2. The Routing Policies Document

Version: 1.0
Status: Draft

2.1 Domain Policy

- Rule: A request is classified into a domain if it contains ≥2 keywords associated with that domain.
- Precedence: coding_architecture > coding_implementation > reasoning > creative > documentation.
- Fallback: If no domain matches, classify as unknown and utilize embedding similarity.

2.2 Stakes Policy

- Calculation: Stakes Score = Sum(Complexity Indicators) + Sum(Risk Indicators).
- Thresholds:
    - Score ≥ 5 → High Stakes
    - Score 2-4 → Medium Stakes
    - Score < 2 → Low Stakes
- Override: coding_architecture is always at least High Stakes (Score ≥ 3).

2.3 Modality & Tool Policy

- OCR: Required if request includes PDF/Image attachments AND keywords (scan, document).
- Vision: Required if request includes Image attachments AND keywords (image, visual).
- Embeddings: Required if request implies retrieval (find similar, context) OR if Domain is unknown.

2.4 Model Policy

- Selection: Select the model with the highest proficiency score for the determined Domain.
- Assignments:
    - Qwen Coder 32B: Primary for coding_architecture (0.95) and coding_implementation (0.85).
    - Nemotron 30B: Primary for coding_implementation (0.95) if performance/optimization is requested.
    - GPT-OSS 20B: Primary for reasoning (0.90).
    - MythoMax 13B: Primary for creative (0.95).

2.5 Validator Policy

- High Stakes: Enforce block_by_block validation (validate every 5-10 lines/logical block).

- Medium Stakes: Enforce end_stage validation (validate entire output once complete).
- Low Stakes: No validation (none) required, unless explicitly requested.

2.6 Depth Policy (Validation Granularity)

- Per-Line: Applied to coding_architecture in High Stakes.
- Per-Function: Applied to coding_implementation in High Stakes.
- Per-Output: Applied to all Medium Stakes tasks.
- No Validation: Applied to Low Stakes tasks.

2.7 Confidence Policy

- Acceptable (≥ 0.75): Proceed with routing decision.
- Uncertain (0.60 - 0.74): Trigger embedding fallback to find similar past successful routes.
- Critical (< 0.60): Halt and request user clarification.