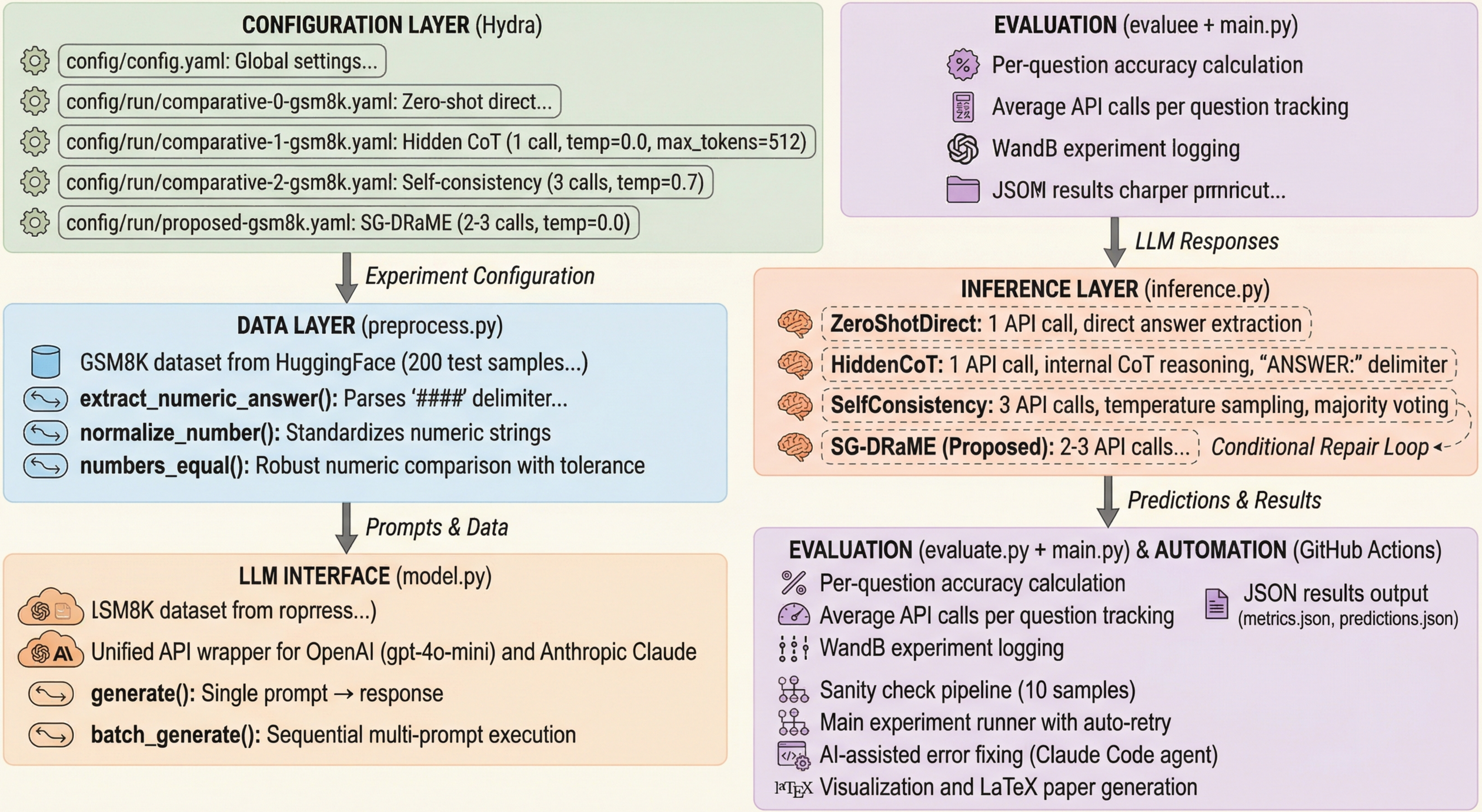
CONFIGURATION LAYER (Hydra)
config/config.yaml: Global settings...
config/run/comparative-0-gsm8k.yaml: Zero-shot direct...
config/run/comparative-1-gsm8k.yaml: Hidden CoT (1 call, temp=0.0, max_tokens=512)
config/run/comparative-2-gsm8k.yaml: Self-consistency (3 calls, temp=0.7)
config/run/proposed-gsm8k.yaml: SG-DRaME (2-3 calls, temp=0.0)
Experiment Configuration
DATA LAYER (preprocess.py)
GSM8K dataset from HuggingFace (200 test samples...)
extract_numeric_answer(): Parses '####' delimiter...
normalize_number(): Standardizes numeric strings
numbers_equal(): Robust numeric comparison with tolerance
Prompts & Data
LLM INTERFACE (model.py)
LSM8K dataset from roprress...)
Unified API wrapper for OpenAI (gpt-4o-mini) and Anthropic Claude
generate(): Single prompt → response
batch_generate(): Sequential multi-prompt execution
EVALUATION (evaluee + main.py)
Per-question accuracy calculation
Average API calls per question tracking
WandB experiment logging
JSOM results charper prrnricut...
LLM Responses
INFERENCE LAYER (inference.py)
ZeroShotDirect: 1 API call, direct answer extraction
HiddenCoT: 1 API call, internal CoT reasoning, "ANSWER:" delimiter
SelfConsistency: 3 API calls, temperature sampling, majority voting
SG-DRaME (Proposed): 2-3 API calls...
Conditional Repair Loop
Predictions & Results
EVALUATION (evaluate.py + main.py) & AUTOMATION (GitHub Actions)
Per-question accuracy calculation
Average API calls per question tracking
WandB experiment logging
JSON results output
(metrics.json, predictions.json)
Sanity check pipeline (10 samples)
Main experiment runner with auto-retry
AI-assisted error fixing (Claude Code agent)
Visualization and LaTeX paper generation