Supporting Information for

Orchestrator-Agent Trust: An Agentic AI Visual Classification System with Confidence-Aware Orchestration and RAG Reasoning

Konstantinos I. Roumeliotis, Ranjan Sapkota, Manoj Karkee and Nikolaos D. Tselikas

Corresponding Authors.
E-mail: k.roumeliotis@uop.gr, rs2672@cornell.edu

This PDF file includes:

Figs. S1 to S13
Tables S1 to S5

```
"agent_prompt": """Analyze the provided image of an apple leaf using your
  computer vision capabilities. Classify the leaf into the most appropriate
  category based on its condition, choosing from the predefined list:
  {
  "categories": [
    "black-rot",
    "healthy",
    "rust",
    "scab"
    ]
  }
  Provide your final classification in the following JSON format without
  explanations:
  {
    "category": "chosen_category_name",
    "reason": "brief_reason_for_this_classification",
    "confidence": "confidence_score_between_0_and_1"
  }""",

"agent_response": {'category': 'scab',
                   'reason': 'The presence of brown spots and scarring on the'
                   'leaf surfaceis characteristic of scab disease.',
                   'confidence': 0.95, 'time': 4.0999331474, 'cost': 0.001035}
```

Fig. S1. The AI agent prompt for classifying apple leaf diseases using computer vision, alongside the agent's JSON-formatted response including classification, reasoning, confidence score, processing time, and cost.

```
"orchestrator_prompt": """You are acting as the orchestrator in an agentic AI
  system. Your role is to make an informed and well-reasoned final decision
  for a classification task, based on the outputs of multiple AI agents.
  Task:
  Classify an image of an apple plant leaf into the correct disease category.
  Context:
  You have received independent responses from two specialized AI agents.
  Each agent may use different modalities, models, or reasoning approaches.
  Your responsibility is to analyze, compare, and synthesize their outputs to
  make a final, justified classification decision.

  $agents_response

  Your Task as Orchestrator:
  Based on the agents' responses, provide your final classification decision.

  You must return your answer in the following JSON format:
  {
    "category": "chosen_category_name",
    "reason": "detailed explanation of your reasoning process",
    "confidence": "confidence_score_between_0_and_1"
  }

  Be sure your reasoning includes:
  - How you evaluated the agents' outputs
  - Any conflicts or agreements you observed
  - The rationale behind your chosen decision"""

"orchestrator_response": {'category': 'scab', 'reason': 'Both AI Agent 1 and'
  'AI Agent 2 identified the disease as scab based on the presence of dark,'
  'scabby spots and characteristic brown spots with scarring on the leaf'
  'surface. Given the strong agreement between both agents, and their high'
  'confidence levels (0.95 and 0.85), the evidence clearly points to scab'
  'disease. This consensus, along with the corroborating descriptions, supports'
  'a reliable classification.',
  'confidence': 0.93, 'time': 3.1174688339, 'cost': 0.000787}
```

Fig. S2. The orchestrator prompt and response in an agentic AI system. The orchestrator synthesizes independent classifications from multiple AI agents to produce a final, justified decision for apple leaf disease diagnosis. The response includes the chosen category, detailed reasoning explaining agreement between agents, confidence score, inference time, and cost.

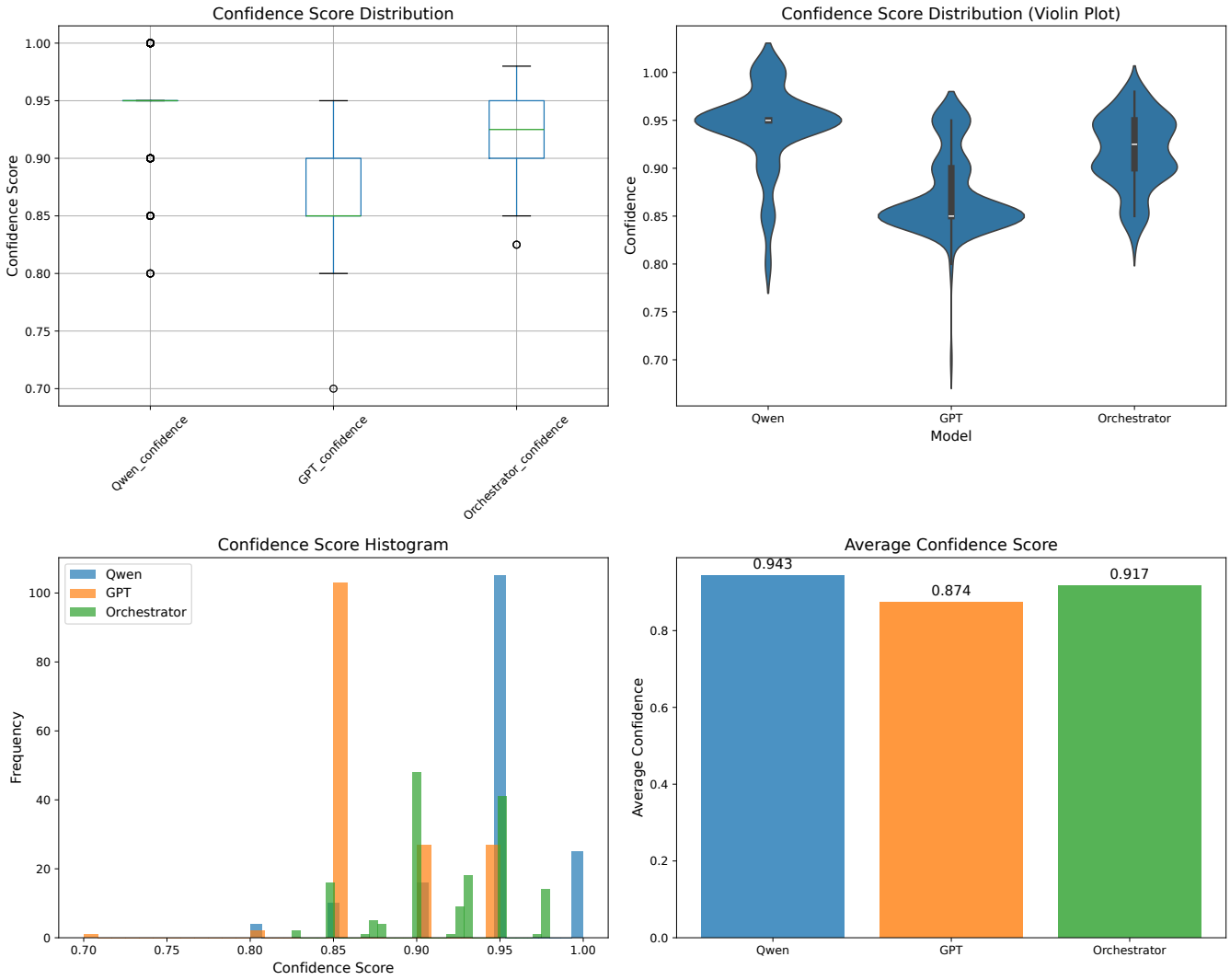

Fig. S3. Experiment I. Confidence Analysis

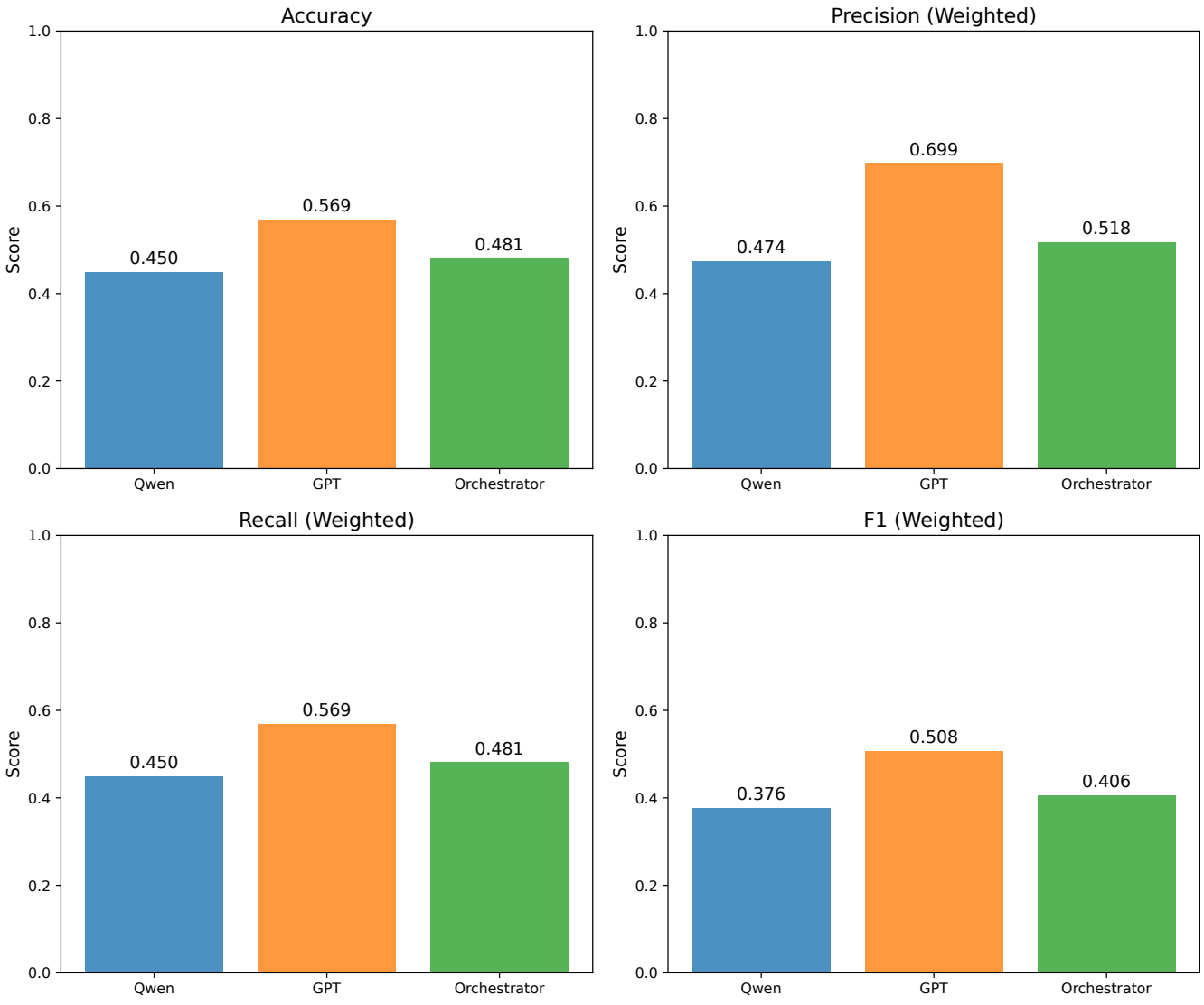

Fig. S4. Experiment I. Classification Metrics Comparison

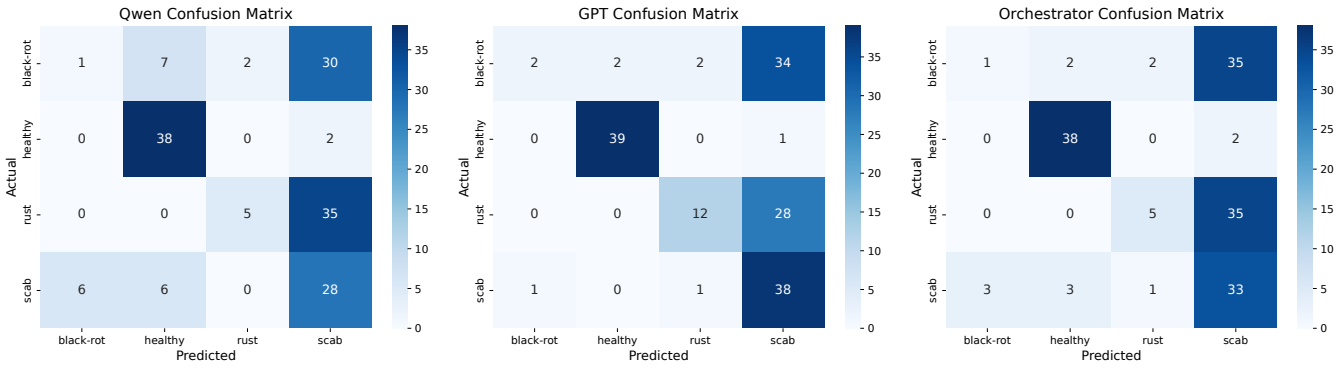

Fig. S5. Experiment I. Confusion Matrices

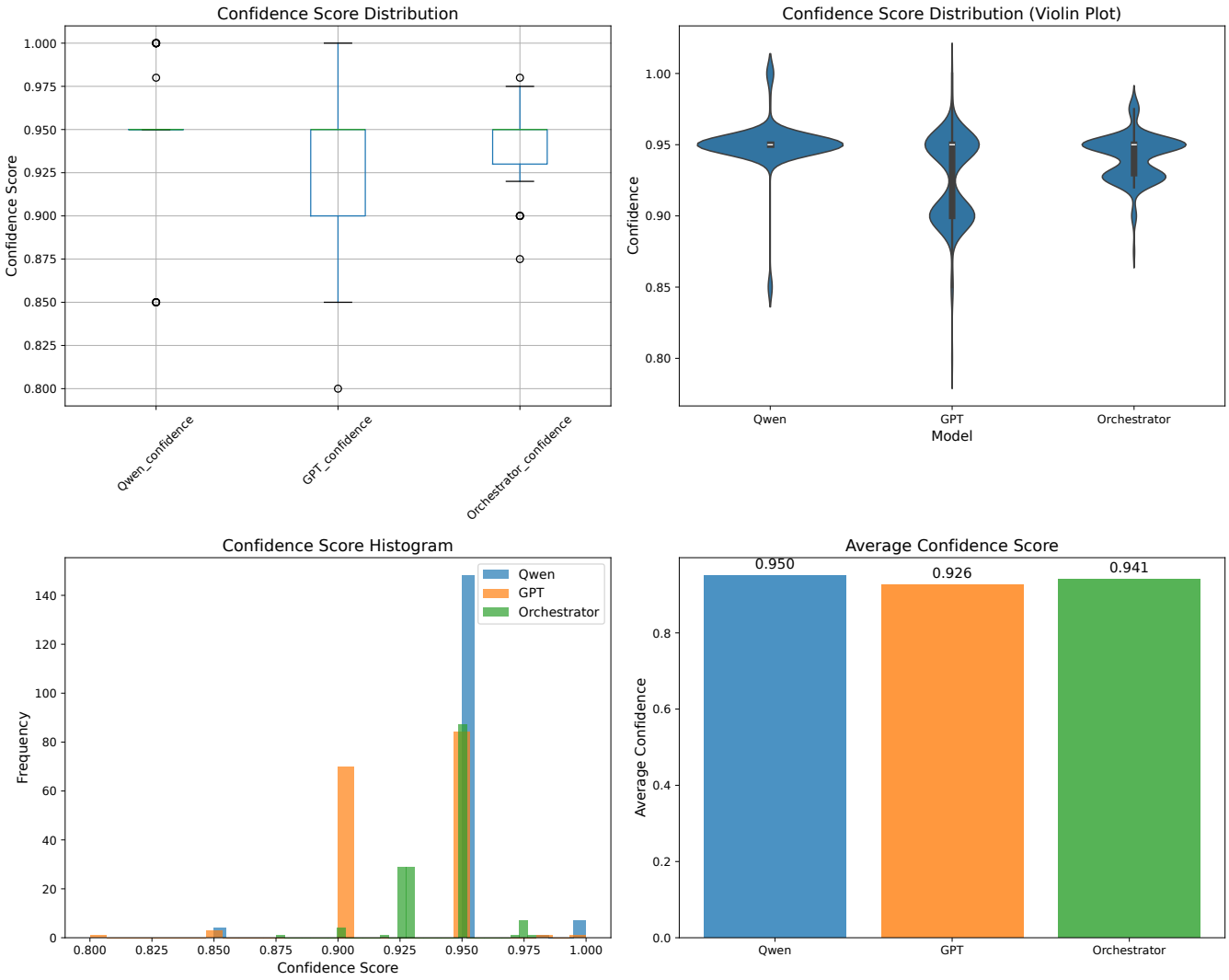

Fig. S6. Experiment II. Confidence Analysis

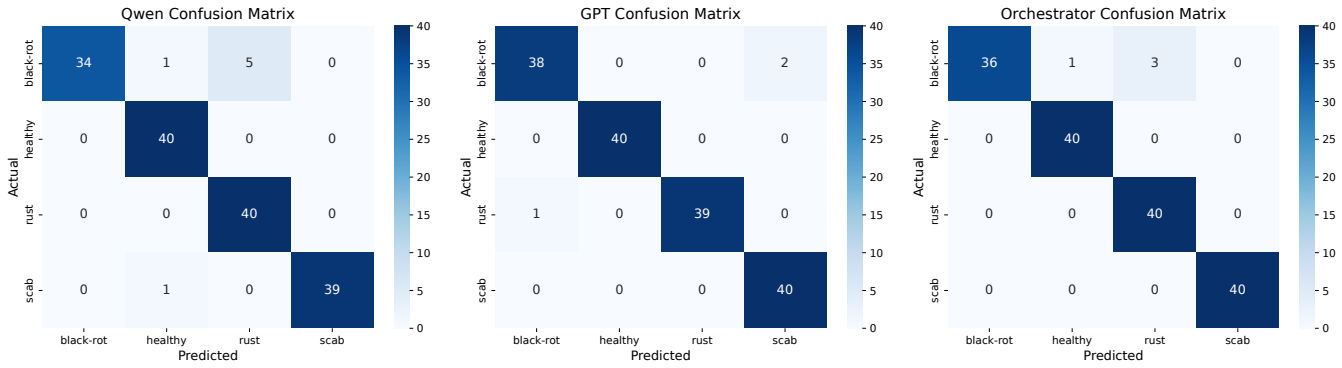

Fig. S7. Experiment II. Confusion Matrices

```
Using RAG...
{'response': [{'category': 'rust', 'confidence': 1.0}], 'time': 0.1972808838,
  'cost': 5e-05}
Using Qwen Unsloth Base Model | Zero-Shot...
{'category': 'scab', 'reason': 'Presence of brown spots that resemble scab'
  'fungal infection.', 'confidence': 0.85, 'time': 3.7545511723,
  'cost': 0.000948}
Using GPT OpenAI Base Model | Zero-Shot...
{'category': 'rust', 'reason': 'Presence of small, rust-colored spots on the'
  'leaf surface.', 'confidence': 0.85, 'time': 2.1049318314, 'cost': 0.001357}
Using GPT Reasoning Model as an Orchestrator...
{'agent_1': 'Please reconsider your selection, taking into account the'
  'Image-RAG. You are not required to change your selection if you are'
  'confident in it. However, if you do change your answer, please explain in'
  'detail why you changed your mind, providing reasoning based on the image.',
'agent_2': 'Please reconsider your selection, taking into account the'
  'Image-RAG. You are not required to change your selection if you are'
  'confident in it. However, if you do change your answer, please explain in'
  'detail why you changed your mind, providing reasoning based on the image.',}
Using Qwen Unsloth Base Model | Zero-Shot...
{'category': 'rust', 'reason': 'The leaf exhibits characteristic spots'
  'indicative of a fungal infection, which aligns closely with the appearance'
  'of rust symptoms.', 'confidence': 0.95, 'time': 4.6233711243,
  'cost': 0.001167}
Using GPT OpenAI Base Model | Zero-Shot...
{'category': 'rust', 'reason': 'The leaf displays small, rust-colored spots'
'consistent with rust disease.', 'confidence': 1.0, 'time': 2.9448902607,
 'cost': 0.001357}
Using GPT Reasoning Model as an Orchestrator...
{'category': 'rust', 'reason': 'Both AI Agent 1 and AI Agent 2 independently'
  'classified the disease as rust based on clear visual indicators such as'
  'characteristic rust-colored spots on the leaf. There is complete agreement'
  'between the agents, with each noting the visible signs of fungal infection'
  'typical of rust disease. The high confidence scores (0.95 and 1.0) further'
  'reinforce the decision to classify the disease as rust.',
  'confidence': 0.975, 'time': 2.5999305248, 'cost': 0.001569}
```

Fig. S8. Experiment III. Agentic AI reasoning orchestration combining Image-RAG with iterative re-evaluation from multiple AI agents. The process enhances trust and accuracy in plant disease classification by reconciling outputs and confidence levels from diverse models.

```
"orchestrator_case_3_prompt": """You are the orchestrator in an agentic AI system.
The system's task is to classify an image of an apple plant leaf into the correct disease category.
The AI agents in this system have returned their classifications along with *self-confidence* scores for this task. However, we are
     unsure whether their classifications and confidence scores are reliable.
To address this, we conducted an independent evaluation of each agent's calibration metrics, indicating how trustworthy their
     confidence scores are.

Think carefully and decide whether you trust the agents' outputs, or whether they need to re-evaluate their responses using our
     Image-RAG.

- **Agents' responses**:
  $agents_response
- **Agents' independent evaluation**:
  $agents_confidence

If you trust them, return:
{
  "category": "chosen_category_name",
  "reason": "detailed explanation of your reasoning process",
  "confidence": "confidence_score_between_0_and_1"
}

If you do not trust one or both of them, return:
{
  "agent_1": "Please reconsider your selection, taking into account the Image-RAG. You are not required to change your selection if
       you are confident in it. However, if you do change your answer, please explain in detail why you changed your mind,
       providing reasoning based on the image.",
  "agent_2": "Please reconsider your selection, taking into account the Image-RAG. You are not required to change your selection if
       you are confident in it. However, if you do change your answer, please explain in detail why you changed your mind,
       providing reasoning based on the image."
}
"""
```

Fig. S9. Experiment III. Prompt used by the Orchestrator agent in Experiment III. The orchestrator assesses the trustworthiness of AI agents' classifications based on their responses and independently evaluated trust metrics. Depending on this evaluation, it either accepts a classification or initiates a re-evaluation using Image-RAG.

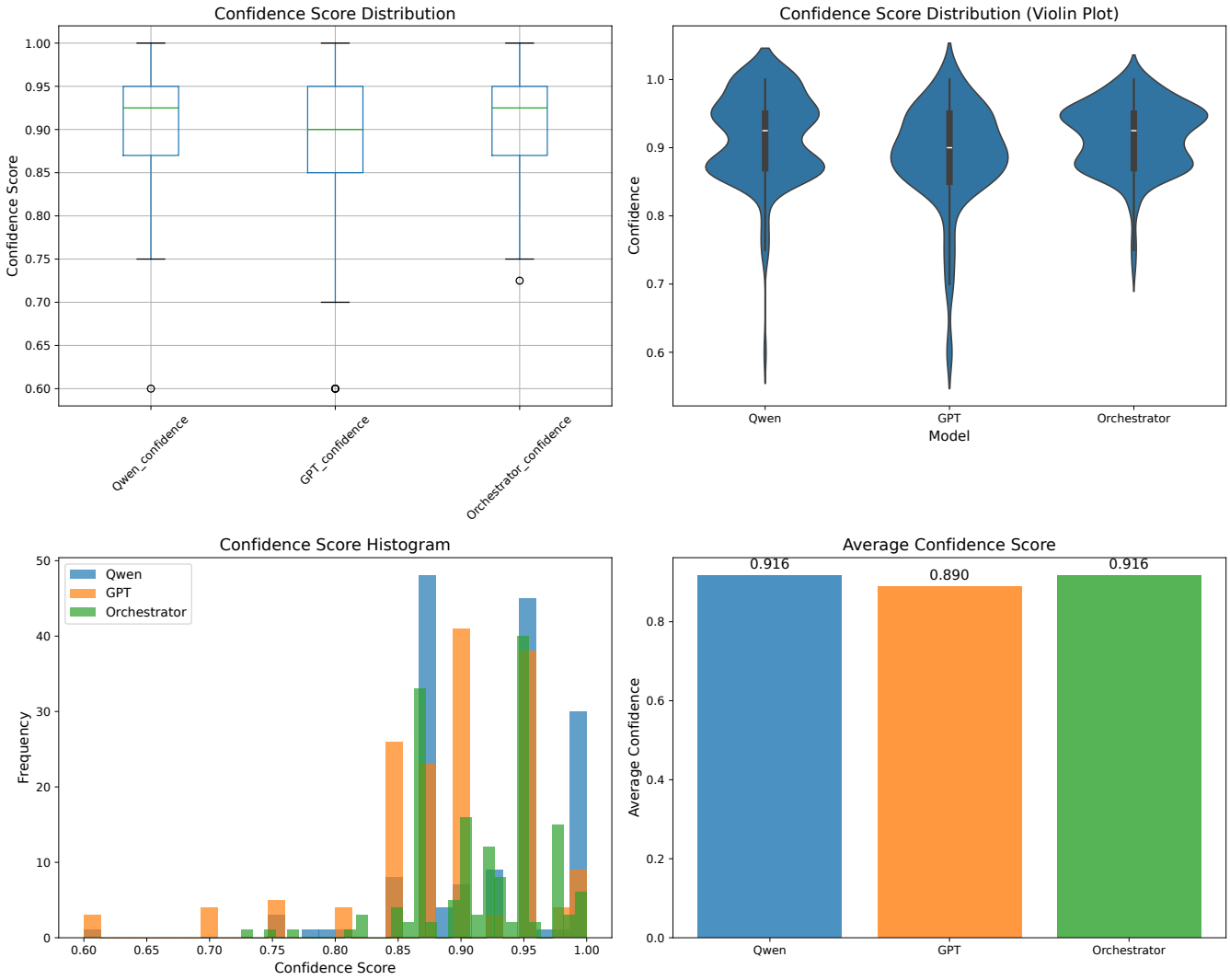

Fig. S10. Experiment III. Confidence Analysis

Table S1. Experiment I. Summary statistics of confidence scores for each agent during zero-shot classification.

Statistic	Qwen Confidence	GPT-4o Confidence	Orchestrator Confidence
Count	160	160	160
Mean	0.943	0.874	0.917
Std. Dev.	0.042	0.042	0.037
Min	0.800	0.700	0.825
25th Pctl	0.950	0.850	0.900
Median	0.950	0.850	0.925
75th Pctl	0.950	0.900	0.950
Max	1.000	0.950	0.980

Table S2. Experiment II. Comparison of fine-tuning configurations for GPT-4o using ResNet-50–derived and default hyperparameters.

Base Model	Epochs	Batch Size	Val. Loss	Duration (s)	Cost (USD)
GPT-4o (ResNet-50 tuned)	10	16	0.0088	1,778	47.53
GPT-4o (Default settings)	3	1	0.0617	1,652	13.09

Table S3. Experiment II. Summary statistics of confidence scores for each agent during few-shot classification.

Statistic	Qwen Confidence	GPT Confidence	Orchestrator Confidence
Count	160	160	160
Mean	0.9499	0.9258	0.9413
Standard Deviation	0.0192	0.0294	0.0158
Minimum	0.85	0.80	0.875
25th Percentile	0.95	0.90	0.93
Median (50th Percentile)	0.95	0.95	0.95
75th Percentile	0.95	0.95	0.95
Maximum	1.00	1.00	0.98

```
"agent_re_evaluation_prompt": """Context from the previous conversation:
  $previous_context
  Your previous answer:
  $previous_response

  Please reconsider your selection, taking into account the Image-RAG:
  $agentic_rag

  You are not required to change your selection if you are confident in it. However, if you do change your answer, please explain in
        detail why you changed your mind, providing reasoning based on the image.

  CLASSIFICATION REQUIREMENTS:
  - You must select exactly ONE category: black-rot, healthy, rust, or scab
  - Confidence must be a decimal number between 0.0 and 1.0
  - Provide a clear, detailed reason for your classification
  Your response should end with this exact JSON format (complete and valid):
  {
    "category": "black-rot",
    "reason": "The leaf shows characteristic dark spots with concentric rings typical of black rot disease",
    "confidence": 0.87
  }

  Replace the values above with your actual classification. Ensure the JSON is complete with all closing braces and quotes."""
```

Fig. S11. Experiment III. Prompt used during the agent re-evaluation loop in Experiment III. After reviewing their previous answer and receiving additional image-grounded context (Image-RAG), agents are asked to reassess their classification, provide reasoning, and output a structured response in JSON format.

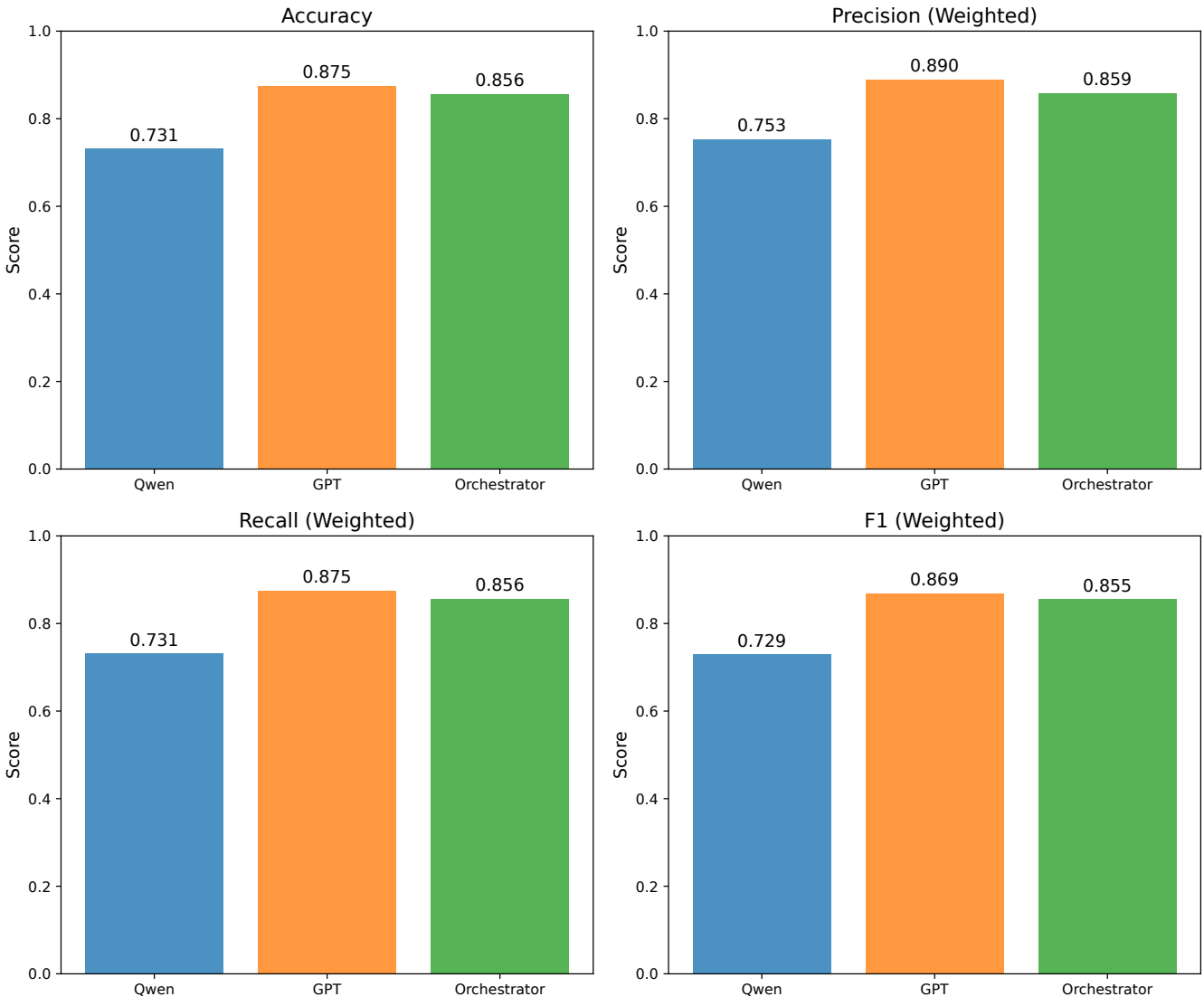

Fig. S12. Experiment III. Classification Metrics Comparison

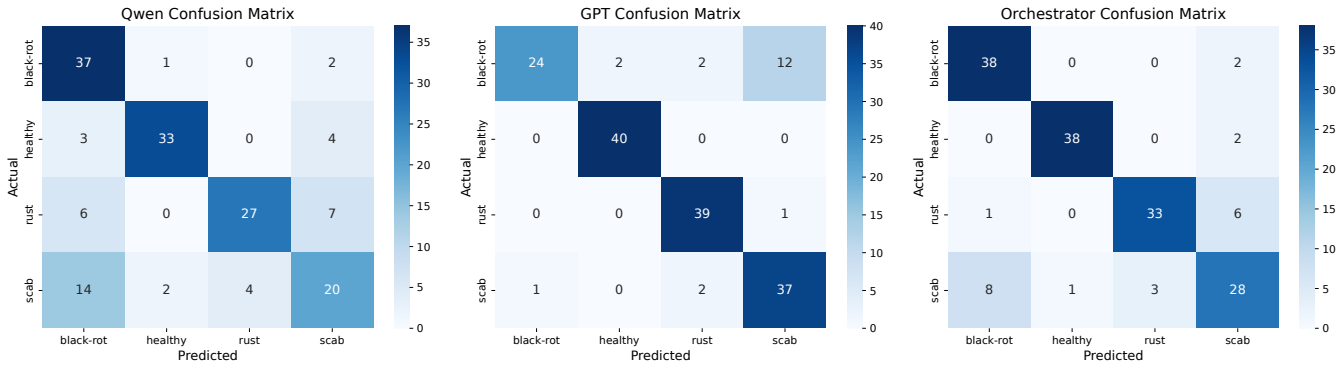

Fig. S13. Experiment III. Confusion Matrices

Table S4. Experiment III. Trust profiling results of Qwen and GPT agents in a zero-shot setting using multiple calibration and reliability metrics.

Model	Acc.	Avg. Conf.	$\text{Conf}_{\text{corr}}$	$\text{Conf}_{\text{incorr}}$	CG	OCR	HCW	THC	CCC	p-val	ECE	CWA
Qwen	0.492	0.945	0.950	0.941	0.009	0.508	260	512	0.126	0.0042	0.453	0.495
GPT	0.584	0.877	0.890	0.860	0.030	0.416	213	512	0.361	0.0000	0.293	0.592

Acc., overall accuracy; Avg. Conf., average predicted confidence; $\text{Conf}_{\text{corr}}$, confidence when correct; $\text{Conf}_{\text{incorr}}$, confidence when incorrect; CG, confidence gap (mean difference between confidence and correctness); OCR, overconfidence rate (fraction of incorrect predictions with confidence > 0.9); HCW, high-confidence wrong predictions; THC, total high-confidence predictions; CCC, Pearson correlation between confidence and correctness; ECE, expected calibration error; CWA, confidence-weighted accuracy.

Table S5. Experiment III. Summary statistics of the confidence scores for each agent after the re-evaluation loop.

Statistic	Qwen Confidence	GPT Confidence	Orchestrator Confidence
Count	160	160	160
Mean	0.9164	0.8896	0.9164
Standard Deviation	0.0635	0.0737	0.0497
Minimum	0.60	0.60	0.725
25th Percentile	0.87	0.85	0.87
Median (50th Percentile)	0.925	0.90	0.925
75th Percentile	0.95	0.95	0.95
Maximum	1.00	1.00	1.00

References