Case Study Report — Context Engineering with Multimodal Image Understanding

Alara Zindancıoğlu

The system was evaluated using the provided `eval_samples.jsonl` script. The full results are available in `evaluation_results/latest_run.json`.

Accuracy: 100% (6/6 passed). The system correctly handled all evaluation samples.
Compliance: 100%. Correctly identified and failed the compliance check for the "orthopedic" claim on `SKU-010`, citing rules `R-103` and `R-202`.
Latency: avg ~16.3 seconds (range was 12.7s to 23.3s)

Conflicts & Edge Cases

- The system was effective in **handling conflict and identifying discrepancies** between the visual evidence and the data provided in the spec catalog. For example, for SKU-001, the model responded as: “Based on the image, the shoe features a lace-up closure. However, the product attributes in the context card state the closure type as 'velcro'. There is a discrepancy…”

- **Handling Ambiguity**: When an image is ambiguous, the model responds with uncertainty. For example, for SKU-005, it responded as: "Based on the product attributes, the closure type is velcro. However, the image provided for SKU-005 shows a boot that appears to be a pull-on or have a hidden zipper, and a second shoe with laces. Neither shoe in the image displays a velcro closure, indicating a discrepancy between the product attributes and the visual evidence.”. The model also cited ‘insufficient visual evidence for R-102 in this task.

Sample run:

```
--- Starting Evaluation ---

[1/1] Evaluating SKU: SKU-001
  Question: What is the closure type?
  -> Result: PASS

[2/2] Evaluating SKU: SKU-002
  Question: What is the closure type?
  -> Result: PASS
```

[3/3] Evaluating SKU: SKU-003
  Question: What is the closure type?
  -> Result: PASS

[4/4] Evaluating SKU: SKU-004
  Question: What is the closure type?
  -> Result: PASS

[5/5] Evaluating SKU: SKU-005
  Question: What is the closure type?
  -> Result: PASS

[6/6] Evaluating SKU: SKU-010
  Question: Can we claim it is orthopedic?
  -> Result: PASS

Detailed evaluation results saved to: evaluation_results/latest_run.json

--- Evaluation Summary ---
Total Samples: 6
Passed: 6
Failed/Errored: 0
Skipped: 0
Accuracy (on evaluated samples): 100.00%

Brief Summary:

	sku	pass	reason
0	SKU-001	True	API returned a valid answer.
1	SKU-002	True	API returned a valid answer.
2	SKU-003	True	API returned a valid answer.
3	SKU-004	True	API returned a valid answer.
4	SKU-005	True	API returned a valid answer.
5	SKU-010	True	Correctly failed compliance rule R-103.