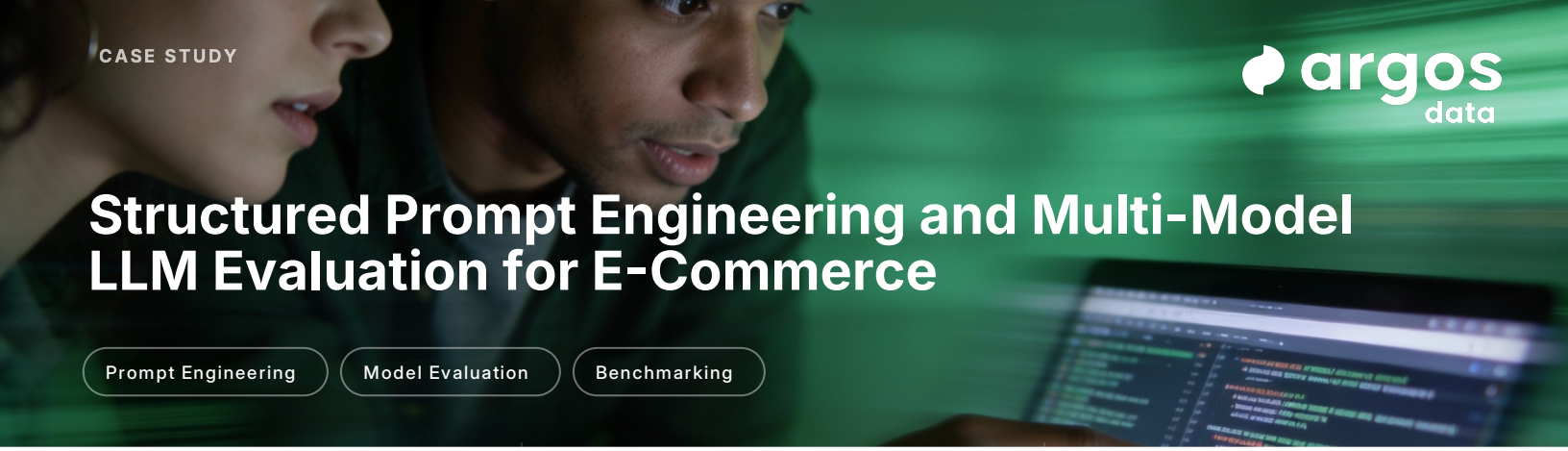

CASE STUDY

Structured Prompt Engineering and Multi-Model LLM Evaluation for E-Commerce

Prompt Engineering | Model Evaluation | Benchmarking

THE CLIENT

A global e-commerce and technology organization developing AI systems that interpret and act on customer interactions, including intent classification, agent action prediction, content compliance, product-query matching, and structured response generation.

THE CHALLENGE

The client needed a structured dataset of high-complexity prompts designed to stress-test language models across 5 e-commerce task categories, each with an input prompt, context, reasoning trace, ground-truth classification, and rich metadata. Without a unified environment, testing meant error-prone manual copy-paste cycles across separate model interfaces, undermining consistency and cross-model comparison.

THE ARGOS DATA SOLUTION

Argos Data built a centralized prompt engineering and multi-model evaluation environment designed to support controlled creation, submission, evaluation, and review of e-commerce prompts. The environment was deployed inside Argos Myriad and configured to the client's specific dataset specifications.

The program embedded structured review controls, taxonomy distribution validation, metadata consistency verification, difficulty calibration checks based on cross-model performance, and controlled reasoning trace evaluation. This ensured dataset integrity and scientific comparability across evaluated models.

CAPABILITIES DELIVERED

Structured Prompt Authoring Interface

Linguists generated prompts within a guided environment aligned to dataset requirements.

Multi-Model Access Integration

Prompts could be submitted directly to 4 designated LLMs within 1 system, eliminating manual cross-platform handling.

Automated Response Capture

Model outputs were stored automatically for structured comparison.

Ground Truth Alignment Workflow

Reviewers evaluated model responses against expected classifications and reasoning traces.

Complexity Calibration Support

The system tracked how models performed on each prompt, supporting Level 1, Level 2, and Level 3 difficulty categorization.

Specialized Review Layer

Expert reviewers validated prompt construction quality, structural integrity, reasoning trace adequacy, and taxonomy compliance.

RESULTS

- Structured generation of high-complexity e-commerce prompts across five task categories
- Controlled multi-model evaluation in a single unified environment
- Accurate difficulty level calibration grounded in cross-model performance data
- Reduced operational friction by eliminating manual cross-platform submission
- Reproducible model failure measurement methodology
- High-quality dataset delivery aligned to specification requirements