Mitigating Knowledge Entropy: A Multi-Agent Framework with Decoupled Reranking and Governance-by-Design

Problem Statement: Enterprise support teams manually curate knowledge base (KB) articles, but this process is unscalable, leading to **knowledge entropy**—a state where KBs become saturated with outdated or overlapping information. AI-driven systems offer automation but often lack the accuracy and mandatory human-in-the-loop (HITL) process, creating significant risks of propagating unverified information and failing to provide governance over the AI's impact on the knowledge lifecycle.

Solution: We propose a novel, multi-agent orchestration framework for the end-to-end intelligent automation of the enterprise knowledge lifecycle. This system systematically distills insights from resolved support tickets via three orchestrated agents: (1) a ***Ticket Analyzer*** that de-identifies and summarizes support interactions; (2) a ***KB Matching Tool*** that identifies the most relevant knowledge and assigns an actionable verdict; and (3) an ***Article Generation Tool*** that drafts or revises content accordingly. The core of our framework is the *KB Matching Tool*, which comprises three integrated components: **1. Context-Aware Reranking (CAR):** We use a two-component score to rerank an initial set of retrieved documents (d_i):

$$S_{\text{rerank}}(d_i) = \alpha \cdot \underbrace{\text{Sim}(q, d_i)}_{S_{\text{text}}} + \beta \cdot \underbrace{\frac{1}{K-1} \sum_{j=1, j\neq i}^{K} \text{Sim}(d_i, d_j)}_{S_{\text{context}}}$$

where q is the query, d_i is a candidate from the top-K set, and α, β are weighting factors. This approach is highly effective in corpora with content overlap, as it prioritizes documents that are not only relevant to the query (S_{text}) but also central to the context of other top candidates (S_{context}). **2. Actionable Verdict Engine (AVE):** The top-ranked article is fed into the *AVE*, a structured decision algorithm that assigns a verdict (*Update Article, Draft New, Article Exists,* or *Human Intervention*). Instead of a single classification prompt (baseline, *B*), the AVE utilizes a multi-step logical pipeline, which significantly improves accuracy by breaking the decision into simpler, verifiable sub-tasks. **3. AI Reliability Score for Governance:** For the HITL process, a dynamic confidence score, $R_t(d)$, is maintained for every article and updated via an Exponential Moving Average:

$$R_{t+1}(d) = (1 - \lambda)R_t(d) + \lambda F_t$$

where $R_t(d)$ is the score at time t, F_t is quantified by the semantic similarity between the suggested AI draft and the final SME-approved text, and λ is a learning rate (set to 0.1 to prioritize stability). This score provides a governance signal, enabling reviewers to prioritize their efforts on articles with lower reliability over-time.

Experimental Results: The system is powered by `GPT-4o`. Our *CAR* reranking method (T1) significantly improved retrieval effectiveness over a standard vector search baseline. In *AVE* powered by (T2), `GPT-4o` demonstrated superior overall accuracy, achieving **88.2%** accuracy across all categories & signigicantly improved from baseline (*B*) classification accuracy of **65.71%**. Currently in production, the system processes ∼1,500 tickets/week, reduced average authoring time by **75%** (from 40.3 to 10.3 minutes), and increased KB article generation efficiency by **67%**, yielding an estimated **$430k** in annualized savings.

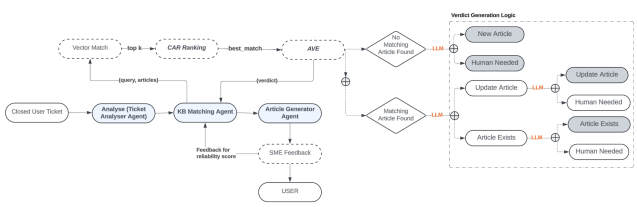

Figure 1: System architecture showing the decoupled pipeline for retrieval and governance.

(a) T1: Article Matching (Recall@K)

Metric	Baseline	Ours (CAR)
Recall@1	58.82%	**79.41%**
Recall@5	85.15%	**94.20%**

(b) T2: Verdict Generation Accuracy (%)

Model	Draft	Update	Overall
GPT-4o (Ours)	**70.7**	**80.3**	**88.2**
GPT-OSS 120B	70.3	70.32	75.00
Llama 3.1 405B	70.4	30.4	55.8
Mistral Large	50.1	70.2	60.6
Claude 3.5 Sonnet	80.0	60.1	70.5

Conclusion: Our production framework successfully bridges automated ticket resolution with knowledge creation at scale, providing a practical model for deploying generative AI safely and efficiently in the enterprise through essential human-in-the-loop governance.