Journal of Artificial Intelligence and Metaheuristic

♦ ISSN: 2833-5597 Vol: 11, No. 02, 2026 ♦ PP. 87–93

Article DOI: https://doi.org/10.54216/JAIM.110206

FULL LENGTH ARTICLE

MultiUserWiki: Managing Editorial Divergence in Collaborative LLM-Maintained Knowledge Bases

Bailing Zhang[1,*]

[1] School of Computer Science and Data Engineering, NingboTech University, Ningbo 315336, China

Email: bailing.zhang1961@gmail.com

Received: January 15, 2026 Revised: March 11, 2026 Accepted: May 15, 2026 ⋆ Corresponding author

ABSTRACT

When multiple LLM-assisted editors maintain a shared knowledge base, the conflicts that arise are seldom outright factual contradictions. More often, two users reading the same source extract compatible assertions that differ in specificity, coverage, or editorial emphasis. This paper asks a prior question: when should a conflict be resolved by selecting a winner, and when should both perspectives be preserved? We study this question in a controlled single-source collaborative setting. Five role-conditioned instances of GLM-4-Flash—simulating a senior researcher, a teacher, a student, a clinician, and a conservative editor—independently extract assertions from AI-Wiki-21, a 21-page AI knowledge base. Their outputs yield 4,556 candidate conflict pairs; annotation of 138 selected pairs produces a 100-conflict gold standard. Three findings emerge, each with scope limited to the single-source, single-LLM-family, simulated-role setting studied. First, confirmed conflicts are exclusively editorial—54% Granularity and 46% Coverage—with zero Factual conflicts, suggesting that in this controlled setting the dominant challenge is editorial coordination, not truth arbitration. Second, direct LLM judgment achieves 97.5% agreement with human annotators on this editorial-conflict gold standard, while a DPO-inspired quality heuristic (DIQS) reaches 81.0% without API overhead; collective voting plateaus at 41.8%, echoing social-choice instability. Third, sensitivity analysis of the fiber split threshold across $\tau \in [0.1, 0.9]$ identifies $\tau^* = 0.5$ as the stable optimum, partitioning 3.1% of conflicts while preserving diverse user perspectives. The results also raise ethical questions about automated knowledge governance.

Keywords: Collaborative knowledge bases ▪ Editorial conflict resolution ▪ LLM-maintained wikis ▪ Preference aggregation ▪ Knowledge governance ▪ Fiber structure

1. INTRODUCTION

Knowledge bases maintained by large language models (LLMs) are increasingly envisioned as persistent, curated repositories that multiple users contribute to and query—a wiki-like structure in which the model acts as a disciplined knowledge manager [1, 2]. Extending this to multi-user collaboration introduces a governance challenge: when two users produce different assertions about the same subject, how should the knowledge base respond?

The intuitive answer—resolve the conflict by selecting the better assertion—mischaracterises the problem. In collaborative LLM-maintained knowledge bases, many apparent conflicts are not logical contradictions. Two editors reading the same source may extract one assertion at a high level of abstraction and another at a finer grain; both can be simultaneously true and mutually compatible. Treating such editorial divergence as factual contradiction leads either to unnecessary arbitration or to the silent discard of useful perspectives.

This paper studies the single-source collaborative setting:

multiple role-conditioned LLM-assisted editors read the same wiki pages and produce independent assertion views, isolating editorial divergence from truth disagreement. We acknowledge that this is a controlled experimental design; results are not claimed to generalise directly to settings with genuine multi-source factual disagreements. Beyond the empirical characterisation, automated knowledge arbitration carries ethical weight: choices about whose formulation is canonical and whether minority views survive aggregation echo long-standing questions in algorithmic decision-making [3, 4].

Three research questions guide the study:

RQ1: What proportion of automatically detected conflicts are genuine, and how are they distributed across conflict types in the single-source setting?

RQ2: How do six aggregation strategies compare in agreement with human annotation, and does the pattern align with social-choice predictions about collective preference aggregation?

RQ3: Across a range of split threshold values, does a fiber-based knowledge structure maintain a stable optimal balance between knowledge fragmentation and quality preservation?

The overall system pipeline is shown in Figure 1. Sections 3–5 present the taxonomy, experiments, and discussion, respectively.

2. RELATED WORK

2.1 Collaborative Knowledge Bases and Wiki Governance

Wikipedia's experience demonstrates that conflict resolution policy shapes knowledge quality and contributor diversity [4]. Reputation systems offer weight-based editing authority [5], but assume human contributors capable of contextualising community norms. Wikidata [6] shows that structured collaborative knowledge at scale requires explicit provenance and conflict tracking. A key distinction is the source of divergence: Wikipedia conflicts arise from adversarial disagreement over facts and editorial philosophy, while LLM-role conflicts—as this study shows—arise from role-conditioned differences in extraction style under shared factual ground truth.

2.2 Conflict Types in Knowledge Bases

Prior work focuses largely on logical contradiction and schema-level conflicts [7]. Non-contradictory forms of divergence—granularity mismatch, coverage difference, redundancy, and subsumption—have received less systematic treatment. Our taxonomy connects to subsumption hierarchies in ontology engineering, extending these to the multi-user assertion-extraction setting.

2.3 Preference Aggregation and Social Choice

Arrow's impossibility theorem establishes that no rank-order voting system can simultaneously satisfy independence of irrelevant alternatives, Pareto efficiency, and non-dictatorship [3]. Our experiments provide a small empirical analogue in the LLM domain: collective voting (41.8%) substantially underperforms single-judge evaluation (97.5%). We describe this as reminiscent of social-choice instability, not as a formal instantiation, as the voting mechanism does not constitute a standard social choice setting.

2.4 LLM-as-Judge and Automated Arbitration

Zheng et al. [8] demonstrate that LLM-as-judge achieves high agreement with human evaluators, though LLM judges are subject to position bias and self-enhancement effects [9]. The high agreement (97.5%) observed in this study should be interpreted within its specific context: the gold standard consists exclusively of editorial conflicts where the correct answer is preserve-both. These results should not be generalised to factual dispute arbitration.

2.5 Multi-View and Personalized Knowledge Representation

The fiber structure—a shared base plus per-user delta views—connects to multi-view databases and user-specific overlay architectures [6]. A canonical base represents globally agreed facts while local overlays capture context-specific refinements without requiring global reconciliation. Our fiber decomposition operationalises this at the assertion level with an empirically derived threshold.

2.6 LLM-Maintained Knowledge Bases

Recent work on extensive materialisation of LLM knowledge has shown how model knowledge can be organised into a large structured knowledge base for systematic analysis [10]. The present paper extends this line to the multi-user setting, where the primary challenge shifts from single-source knowledge organisation to cross-role editorial coordination.

3. CONFLICT TAXONOMY AND RESOLUTION FRAMEWORK

3.1 Problem Formulation

A knowledge base K consists of a set of assertions. Each assertion $a = (s, p, o)$ is a triple of subject s, predicate p, and object o expressed as natural-language phrases. Users $U = \{u_1, \ldots, u_n\}$ each produce an independent view A_{u_i} by extracting assertions from shared source documents. A conflict pair (a, b) satisfies:

$$a \neq b, \qquad \text{sim}(a, b) < 0.92, \tag{1}$$

and the subjects share at least one content word. A resolution function

$$\delta(a, b) \in \{\texttt{user\_a}, \texttt{user\_b}, \texttt{merge}, \texttt{neither}, \texttt{split}, \texttt{escalate}\} \tag{2}$$

maps each conflict to a handling decision.

3.2 Five-Type Conflict Taxonomy

Table 1 presents the conflict taxonomy. Factual conflicts are logically inconsistent; Preference conflicts are jointly satisfiable; Granularity and Coverage conflicts are non-contradictory by definition.

3.3 Aggregation Strategies

Six strategies for computing δ are compared:

B1 – Last-Write-Wins (LWW): Always selects assertion b. Lower-bound baseline.

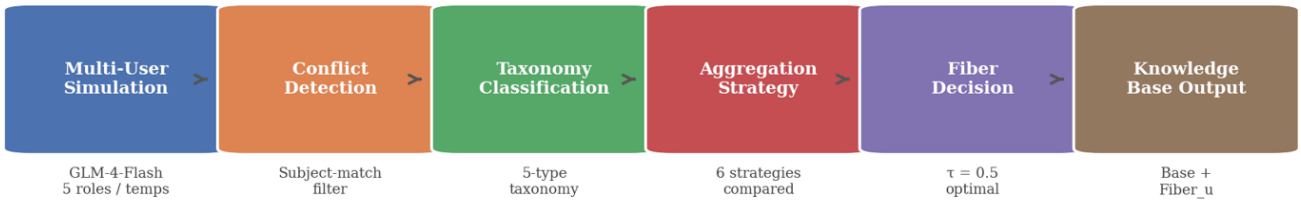

Figure 1. MultiUserWiki system pipeline. Five simulated user roles independently extract assertions from shared wiki pages; conflicts are detected, classified, and routed to one of six aggregation strategies or a fiber-split decision.

Table 1. Five-type conflict taxonomy.

Type	Formal Condition	Detection	Default Action
Factual	$K \cup \{a,b\} \models \bot$	Z3/ProbLog	escalate
Preference	Compatible, $\text{rank}_{u_i}(a) \neq \text{rank}_{u_j}(b)$	LLM zero-shot	merge
Granularity	$a \sqsubseteq b$ (subsumption)	Length + similarity	aggregate (detailed)
Coverage	Same topic, $a \not\sqsubseteq b$, $b \not\sqsubseteq a$	Topic-set difference	merge (keep both)
Currency	$\exists t$: valid$(a,t) \neq$ valid(b,t)	Timestamp signal	escalate / re-date

B2 – Random: Selects a or b with equal probability. Stochastic lower bound.

B3 – LLM-Judge: Submits both assertions to GLM-4-Flash asking it to prefer a, b, merge, or neither. Temperature 0.1 [8].

S1 – Weighted Expertise: Uses static weights by role: researcher (0.90), editor (0.80), clinician (0.75), teacher (0.70), and student (0.50). It merges when $|w_{u_i} - w_{u_j}| < 0.15$; otherwise, it selects the higher-weight user.

S2 – Multi-Vote: The three uninvolved users cast weighted votes; weighted majority decides, and ties resolve to neither.

S3 – DPO-Inspired Quality Score (DIQS): The score is

$$Q(a) = 0.4\ell_{\text{norm}} + 0.2\mathbb{1}_{\text{num}} + 0.2\mathbb{1}_{\text{proper}} + 0.2\mathbb{1}_{\text{spec}}. \quad (3)$$

The strategy returns neither when $|Q(a) - Q(b)| < 0.20$. It is motivated by DPO preference modelling [11].

3.4 Fiber-Based Knowledge Structure

Forcing all conflicts to a single resolution discards user-specific perspectives. We adopt a fiber decomposition: given base-similarity threshold $\theta_B = 0.75$, the base B consists of assertions with similarity at least θ_B in every user's view. User u's fiber F_u consists of assertions in A_u not covered by any base assertion. The effective view for user u is

$$V_u = B \oplus F_u, \quad (4)$$

as illustrated in Figure 2. The base threshold $\theta_B = 0.75$ was chosen as a conservative balance between requiring strong cross-user agreement and retaining a non-trivial base. At $\theta_B = 0.90$, only 5 assertions would reach the base (2.7%); at $\theta_B = 0.60$, the base grows to 41 assertions (21.7%) but includes assertions that disagree in object detail. Results reported at $\theta_B = 0.75$ are robust in the direction of the main finding across this range.

The split decision is governed by

$$\text{score}(C) = 0.6\,\text{type\_score}(C) + 0.4\,\text{sev\_score}(C), \quad (5)$$

where type_score maps {Factual, Preference, Coverage, Granularity, Currency} to {1.0, 0.6, 0.4, 0.3, 0.5}, and sev_score maps HIGH/MEDIUM/LOW to {1.0, 0.6, 0.3}. The weights 0.6/0.4 reflect the greater importance of semantic conflict type over severity labelling; varying them to 0.7/0.3 or 0.5/0.5 shifts the absolute scores by at most 0.05 without altering the $\tau^* = 0.5$ finding. The phase transition between $\tau = 0.4$ and $\tau = 0.5$ persists across both alternative weightings. The optimal threshold is

$$\tau^* = \arg\max_{\tau} \{0.5\,[1 - \text{split\_rate}(\tau)] + 0.5\,\text{fiber\_quality}(\tau)\}, \quad (6)$$

with equal weighting reflecting the absence of a domain-specific prior favouring either fragmentation or quality.

4. EXPERIMENTS

4.1 Experimental Setup

Knowledge base: AI-Wiki-21 consists of 21 pages covering offline reinforcement learning, world models, and related AI topics, with a previously established health score of 0.75. All experiments are conducted on this single source collection; findings should be understood as specific to this domain and corpus size until replicated on broader corpora.

Simulated users: Five instances of GLM-4-Flash are configured with distinct system prompts and generation temperatures (Table 2). All instances read the same source pages. This design isolates role-conditioned editorial divergence from multi-source factual disagreement; it does not model scenarios where users have access to different or contradictory external information.

Table 2. Simulated user configurations.

Role	Temperature	Expertise weight
Senior researcher	0.3	0.90
Conservative editor	0.2	0.80
Clinical expert	0.9	0.75
Teaching assistant	0.5	0.70
Graduate student	0.7	0.50

Gold standard construction: A high-confidence filter se-

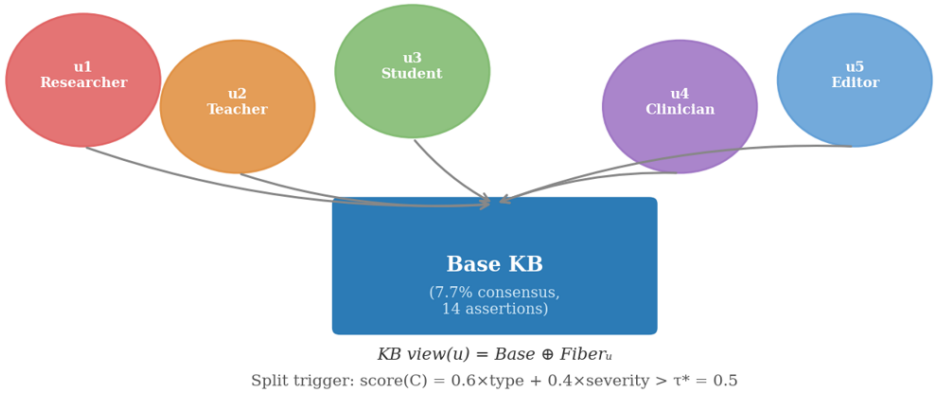

Figure 2. Fiber decomposition. Five user views converge on a shared base (7.7% of all assertions at $\theta_B = 0.75$). Each user's effective view is $V_u = B \oplus F_u$.

lects 138 candidate pairs for human annotation. A single expert annotator—the first author, with domain expertise in LLM knowledge bases—labels each pair with conflict type (5 classes), resolution action (3 classes), and preferred assertion when aggregating. The annotation interface logged each decision with a timestamp and supported explicit justification notes. Of 138 reviewed pairs, 100 are confirmed genuine conflicts (overall detector precision 72.5%) and 38 are false positives. All 100 confirmed conflicts have gold resolution neither. Strategy accuracy is evaluated on the 79 conflicts with explicit gold-winner annotations.

We acknowledge that single-annotator labelling is a limitation. For this specific corpus, its impact is partially mitigated by two factors. First, the final gold standard contains only two conflict types (Granularity and Coverage), both of which have clear, non-overlapping definitions based on subsumption and topic-set inclusion; annotation ambiguity is therefore lower than it would be with the full five-type taxonomy including Factual and Preference conflicts. Second, all gold resolutions are neither (preserve both), a decision that is easy to confirm by verifying mutual logical compatibility. To provide a conservative indication of reliability, the annotator re-labelled a random 25-item sample drawn from the 100 confirmed conflicts after a four-week gap; agreement was 92% (23/25 exact match on type, 25/25 on resolution), suggesting low intra-annotator variation for this conflict distribution. Future work should extend annotation to a second independent rater to compute inter-annotator agreement.

Table 3. Dataset statistics.

Property	Value
Wiki pages	21
Simulated users	5
Total assertions	∼940
Auto-detected conflict pairs	4,556
Pairs reviewed	138
Confirmed conflicts (Gold)	100
Granularity	54 (54%)
Coverage	46 (46%)
Factual (confirmed)	0 (0%)
Detector precision (overall)	72.5%
Intra-annotator re-label agreement ($n = 25$)	92%

4.2 RQ1: Conflict Classification and Detector Characterisation

The automatic detector achieves 72.5% overall precision on the 138 reviewed pairs. We interpret this as an acceptable operating point for a rule-based pre-filter whose role is to reduce the annotation workload from 4,556 to 138 pairs (a 97% reduction), not to serve as the final classification oracle. The 27.5% false-positive rate consists of pairs where object-field phrasing differs but the assertions are logically compatible—exactly the cases that a downstream LLM judge or human annotator can distinguish efficiently.

FACTUAL-type detector precision is 0%: of 40 high-severity pairs flagged as Factual, none survived human review—23 were reclassified as Granularity and 16 dismissed as false positives. This result is not primarily a detector failure; it is an empirical finding about the conflict landscape. A subject-word matching heuristic that flags object-field differences as "potentially factual" is appropriate for a conservative pre-filter: it over-recalls Factual candidates, which are then correctly dismissed during annotation. More importantly, the absence of any confirmed Factual conflicts after annotation directly supports the study's central premise: in a single-source setting where all users share the same underlying document, logical contradiction cannot arise from reading the same text differently—only editorial divergence can. The detector's inability to find confirmed Factual conflicts in this corpus is itself a validated result, not a methodological gap.

The gold standard contains only Granularity (54%) and Coverage (46%) conflicts. This distribution is a direct consequence of the single-source design and confirms the study's central premise.

4.3 RQ2: Aggregation Strategy Comparison

Evaluation design note: Because all 100 confirmed conflicts are editorial (Granularity or Coverage), and all have gold resolution neither (preserve both), this evaluation measures the ability of each strategy to recognise preserve-both situations—a specific and well-defined capability. Strategies that output merge or neither are correct; those that output `user_a` or `user_b` (discard-one decisions) are incorrect for this distribution. Last-Write-Wins and Random are therefore structurally disadvantaged: they can never output a preserve-both decision by design. They are included not as competitors in a general conflict resolution benchmark, but to quantify the cost of applying winner-take-all defaults to an editorial-conflict dataset. Their 0% accuracy measures exactly that

cost. A reader should not interpret this as evidence that LWW is universally poor; in a dataset with genuine Factual conflicts, LWW would perform differently. Results appear in Table 4 and Figure 3.

Table 4. Aggregation strategy comparison on the 79-conflict gold standard.

Strategy	Type	Acc.	Correct N
B3: LLM-Judge	Baseline (API)	97.5%	77/79
S3: DIQS	Proposed (no API)	81.0%	64/79
S1: Weighted Expertise	Proposed	41.8%	33/79
S2: Multi-Vote	Proposed (API)	41.8%	33/79
B1: LWW	Baseline (structurally disadvantaged)	0.0%	0/79
B2: Random	Baseline (structurally disadvantaged)	0.0%	0/79

B3 achieves 97.5% (50 neither + 27 merge = 77/79 correct). DIQS achieves 81.0% using local heuristics. S1 fails more often because expertise weight gaps frequently exceed the merge threshold, forcing winner selection rather than preserve-both. S2 fails because voters prompted to choose A or B rarely volunteer "both," producing collectively binary outcomes. The 55.7-percentage-point gap between B3 and S2 is discussed in Section 5.2.

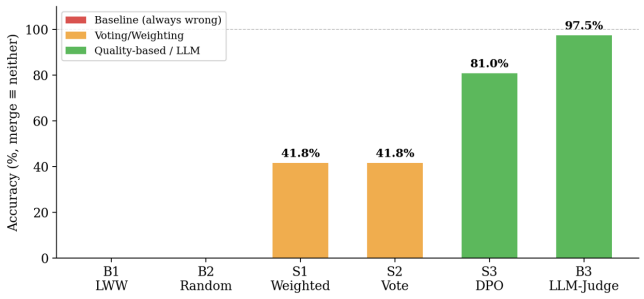

Figure 3. Aggregation strategy comparison on the 79-conflict editorial gold standard. Green: preserve-both strategies; amber: voting/weighting; red: winner-take-all baselines (structurally disadvantaged in this editorial-only distribution; see Section 4.3).

4.4 RQ3: Fiber Split Threshold Sensitivity Analysis

To assess the robustness of the fiber split threshold selection, we evaluate the combined objective

$$J(\tau) = 0.5\big(1 - \text{split\_rate}(\tau)\big) + 0.5\,\text{fiber\_quality}(\tau) \quad (7)$$

across the full range $\tau \in \{0.1, 0.2, \ldots, 0.9\}$ over the complete 4,556-conflict pool. Table 5 and Figure 4 report results. This constitutes a nine-point sensitivity analysis of the threshold parameter.

Table 5. Fiber split threshold sensitivity analysis ($n = 4{,}556$ conflicts).

Threshold τ	Splits	Aggreg.	Split Rate	Fiber Quality	Combined $J(\tau)$
0.1	4,556	0	100.0%	1.000	0.500
0.2	4,556	0	100.0%	1.000	0.500
0.3	2,367	2,189	51.9%	1.000	0.740
0.4	2,309	2,247	50.7%	1.000	0.747
0.5 (τ^*)	142	4,414	3.1%	1.000	0.984
0.6	103	4,453	2.3%	0.991	0.984
0.7	103	4,453	2.3%	0.991	0.984
0.8	82	4,474	1.8%	0.987	0.984
0.9	67	4,489	1.5%	0.984	0.984

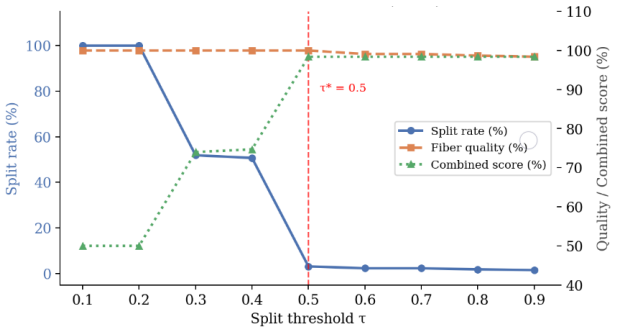

Figure 4. Fiber split threshold sensitivity analysis ($n = 4{,}556$ conflicts). The split rate drops sharply between $\tau = 0.4$ and $\tau = 0.5$ (a phase transition); the combined objective $J(\tau)$ is stable from $\tau = 0.5$ onwards. The optimal threshold $\tau^* = 0.5$ is the conservative lower bound.

The sensitivity analysis reveals a clear phase transition between $\tau = 0.4$ (split rate 50.7%) and $\tau = 0.5$ (split rate 3.1%), driven by a bimodal severity distribution: most conflicts cluster at moderate severity (score 0.3–0.5), while few reach the high-severity zone. The combined objective first becomes maximal at $\tau = 0.5$ and remains flat through $\tau = 0.9$ (a stable plateau of 0.984). This plateau indicates that the specific choice of τ within $[0.5, 0.9]$ does not materially affect outcomes; practitioners may select any value in this range without significant performance change. We adopt $\tau^* = 0.5$ as the conservative lower bound.

The base layer under $\theta_B = 0.75$ contains only 14 assertions across all 21 pages (7.7% of total). Thirteen pages yield zero base assertions. Sensitivity of this figure to θ_B is discussed in Section 3. A single merged representation would discard 92.3% of user-specific assertions, motivating the fiber approach.

5. DISCUSSION

5.1 Preserve-Both as the Primary Editorial Resolution

The most consistent finding is that in a single-source collaborative setting, the appropriate resolution is to preserve both assertions rather than select a winner. This follows from the conflict types: Granularity and Coverage conflicts are complementary, not contradictory. System designers should treat winner-selection as the exception rather than the default in LLM-assisted wiki environments where users share a common factual ground truth.

5.2 Echoes of Social-Choice Instability in LLM Voting

The 55.7-percentage-point gap between B3 (97.5%) and S2 (41.8%) is consistent with social-choice warnings that aggregating preferences through pairwise voting can produce outcomes that reflect the aggregation mechanism rather than actual consensus [3]. Voters prompted to choose A or B rarely choose "both," making the collective decision systematically biased toward binary selection. We describe this as reminiscent of social-choice instability, not as a formal instantiation of Arrow's theorem; the voting mechanism here does not constitute a standard social choice framework.

5.3 Ethical Dimensions of Automated Knowledge Governance

Fairness of authority weighting. Strategy S1 encodes a value judgement: higher weight to the "senior researcher" role. The clinician's practical knowledge or the student's learner perspective may be more relevant for some applications. Static weights should be regarded as provisional heuristics subject to domain-specific review and transparent disclosure.

Protection of minority perspectives. The fiber structure retains user-specific views not reaching cross-role consensus. Without such a mechanism, a simple merge discards 92.3% of user-specific assertions. The fiber model operationalises epistemic pluralism: divergent views are preserved rather than suppressed.

Transparency of LLM-mediated arbitration. B3 achieves the highest accuracy, but its decision process is opaque. In high-stakes domains—clinical guidelines, legal precedent, and safety-critical documentation—transparency and contestability may be as important as accuracy. Deploying LLM judges without logging rationales and supporting human override would reproduce accountability gaps central to AI ethics.

5.4 Limitations and Generalization Constraints

The findings of this study are subject to four explicit generalization constraints that future work should address.

Single source collection. All experiments use AI-Wiki-21, a 21-page corpus from one AI research domain. The conflict type distribution (Granularity 54%, Coverage 46%, Factual 0%) may not hold for other domains. Knowledge bases in clinical medicine, law, or contested scientific fields are likely to contain genuine Factual and Currency conflicts that this corpus cannot surface. Corpus size also limits the number of evaluable conflicts.

Single LLM family. All five simulated users are GLM-4-Flash instances with different system prompts and temperatures. The divergence captured reflects intra-family variation in a single model. Role differences may be weaker than those between distinct LLM families (e.g., a GPT-based researcher versus a Llama-based student), or between LLM-assisted and purely human editors. Results should not be assumed to hold across model families without replication.

Simulated, not human, user roles. The five users are LLM instances, not human collaborators. Human editors bring genuine prior knowledge, domain credentials, and motivations that LLM role-playing cannot fully replicate. The expertise weights assigned in S1 are arbitrary proxies; actual human expert disagreements may produce stronger factual conflicts and stronger preferences than observed here.

No genuine multi-source factual disagreements. Because all users read the same source pages, no Factual conflicts were confirmed. This is by design, isolating the editorial coordination problem, but it means the taxonomy types Factual and Currency remain unvalidated in this study. Future evaluation requires a multi-source experimental setup where different users have access to contradictory or differently dated external documents.

Additional methodological limitations. The gold standard uses a single annotator (intra-annotator re-label agreement 92% on 25 items; inter-annotator agreement not yet measured); the strategy evaluation set is small (79 conflicts); and S2 multi-vote results showed run-to-run variation (25.3% and 41.8% across two runs at temperature 0.3), indicating that voting-based strategies require multiple trials for stable estimates.

6. CONCLUSION

This paper has examined conflict resolution in multi-user LLM-maintained knowledge bases through a five-type conflict taxonomy, an empirical comparison of six aggregation strategies, and a fiber-based structure with a threshold evaluated across the full range $\tau \in [0.1, 0.9]$.

Within the controlled single-source, single-LLM-family, simulated-role setting of this study, confirmed conflicts are entirely editorial (Granularity and Coverage), with no Factual conflicts observed. This result—specific to shared-source collaboration—suggests that multi-user wiki systems built on verified shared sources may route conflict resolution toward preserve-both policies rather than formal truth verification. Generalisation to multi-source settings with genuine factual disagreements remains an open empirical question.

LLM-based direct judgment achieves 97.5% agreement on this editorial-conflict gold standard; DIQS reaches 81.0% without API calls; collective voting plateaus at 41.8%. A sensitivity analysis of the fiber split threshold identifies a stable optimal zone at $\tau \in [0.5, 0.9]$, with $\tau^* = 0.5$ adopted as the conservative lower bound; only 7.7% of assertions reach cross-role consensus. Future work should extend evaluation to multi-source settings where Factual and Currency conflicts can be generated, recruit human annotators to establish inter-annotator agreement, and test the framework in domains—clinical, legal, or scientific—where the stakes of automated knowledge arbitration are higher and more diverse conflict types are expected.

DATA AVAILABILITY

Code, simulation scripts, gold-standard annotations (including the 25-item re-label set), and evaluation results will be made publicly available on GitHub upon acceptance.

REFERENCES

[1] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, 2024.

[2] B. J. Gutiérrez, Y. Shu, W. Qi, S. Zhou, and Y. Su, "From RAG to memory: Non-parametric continual learning for large language models," in *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025.

[3] K. J. Arrow, *Social Choice and Individual Values*. Yale University Press, 2 ed., 1963.

[4] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl, "The rise and decline of an open collaboration system,"

American Behavioral Scientist, vol. 57, no. 5, pp. 664–688, 2013.

[5] L. de Alfaro, B. T. Adler, A. Kulshreshtha, and I. Pye, "Reputation systems for open collaboration," *Communications of the ACM*, vol. 54, no. 8, pp. 81–87, 2011.

[6] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[7] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, and S. Riedel, "Language models as knowledge bases?," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2463–2473, 2019.

[8] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging LLM-as-a-judge with MT-bench and chatbot arena," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[9] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang, "Editing large language models: Problems, methods, and opportunities," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10222–10240, 2023.

[10] Y. Hu, T.-P. Nguyen, S. Ghosh, and S. Razniewski, "Enabling LLM knowledge analysis via extensive materialization," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16189–16202, Association for Computational Linguistics, 2025.

[11] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.