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.