the full configuration. Regularization is deliberately conservative for the small tabular problem (∼800–2600 training rows per split), shallow trees (15 leaves), an L2 penalty, a minimum of 20 samples per leaf and no row or feature subsampling. The model is retrained from scratch on each document-level split.

G Qualitative Case Studies

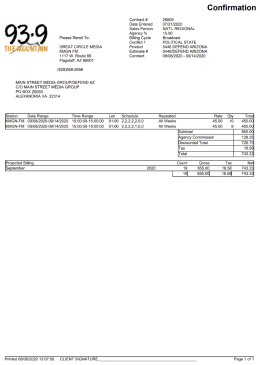
We show four cases where one VLM extracts a wrong value with high verbalized confidence, but the confidence layer assigns it a low probability. In every case the wrong and right values have nearly indistinguishable verbalized confidence (0.90–0.998), yet the fused score separates



Customer_Billing_Name

	Gemini-3.1-Flash	QWen3.6-27B
VLM-extraction	GMMB INC	Katz Media Group
VLM-score	0.988 ✗	0.950 ✓
Our score	0.145 ✓	0.953 ✓
Ground Truth	Katz Media Group	

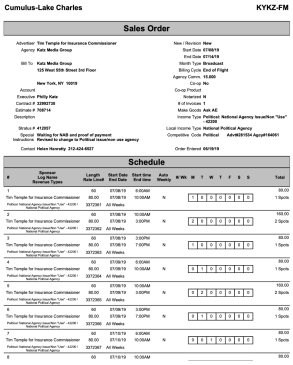
(a) Sender vs. recipient confusion (`customer_billing_name`): fused 0.145 vs. verbalized 0.953 on Gemini result.



Date_Issue

	Gemini-3.1-Flash	QWen3.6-27B
VLM-extraction	08/06/2020	07/21/2020
VLM-score	0.995 ✗	0.940 ✓
Our score	0.083 ✓	0.943 ✓
Ground Truth	07/21/2020	

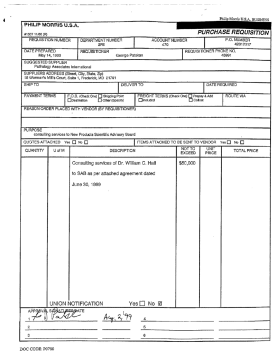
(b) Print-timestamp vs. issue date (`date_issue`): fused 0.083 vs. verbalized 0.943 on Gemini result.



Vendor_Name

	Gemini-3.1-Flash	QWen3.6-27B
VLM-extraction	Cumulus-Lake Charles	Cumulus-Lake Charles KYKZ-FM
VLM-score	0.986 ✓	0.930 ✗
Our score	0.954 ✓	0.178 ✓
Ground Truth	Cumulus-Lake Charles	

(c) Call-sign over-inclusion (`vendor_name`): fused 0.178 vs. verbalized 0.954 on Qwen result.



Vendor_Name

	Gemini-3.1-Flash	QWen3.6-27B
VLM-extraction	PHILIP MORRIS U.S.A.	George Patskan
VLM-score	0.998 ✓	0.900 ✗
Our score	0.803 ✓	0.351 ✓
Ground Truth	PHILIP MORRIS U.S.A.	

(d) Person vs. company (`vendor_name`): fused 0.351 vs. verbalized 0.803 on Qwen result.

Figure 4: Qualitative case studies. In each, one VLM is confidently wrong (verbalized ≥ 0.90) and the other is correct. Our fused score separates them (wrong value listed first). The confidently-wrong extraction is assigned a low fused probability in every case, while the correct value scores high.

them cleanly. This is the qualitative counterpart of the aggregate AUROC gap. It also shows why cross-model selection works by keeping the value our layer scores higher to recover the correct answer in all four cases. Each case illustrates a different error mode.

Case 1. Sender vs. recipient confusion in Figure 4a on `customer_billing_name` field. On a payment-advice document, Gemini returns the paying party "GMMB Inc" (verbalized 0.988) instead of the billed customer "Katz Media Group" (gold). Qwen returns the correct value (0.950). Verbalized confidence cannot tell them apart, but our layer scores Gemini's value 0.145 and Qwen's 0.953: the sender's name sits in the letterhead block, where a billing-name field rarely lands.

Case 2. Wrong date field as shown in Figure 4b on `date_issue` extraction. On a broadcast confirmation, Gemini returns "08/06/2020" (verbalized 0.995), which is the footer *print* timestamp, not the "Date Entered" issue date "07/21/2020" (gold). The reading is crisp and unambiguous, so perception signals do not help; the layout channel penalizes the value because a footer position is far from where an issue date normally appears, scoring it 0.083 against Qwen's 0.943.

Case 3. Call-sign over-inclusion in Figure 4c on `vendor_name`. Qwen returns "Cumulus-Lake Charles KYKZ-FM" concatenating the station call sign onto the vendor name (verbalized 0.930), while gold is "Cumulus-Lake Charles" aligned by Gemini. Our layer scores Qwen's over-inclusive value 0.178 and Gemini's value 0.954, driven by the span/OCR-coverage and value-string cues.

Case 4. Person vs. company in Figure 4d on `vendor_name`. On a purchase requisition, Qwen returns the requisitioner person name "George Patskan" (verbalized 0.900) instead of the vendor company "PHILIP MORRIS U.S.A." (gold). Our layer scores Qwen's 0.351 and Gemini's 0.803, where the requisitioner name sits in an internal form field, not where a vendor name is expected.