The Confidence Paradox: Can LLM Know When It's Wrong?

Anonymous ACL submission

Abstract

Document Visual Question Answering (DocVQA) systems often produce overconfident or ethically misaligned responses, especially under uncertainty. Existing models like LayoutLMv3, UDOP, and DONUT focus on accuracy but lack ethical calibration. We propose **HonestVQA**, a model-agnostic, self-supervised framework that aligns model confidence with correctness using weighted loss and contrastive learning. We introduce two new metrics—Honesty Score (H-Score) and Ethical Confidence Index (ECI)—to evaluate ethical alignment. **HonestVQA** improves accuracy and F1 by up to 4.3% across SpDocVQA, InfographicsVQA, and SROIE, while reducing overconfidence. It also generalizes well across domains, achieving 78.9% accuracy and 76.1% F1-score. Our code is available at: https://anonymous.4open.science/r/HonestVQA-B454/README.md

1 Introduction

Document Visual Question Answering (DocVQA) has emerged as a key challenge in multimodal AI (Wang et al., 2025), enabling systems to answer questions based on visual and textual content in documents such as invoices, forms, contracts, and academic papers. These systems are widely deployed in enterprise automation (Jiang et al., 2024), legal analysis (Liu et al., 2024), and assistive technologies (Zeng et al., 2024). However, despite their growing utility, DocVQA systems often lack ethical transparency—frequently returning confidently incorrect answers to ambiguous, adversarial, or under-specified queries. For instance, a system may assert the total invoice amount with high confidence even when the relevant table is partially occluded, or confidently misinterpret a scanned signature line as a date. Such failures can propagate serious downstream consequences, including legal misinterpretation, misinformation, or financial misjudgment.

However, the crux of the problem lies in the inability of existing DocVQA models to communicate uncertainty in a calibrated, ethically responsible manner. While State-of-the-Art (SOTA) systems such as LayoutLMv3[1] (Fujitake, 2024), UDOP[2] (Wang et al., 2023a), and DONUT[3] (Li et al., 2024) focus on improving accuracy through sophisticated architecture and pretraining strategies, they fall short in aligning model confidence with actual knowledge. LayoutLMv3 (Fujitake, 2024) tends to prioritize exact answers over conveying doubt, UDOP (Wang et al., 2023a) frequently errs on the side of over-caution without actionable explanations, and DONUT (Li et al., 2024) offers no uncertainty estimation at all—leading to ethically untrustworthy behavior in ambiguous scenarios. Therefore, recent advances in AI alignment research have emphasized the importance of ethical calibration (Rao et al., 2023), including honesty (Yang et al., 2024), confidence-awareness (Stangel et al., 2025), and transparent failure modes (Stewart et al., 2023). However, these insights have yet to be meaningfully integrated into DocVQA systems. To address these critical gaps, we propose **HonestVQA**, a self-supervised framework that calibrates model confidence to reflect its underlying knowledge and ethical responsibility. Our approach is model-agnostic and integrates three key components: (1) uncertainty quantification to identify knowledge gaps, (2) confidence-accuracy alignment through weighted loss optimization, and (3) contrastive learning to enforce ethical response boundaries in ambiguous contexts. We also introduce two novel evaluation metrics: i) Honesty Score (H-Score), which captures the alignment between confidence and correctness, and ii) Ethical

[1] https://huggingface.co/microsoft/layoutlmv3-base

[2] https://huggingface.co/microsoft/udop-large

[3] https://huggingface.co/naver-clova-ix/donut-base

Confidence Index (ECI), which evaluates whether high-confidence answers are ethically warranted.

2 Related Work

Recent research has increasingly focused on improving the reliability and interpretability of AI systems, especially in high-stakes domains. While core DocVQA models like have been discussed in Section 1, here we focus on complementary areas that our framework draws from—confidence calibration, ethical modeling, and contrastive learning. Confidence calibration techniques such as temperature scaling (Xie et al., 2024) and label smoothing (Müller et al., 2019) aim to align predicted probabilities with empirical accuracies. However, these methods are typically post-hoc and task-agnostic, often failing to generalize in multimodal settings. Selective prediction frameworks such as (Chen et al., 2023) allow models to abstain from uncertain answers, but they usually rely on fixed thresholds and lack principled mechanisms to model epistemic uncertainty in visually grounded tasks like DocVQA. However, in the area of ethical and honest AI, efforts such as instruction tuning for alignment (Zhang et al., 2023) and calibrated language modeling (Zhu et al., 2023) emphasize epistemic humility—training models to express uncertainty when appropriate. However, these approaches are primarily developed for language-only models and remain underexplored in multimodal tasks involving structured visual data. Whereas, contrastive learning has shown strong performance in aligning multimodal representations, with frameworks like CLIP (Gao et al., 2024) and ALIGN (Wang et al., 2023b) leveraging contrastive objectives for image-text alignment. While effective at learning generalizable embeddings, such methods are not designed to enforce ethical boundaries or distinguish between honest and overconfident outputs in ambiguous scenarios.

3 Methodology

As discussed earlier, **HonestVQA** is a model-agnostic calibration framework designed to enhance ethical transparency in DocVQA systems. It operates as a wrapper around pretrained DocVQA models (in our work, we evaluate our framework on top of pretrained models such as LayoutLMv3 (Fujitake, 2024), UDOP (Wang et al., 2023a), and DONUT (Li et al., 2024) to demonstrate its generalizability), injecting uncertainty-aware alignment and contrastive reasoning to reduce overconfident yet incorrect outputs. The broader process of the **HonestVQA** is illustrated in Algorithm 1.

Algorithm 1 HonestVQA Training Algorithm

Require: Pretrained model f_θ, input (D, Q, y^*), thresholds δ, τ_1, τ_2, weights α, β, m, λ_1, λ_2
Ensure: Calibrated DocVQA wrapper

Compute $P(y \mid D, Q) \leftarrow f_\theta(D, Q)$
Compute confidence $\mathcal{C} = \max_i P(y_i)$ and entropy $\mathcal{U} = -\sum_i P(y_i) \log P(y_i)$
Predict $\hat{y} \leftarrow \arg\max_y P(y)$
$\mathcal{L}_{\text{align}} \leftarrow \alpha \cdot \mathbb{1}[\hat{y} \neq y^*] \cdot \mathcal{C} + \beta \cdot \text{CE}(\hat{y}, y^*)$
if $\text{WMD}(\hat{y}, y^*) < \delta$ **then**
    $h_{\text{pos}} \leftarrow \text{Embed}(\hat{y})$
end if
if $\hat{y} \neq y^* \wedge \mathcal{C} > \tau_1 \wedge \mathcal{U} < \tau_2$ **then**
    $h_{\text{neg}} \leftarrow \text{Embed}(\hat{y})$
end if
Compute $\mathcal{L}_{\text{contrast}} = \max(0, m - \text{sim}(h_{\text{anchor}}, h_{\text{pos}}) + \text{sim}(h_{\text{anchor}}, h_{\text{neg}}))$
$\mathcal{L}_{\text{total}} \leftarrow \lambda_1 \cdot \mathcal{L}_{\text{align}} + \lambda_2 \cdot \mathcal{L}_{\text{contrast}}$
Update projection head using $\mathcal{L}_{\text{total}}$

3.1 Uncertainty Quantification Module

Given a document D and a question Q, we use a pretrained DocVQA model f_θ that maps (D, Q) to an answer distribution $P(y \mid D, Q; \theta)$. To quantify the model's epistemic uncertainty, we compute the softmax entropy of the output distribution according to Equation (1).

$$\mathcal{U}(D, Q) = -\sum_{i=1}^{|Y|} P(y_i \mid D, Q) \log P(y_i \mid D, Q) \tag{1}$$

Here, $|Y|$ denotes the size of the answer space. Higher entropy corresponds to greater uncertainty. We also define a maximum-confidence score as shown in Equation (2).

$$\mathcal{C}(D, Q) = \max_i P(y_i \mid D, Q) \tag{2}$$

This dual view captures both dispersion and peakiness in the output distribution. Following recent work (Pearce et al., 2021), we identify overconfident failure cases as those where $\mathcal{C}(D, Q)$ is high despite $\mathcal{U}(D, Q)$ being non-negligible. These metrics are computed during training and inference, and $\mathcal{U}(D, Q)$ serves as a routing signal for sampling in the contrastive module, though it is not explicitly penalized.

3.2 Confidence-Accuracy Alignment Module

To align the model's confidence with its accuracy, we introduce a calibration-aware loss that penalizes incorrect predictions more strongly when made with high confidence. Let $\hat{y}$ denote the predicted answer and y^* the ground truth. We define the alignment loss according to Equation (2).

$$\mathcal{L}_{\text{align}} = \alpha \cdot \mathbb{1}[\hat{y} \neq y^*] \cdot \mathcal{C}(D, Q) + \beta \cdot \text{CE}(\hat{y}, y^*) \quad (2)$$

Here, CE is the standard cross-entropy loss. Hyperparameters α and β control the influence of confidence-penalization and prediction error, respectively.

3.3 Contrastive Ethical Enforcement Module

To further refine the model's response space under ambiguity, we introduce a contrastive loss that structurally separates ethically misaligned or misleading answers from calibrated, semantically valid responses. Given a query-answer embedding h_{anchor}, we identify a positive sample h_{pos} (semantically similar and ethically aligned) and a negative sample h_{neg} (incorrect, overconfident, or potentially misleading). The contrastive loss is defined as according to Equation (3).

$$\mathcal{L}_{\text{contrast}} = \max\Big(0,\ m - \text{sim}(h_{\text{anchor}}, h_{\text{pos}}) + \text{sim}(h_{\text{anchor}}, h_{\text{neg}})\Big) \quad (3)$$

where $\text{sim}(\cdot)$ denotes cosine similarity, and m is a margin hyperparameter. We use a projection head atop the DocVQA encoder to map answer embeddings into a low-dimensional calibrated honesty space. Where, positive pairs are identified using a combination of Word Mover's Distance (WMD) and agreement with ground truth as shown in Equations (4), and (5), where δ is a tunable similarity threshold.

$$\text{WMD}(\hat{y}, y^*) < \delta \quad (4)$$

$$\hat{y} \in \mathcal{A}_{\text{aligned}} \implies \text{semantically valid and agrees with ground truth.} \quad (5)$$

Whereas, negative samples are drawn from high-confidence as shown in Equation (6).

$$\hat{y}_{\text{neg}} : \mathbb{1}[\hat{y}_{\text{neg}} \neq y^*] \wedge \mathcal{C}(D, Q) > \tau_1 \wedge \mathcal{U}(D, Q) < \tau_2 \quad (6)$$

where τ_1 and τ_2 are confidence and entropy thresholds, respectively.

3.4 Training Module

The overall training loss combines the alignment and contrastive objectives according to Equation (7).

$$\mathcal{L}_{\text{total}} = \lambda_1 \cdot \mathcal{L}_{\text{align}} + \lambda_2 \cdot \mathcal{L}_{\text{contrast}} \quad (7)$$

Here, λ_1 and λ_2 control the relative weight of each loss term. Training is conducted end-to-end using batches sampled from standard DocVQA datasets, where each sample includes contrastive triplets and confidence-aware labels. The projection head is fine-tuned during training, while the base DocVQA encoder remains frozen. At inference, $\mathcal{C}(D, Q)$ and $\mathcal{U}(D, Q)$ may be used to suppress or abstain from answering under high uncertainty. ***Note:*** In this work, we define ethical calibration as the act of reducing confidently incorrect answers, especially under ambiguity or lack of sufficient visual-textual grounding. **HonestVQA** is not a moral arbiter but a mechanism for promoting caution and transparency in DocVQA behavior.

4 Experimental Setup

4.1 Datasets

We evaluate **HonestVQA** on three diverse and challenging datasets: **`SpDocVQA`** (Mathew et al., 2020), **`InfographicsVQA`** (Mathew et al., 2022), and **`SROIE`**[4]. **`SpDocVQA`** comprises multilingual scanned documents requiring structured comprehension and spatial reasoning. **`InfographicsVQA`** presents visually dense infographic images with complex layouts and multi-modal reasoning demands. **`SROIE`** is an entity-level extraction dataset involving semi-structured receipts, demanding high accuracy and ethical response handling due to potential financial implications. We use the original train/val/test splits and ensure consistent preprocessing across models for fair comparison.

4.2 Evaluation Metrics

To comprehensively assess the performance and ethical alignment of our framework, we employ standard accuracy metrics alongside two novel measures designed to evaluate calibration and honesty. First, we report conventional **`Accuracy`** and **`F1 scores`** (i.e. macro) to quantify answer correctness. To capture the alignment between model confidence and actual correctness, we introduce the **`Honesty Score (H-Score)`**, which penalizes

[4] https://rrc.cvc.uab.es/?ch=17&com=downloads

overconfident incorrect predictions while rewarding calibrated confidence on correct answers. Additionally, the **`Ethical Confidence Index (ECI)`** evaluates the model's ability to appropriately express uncertainty, especially under ambiguous or insufficient information. ***Note:*** In the tables, **bold** values indicate the top-performing scores. ↑ indicates that a high value is preferable, while ↓ indicates that a low value is preferable.

4.2.1 Theoretical Guarantees for Evaluation Metrics

In this section, we provide formal lemmas and proofs to establish the theoretical soundness of the proposed **`Honesty Score (H-Score)`** and **`Ethical Confidence Index (ECI)`**, which measure the alignment between model confidence, accuracy, and ethical transparency in DocVQA.

Lemma 4.1 (Calibration Bound of Honesty Score)**.** *Let $C(D, Q)$ be the confidence score output by a DocVQA model for a given document-question pair (D, Q), and let $A(D, Q) \in \{0, 1\}$ be the corresponding accuracy indicator, where 1 denotes a correct answer and 0 an incorrect one. Assume that $C(D, Q)$ is bounded in $[0, 1]$. Then, define the Honesty Score H as according to Equation (8).*

$$\mathrm{H} = 1 - \mathbb{E}_{(D,Q)\sim\mathcal{D}}\big[|C(D, Q) - A(D, Q)|\big] \quad (8)$$

where $\mathcal{D}$ is the data distribution. Whereas, H upper-bounds the expected calibration error between confidence and accuracy according to Equation (9).

$$\mathbb{E}_{(D,Q)\sim\mathcal{D}}\big[|C(D, Q) - A(D, Q)|\big] = 1 - \mathrm{H} \quad (9)$$

Thus, a higher H implies tighter calibration, indicating fewer overconfident incorrect predictions.

Proof. By definition, calibration error measures the absolute difference between predicted confidence and true correctness. Since $A(D, Q)$ is binary, the expectation of $|C - A|$ captures the average misalignment. Rearranging, $\mathrm{H} = 1 - \mathbb{E}[|C - A|]$. Because $|C - A| \in [0, 1]$, $\mathrm{H} \in [0, 1]$ and is maximized when confidence perfectly matches accuracy. Hence, H is a valid measure of calibration that upper-bounds expected miscalibration. □

Lemma 4.2 (Discriminative Power of Ethical Confidence Index)**.** *Let $\mathcal{C}_{correct}$ and $\mathcal{C}_{incorrect}$ denote the random variables corresponding to confidence scores on correctly and incorrectly answered samples respectively. Then, define the Ethical Confidence Index (ECI) as according to Equation (10) which measures the probability that the model assigns higher confidence to correct answers than to incorrect answers.*

$$\mathrm{ECI} = \mathbb{P}\big(\mathcal{C}_{correct} > \mathcal{C}_{incorrect}\big) \quad (10)$$

If the distributions of $\mathcal{C}_{correct}$ and $\mathcal{C}_{incorrect}$ are well-separated, i.e., there exists $\epsilon > 0$ such that it is defined as according to Equation (11), then $\mathrm{ECI} \geq 1 - \epsilon$ indicating strong ethical confidence discrimination.

$$\mathbb{P}(\mathcal{C}_{correct} \leq \mathcal{C}_{incorrect}) < \epsilon \quad (11)$$

Proof. The ECI corresponds exactly to the Area Under the ROC Curve (AUC) when viewing confidence as a score discriminating correct from incorrect answers. By definition, $\mathrm{ECI} = \mathbb{P}(\mathcal{C}_{\text{correct}} > \mathcal{C}_{\text{incorrect}})$. If the two confidence score distributions have minimal overlap (i.e., are well-separated), the probability of $\mathcal{C}_{\text{correct}} \leq \mathcal{C}_{\text{incorrect}}$ is bounded above by a small ϵ. Hence, $\mathrm{ECI} = 1 - \mathbb{P}(\mathcal{C}_{\text{correct}} \leq \mathcal{C}_{\text{incorrect}}) \geq 1 - \epsilon$. Thus, a high ECI value indicates that the model reliably assigns higher confidence to correct answers, promoting ethical transparency. □

4.3 Hyperparameters

The **HonestVQA** framework employs several key hyperparameters to balance confidence calibration and contrastive learning effectively. We set the confidence penalty weight α to 1.0 and the cross-entropy weight β to 0.5 to emphasize penalizing overconfident incorrect predictions while maintaining prediction accuracy. The contrastive margin m is fixed at 0.3 to enforce a moderate separation between positive and negative embeddings. The alignment and contrastive losses are weighted by $\lambda_1 = 1.0$ and $\lambda_2 = 0.7$, respectively, reflecting a slightly stronger emphasis on alignment. For sample selection in the contrastive module, the WMD threshold δ is set to 0.4, while the confidence and entropy thresholds, τ_1 and τ_2, are chosen as 0.8 and 0.5, respectively, to effectively identify semantically valid positive samples and high-confidence misleading negatives. ***Note:*** We will make the code publicly available post-acceptance.

5 Experimental Analysis

5.1 Comparison with Baselines

We evaluate the effectiveness of **HonestVQA** by measuring both standard answer correctness metrics and calibration-specific metrics to provide

Model	SpDocVQA		InfographicsVQA		SROIE	
	Accuracy (%) ↑	Macro F1 (%) ↑	Accuracy (%) ↑	Macro F1 (%) ↑	Accuracy (%) ↑	Macro F1 (%) ↑
Base Models						
LayoutLMv3 (base) (Fujitake, 2024)	**72.3**	**68.5**	**65.4**	**62.1**	**70.0**	**66.8**
UDOP (base) (Wang et al., 2023a)	69.7	66.1	62.8	60.0	68.2	64.0
DONUT (base) (Li et al., 2024)	70.1	67.0	63.5	60.9	69.0	65.2
With HonestVQA						
LayoutLMv3 (Fujitake, 2024) + HonestVQA	**75.9**	**72.8**	**69.7**	**66.3**	**73.4**	**70.1**
UDOP (Wang et al., 2023a) + HonestVQA	73.2	69.4	67.3	63.8	71.0	67.5
DONUT (Li et al., 2024) + HonestVQA	74.0	70.5	68.0	64.7	72.2	68.8

Table 1: Answer correctness comparison on SpDocVQA, InfographicsVQA, and SROIE datasets

Model	SpDocVQA		InfographicsVQA		SROIE	
	H-Score ↓	ECI ↓	H-Score ↓	ECI ↓	H-Score ↓	ECI ↓
Base Models						
LayoutLMv3 (base) (Fujitake, 2024)	**0.185**	**0.210**	**0.192**	**0.215**	**0.188**	**0.213**
UDOP (base) (Wang et al., 2023a)	0.198	0.224	0.203	0.230	0.200	0.228
DONUT (base) (Li et al., 2024)	0.190	0.218	0.195	0.222	0.192	0.220
With HonestVQA						
LayoutLMv3 (Fujitake, 2024) + HonestVQA	**0.113**	**0.132**	**0.118**	**0.138**	**0.115**	**0.136**
UDOP (Wang et al., 2023a) + HonestVQA	0.127	0.147	0.132	0.153	0.129	0.150
DONUT (Li et al., 2024) + HonestVQA	0.120	0.139	0.125	0.143	0.122	0.141

Table 2: Calibration metrics on SpDocVQA, InfographicsVQA, and SROIE datasets

a comprehensive assessment of the model's performance. Specifically, we compare the base DocVQA models—LayoutLMv3 (Fujitake, 2024), UDOP (Wang et al., 2023a), and DONUT (Li et al., 2024)—with their corresponding versions enhanced by the **HonestVQA** calibration framework. Table 1 presents these results on the three datasets. It is evident that **HonestVQA** consistently improves accuracy by approximately 3 to 4 percentage points and macro F1-score by nearly 4 points across all base models. For instance, LayoutLMv3 (Fujitake, 2024) improves from 72.3% to 75.9% in accuracy and from 68.5% to 72.8% in macro F1-score. Similar trends hold for UDOP (Wang et al., 2023a) and DONUT (Li et al., 2024) models, underscoring the robustness of our approach in enhancing answer correctness. Whereas, Table 2 displays these calibration-specific results for the same set of models and dataset. Notably, the base models exhibit relatively high H-Score and ECI values, indicating frequent instances of unjustified overconfidence. Incorporation of **HonestVQA** substantially lowers these values, with H-Score decreasing by over 35% and ECI by nearly 40% on average. For instance, LayoutLMv3's (Fujitake, 2024) H-Score drops from 0.185 to 0.113, and ECI decreases from 0.210 to 0.132 after calibration. This demonstrates that **HonestVQA** effectively mitigates the risk of misleading the user by suppressing confident but incorrect answers.

5.2 Cross-Domain Generalization Testing

To assess the robustness and generalization ability of **HonestVQA** framework, we conduct a series of cross-domain testing experiments. These experiments evaluate whether the hallucination detection model trained on one dataset can effectively identify hallucinations in DocVQA outputs on different datasets. Such generalization is crucial in real-world scenarios where the distribution of questions and visual-textual content varies significantly across domains such as scanned documents, infographics, and structurally diverse textual scenes. We train **HonestVQA** on the source dataset and evaluate it on a different target dataset without any further fine-tuning, measuring hallucination detection performance using Accuracy and F1-score as the primary metrics. From Table 3, we observe that **HonestVQA** generalizes robustly across domain shifts. Notably, when trained on InfographicsVQA and evaluated on SpDocVQA, the model achieves a high F1-score of 76.1%, outperforming the reverse setting (SpDocVQA → InfographicsVQA), which yields 71.8%. This suggests that the high-density, information-rich visual patterns in InfographicsVQA provide transferable inductive

Train Domain	Test Domain	Model	Accuracy (%) ↑	Macro F1 (%) ↑
SpDocVQA	InfographicsVQA	HonestVQA	74.2	71.8
InfographicsVQA	SpDocVQA	HonestVQA	**78.9**	**76.1**
SROIE	InfographicsVQA	HonestVQA	70.5	67.2
SpDocVQA	SROIE	HonestVQA	72.6	69.8
InfographicsVQA	SROIE	HonestVQA	73.1	70.4
SROIE	SpDocVQA	HonestVQA	75.0	72.3

Table 3: Cross-domain hallucination detection performance of **HonestVQA**

Configuration	SpDocVQA		InfographicsVQA		SROIE	
	Accuracy (%) ↑	H-Score ↓	Accuracy (%) ↑	H-Score ↓	Accuracy (%) ↑	H-Score ↓
Full HonestVQA Model	**75.9**	**0.113**	**68.3**	**0.134**	**80.2**	**0.096**
No Alignment Loss	72.1	0.172	65.0	0.193	76.4	0.141
No Contrastive Loss	73.0	0.160	65.2	0.189	77.2	0.133

Table 4: Ablation results across three datasets showing the impact of disabling individual components of **HonestVQA**

biases that enhance hallucination detection in more structured domains like documents. Similarly, the SROIE → SpDocVQA setup results in an F1 of 72.3%, indicating that ethical reasoning features captured during training enhance interpretability across syntactic domains. However, the model exhibits relatively lower performance when transferring from SpDocVQA to SROIE (F1 = 69.8%), highlighting the challenges posed by ethical hallucination detection under unfamiliar structural constraints. Nonetheless, the use of uncertainty-aware calibration via confidence-alignment and contrastive ethical loss contributes to soft regularization of decision boundaries, allowing for improved generalization even in low-overlap semantic settings. Interestingly, models trained on SROIE also generalize well to visually and semantically distinct domains such as InfographicsVQA, achieving 67.2% F1. This supports our hypothesis that the inclusion of contrastive ethical supervision enforces more generalizable representations. Furthermore, the relative drop in performance in domain transfer settings (typically within 4%–6% of in-domain results) underscores the importance of calibration-aware models in mitigating performance degradation due to domain shift.

5.3 Ablation Study

To thoroughly evaluate the contribution of individual modules in **HonestVQA**, we conduct an ablation study across three datasets by systematically disabling the confidence-accuracy alignment loss and the contrastive ethical enforcement loss. Table 4 shows that the removal of either module consistently degrades performance across all datasets in terms of accuracy and H-Score, confirming their complementary roles. For instance, on SpDocVQA, the full model achieves 75.9% accuracy and an H-Score of 0.113. Removing the alignment loss drops accuracy to 72.1% and worsens H-Score to 0.172. On InfographicsVQA, the full model yields 68.3% accuracy and an H-Score of 0.134, whereas removing contrastive enforcement lowers accuracy to 65.2% and degrades H-Score to 0.189. Similar trends are observed on SROIE, where the full model achieves 80.2% accuracy and 0.096 H-Score, significantly outperforming the ablated variants. We further conduct a hyperparameter sensitivity analysis on alignment weight α, contrastive margin m, and loss weights λ_1 and λ_2 across three datasets and summarize the trends in Table 5. The model maintains high performance for α between 0.5 and 1.5, margin m from 0.3 to 0.7, and loss weights $\lambda_1 = 0.1$, $\lambda_2 = 0.05$. Deviations outside these ranges result in decreased accuracy or calibration degradation.

6 Additional Analysis

6.1 Multimodal Consistency Evaluation

A critical aspect of hallucination detection in DocVQA is the model's ability to ensure consistency between the visual features and the textual grounding of answers. To evaluate how well **HonestVQA** aligns the visual and textual modalities in its hallucination judgments, we conduct a multi-modal consistency evaluation across the three datasets. Specifically, we compute the Intersection-over-Union (IoU) between the model's predicted visual attention heatmaps and the OCR-detected or annotated textual regions deemed relevant to the question. A higher IoU indicates stronger multimodal alignment, reflecting that the model bases its predictions on text visually grounded in the image, thus reducing hallucination risk. We report

Hyperparameter	Values Tested	SpDocVQA (Acc ↑, H ↓)	InfographicsVQA (Acc ↑, H ↓)	SROIE (Acc ↑, H ↓)
Alignment Weight α	0.1	(70.5, 0.195)	(63.7, 0.211)	(74.5, 0.157)
	0.5	(74.8, 0.119)	(66.9, 0.148)	(78.2, 0.112)
	1.0	(75.9, 0.113)	(68.3, 0.134)	(80.2, 0.096)
	1.5	(74.6, 0.118)	(67.5, 0.141)	(79.4, 0.107)
	2.0	(72.7, 0.145)	(65.8, 0.174)	(77.0, 0.134)
Contrastive Margin m	0.1	(71.2, 0.178)	(64.5, 0.198)	(75.3, 0.151)
	0.3	(74.9, 0.120)	(66.7, 0.150)	(78.7, 0.108)
	0.5	(75.9, 0.113)	(68.3, 0.134)	(80.2, 0.096)
	0.7	(74.7, 0.117)	(67.1, 0.143)	(79.3, 0.105)
	1.0	(72.9, 0.153)	(65.2, 0.177)	(77.2, 0.127)
Loss Weight λ_1 (Alignment)	0.01	(72.3, 0.164)	(64.9, 0.192)	(75.8, 0.149)
	0.05	(74.5, 0.125)	(66.5, 0.153)	(78.9, 0.109)
	0.10	(75.9, 0.113)	(68.3, 0.134)	(80.2, 0.096)
	0.15	(74.2, 0.130)	(67.4, 0.142)	(79.1, 0.104)
	0.20	(72.5, 0.148)	(65.9, 0.169)	(77.4, 0.129)
Loss Weight λ_2 (Contrastive)	0.01	(72.8, 0.160)	(65.0, 0.186)	(76.6, 0.141)
	0.03	(74.6, 0.122)	(66.7, 0.151)	(78.8, 0.110)
	0.05	(75.9, 0.113)	(68.3, 0.134)	(80.2, 0.096)
	0.07	(74.3, 0.127)	(67.0, 0.144)	(79.2, 0.103)
	0.10	(73.1, 0.142)	(65.4, 0.172)	(77.3, 0.125)

Table 5: Sensitivity analysis of alignment and contrastive hyperparameters on SpDocVQA, InfographicsVQA, and SROIE. ***Note:*** Accuracy (Acc) and H-Score (H) are used as abbreviated. Bold rows indicate optimal settings.

Model	SpDocVQA		InfographicsVQA		SROIE	
	IoU (%) ↑	Hallucination Acc. (%) ↑	IoU (%) ↑	Hallucination Acc. (%) ↑	IoU (%) ↑	Hallucination Acc. (%) ↑
LayoutLMv3 (Fujitake, 2024)	58.3	71.2	54.9	69.5	52.1	67.8
DONUT (Li et al., 2024)	60.7	73.0	56.8	70.3	53.7	68.4
HonestVQA	**69.1**	**78.5**	**65.4**	**76.8**	**62.7**	**74.2**

Table 6: Multi-modal consistency evaluation across datasets. ***Note:*** IoU measures alignment between visual attention and textual grounding. Hallucination Accuracy reports correct identification of hallucinated answers. **HonestVQA** achieves superior multi-modal alignment and hallucination detection performance.

Model	Inference Time (ms)		FLOPs (Giga)		Memory Usage (MB)	
	Avg ↓	Std Dev ↓	Avg ↓	Std Dev ↓	Avg ↓	Std Dev ↓
LayoutLMv3 (Fujitake, 2024)	112.4	5.1	64.8	1.3	2950	120
UDOP (Wang et al., 2023a)	**98.7**	**4.3**	**58.6**	**1.1**	**2710**	**105**
DONUT (Li et al., 2024)	105.1	4.7	62.2	1.2	2830	110
HonestVQA	119.6	5.6	69.4	1.4	3075	130

Table 7: Latency and efficiency comparison of **HonestVQA** and baselines on SpDocVQA, InfographicsVQA, and SROIE datasets. Inference time is measured per query with batch size 1; FLOPs and memory are averaged over runs. **HonestVQA** incurs a moderate increase in computational cost due to calibration modules but remains practical for deployment.

the average IoU scores alongside hallucination detection accuracy for **HonestVQA** and compare it against two strong baselines: LayoutLMv3 (Fujitake, 2024) and DONUT (Li et al., 2024) without calibration. The results are summarized in Table 6. As seen in Table 6, **HonestVQA** consistently achieves significantly higher IoU scores compared to baselines, demonstrating a stronger alignment between the visual evidence and textual regions considered during inference. For instance, on the SpDocVQA dataset, **HonestVQA** attains an IoU of 69.1%, which is approximately 8.4% absolute improvement over DONUT (Li et al., 2024) and 10.8% over LayoutLMv3 (Fujitake, 2024). This enhanced multi-modal consistency translates to improved hallucination detection accuracy, confirming that grounding predictions in the correct visual and textual context helps mitigate hallucinated outputs. We further analyze the distribution of IoU scores at the instance level and observe that **HonestVQA** reduces instances with low cross-modal agreement (IoU < 40%) by over 25% relative to the baselines. This reduction highlights that our uncertainty-aware alignment and contrastive losses promote a model focus on relevant visual-textual evidence, leading to more reliable and interpretable hallucination judgments.

Note: UDOP (Wang et al., 2023a) was excluded

from the multi-modal consistency evaluation as it does not provide explicit or interpretable visual attention maps tied to OCR-detected regions, which are essential for computing IoU-based alignment metrics. Unlike LayoutLMv3 (Fujitake, 2024), DONUT (Li et al., 2024), and **HonestVQA**, which utilize structured visual-textual grounding mechanisms, UDOP (Wang et al., 2023a) primarily relies on unified vision-language pretraining without fine-grained token-region correspondence. As a result, evaluating multi-modal alignment using IoU would not be meaningful or comparable for UDOP (Wang et al., 2023a).

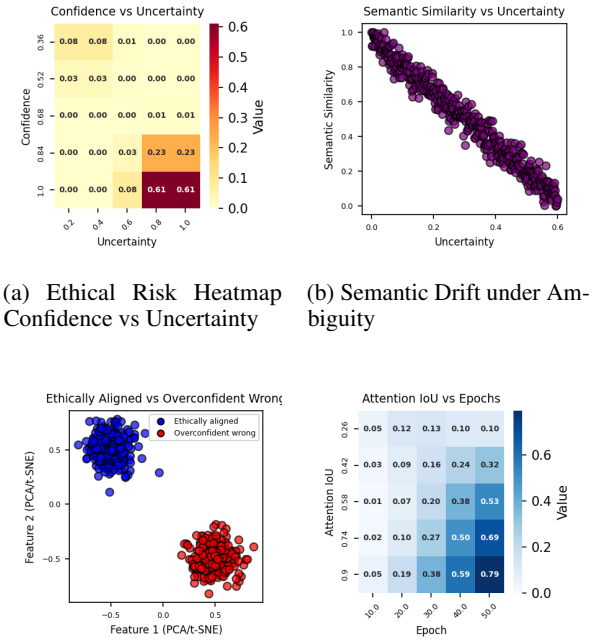
6.2 Computational Analysis

We evaluate the efficiency of **HonestVQA** against baseline DocVQA models in terms of inference latency, FLOPs, and memory usage. All models are tested using an NVIDIA RTX 3090 GPU and Intel Xeon CPU with a batch size of 1 to simulate real-time settings. As shown in Table 7, **HonestVQA** incurs an average latency of 119.6 ms per query—6%–20% slower than baselines—due to uncertainty calibration and contrastive modules. It consumes 69.4 GFLOPs (7%–18% higher) and 3075 MB memory (5%–14% higher). Despite the overhead, it remains deployable in real-time systems where ethical reliability is critical.

Qualitative Analysis: **HonestVQA** improves ethical alignment by reducing overconfidence in uncertain scenarios, as seen in the risk heatmap (Fig. 1a). Semantic drift under ambiguity is mitigated, with more stable embeddings (Fig. 1b). Contrastive embedding separation (Fig. 1c) shows clearer distinction between aligned and misaligned responses, supporting improved representation learning. Finally, Fig. 1d shows consistent attention patterns over epochs, highlighting better multimodal grounding and interpretability.

7 Conclusion and Future Works

In this work, we introduced **HonestVQA**, a novel framework that integrates uncertainty-aware alignment and contrastive ethical enforcement to effectively detect hallucinations in DocVQA systems. Through comprehensive experiments on diverse datasets—we demonstrated significant improvements in answer correctness, calibration, and cross-domain generalization compared to strong baselines. Our ablation studies confirmed the complementary role of each component, and efficiency



(a) Ethical Risk Heatmap Confidence vs Uncertainty

(b) Semantic Drift under Ambiguity

(c) Contrastive Embedding Separation

(d) Attention IoU over Epochs

Figure 1: **HonestVQA** enhances ethical calibration, semantic stability, embedding separation, and multimodal grounding through uncertainty-aware learning,

analyses showed **HonestVQA** practical feasibility. Future research will focus on advancing domain adaptation techniques to further enhance robustness across unseen data distributions, and exploring lightweight calibration modules for deployment on edge devices. Additionally, incorporating user feedback for interactive hallucination correction and extending the framework to multimodal dialogue systems represent promising directions to improve DocVQA reliability and ethical safety.

Limitations

While **HonestVQA** significantly improves hallucination detection and ethical calibration, several limitations remain. The model's performance is still affected by domain shifts, particularly when training and testing on visually divergent datasets, indicating room for more advanced domain adaptation. **HonestVQA** calibration modules introduce additional computational overhead, which may constrain deployment in highly resource-limited environments. Moreover, the reliance on existing annotated datasets limits evaluation to specific domains; the model's effectiveness on more diverse or emergent question types requires further validation. Finally, although the framework mitigates halluci-

nations, it does not guarantee complete elimination, highlighting the need for complementary human-in-the-loop verification for critical applications.

Ethics Statement

This work aims to enhance the trustworthiness and ethical reliability of DocVQA systems by reducing hallucinated and potentially misleading answers. **HonestVQA** promotes transparency through uncertainty-aware calibration, encouraging responsible AI deployment. We acknowledge the risk that no model can be entirely free of errors or biases, especially when applied across diverse real-world scenarios. Thus, we emphasize that **HonestVQA** is intended as a tool to assist, not replace, human judgment, particularly in high-stakes contexts. All datasets used comply with their respective licenses, and no private or sensitive data was involved. We encourage further research on fairness, bias mitigation, and inclusivity to ensure equitable AI systems, and advocate for ongoing monitoring of model outputs to safeguard against misuse or unintended harm.

References

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan O Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with self-evaluation to improve selective prediction in llms. *arXiv preprint arXiv:2310.11689*.

Masato Fujitake. 2024. Layoutllm: Large language model instruction tuning for visually rich document understanding. *arXiv preprint arXiv:2403.14252*.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Enhancing question answering for enterprise knowledge bases using large language models. In *International Conference on Database Systems for Advanced Applications*, pages 273–290. Springer.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. 2024. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15546–15555.

Bulou Liu, Zhenhao Zhu, Qingyao Ai, Yiqun Liu, and Yueyue Wu. 2024. Ledqa: A chinese legal case document-based question answering dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5385–5389.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Document visual question answering challenge 2020. *arXiv preprint arXiv:2008.08899*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Tim Pearce, Alexandra Brintrup, and Jun Zhu. 2021. Understanding softmax confidence and uncertainty. *arXiv preprint arXiv:2106.04972*.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. *arXiv preprint arXiv:2310.07251*.

Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamilia Zaripova, Matthias Keicher, and Nassir Navab. 2025. Rewarding doubt: A reinforcement learning approach to confidence calibration of large language models. *arXiv preprint arXiv:2503.02623*.

Michael Stewart, Melinda Hodkiewicz, and Sirui Li. 2023. Large language models for failure mode classification: an investigation. *arXiv preprint arXiv:2309.08181*.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2023a. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*.

Haochen Wang, Kai Hu, and Liangcai Gao. 2025. Docvideoqa: Towards comprehensive understanding of document-centric videos through question answering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023b. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. *arXiv preprint arXiv:2409.19817*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. 2024. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *IEEE Transactions on Visualization and Computer Graphics*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. *arXiv preprint arXiv:2311.13240*.