http://dx.doi.org/10.62762/NGCST.2026.503855

RESEARCH ARTICLE

WikiConstraintCalibration: Content-Aware Cost Preference Estimation for Constraint Threshold Selection in Wiki-Grounded LLM Agents

Bailing Zhang[1,*]

[1] Institute of Data Science and Intelligent Technology, NingboTech University, Ningbo 315100, China

Abstract

Wiki-grounded LLM agents enforce output constraints through safety rules and scope boundaries derived from structured knowledge bases, yet selecting an appropriate constraint scorer threshold θ remains challenging. Since outputs are blocked when $s(o) \geq \theta$, a low threshold may over-restrict safe responses, whereas a high threshold may admit dangerous outputs. Existing approaches largely rely on manual, domain-agnostic tuning without systematically incorporating wiki content. This paper proposes `WikiConstraintCalibration`, a framework that extracts eight wiki-derived features spanning hazard indicators (dangerous assertion density, immediate-action fraction, domain risk prior, and high-risk constraint fraction) and reliability indicators (citation rate, completeness, staleness mean, and stale fraction). These features determine a domain-specific cost ratio $\rho = C_{fn}/C_{fp}$, representing the relative cost of missing dangerous outputs versus blocking safe ones. The ratio determines a safety weight w_s for navigating an empirical Safety–Function Pareto frontier, while constrained optimisation identifies a recommended operating region $[\theta_l, \theta_h]$. A dynamic module further incorporates staleness signals to adapt class-conditional response policies. Experiments across four deployed wiki-agent domains show that the cost ratio captures domain risk differences, ranging from 5.76 for GP medical to 1.89 for AI course. The GP domain selects the lowest feasible threshold, $\theta^* = 0.32$, driven by a score-distribution cliff near $\theta \approx 0.30$. Ablation analysis identifies domain risk prior as the dominant calibration factor, while content-level features provide additional discrimination. Overall, `WikiConstraintCalibration` provides an auditable, content-grounded approach to threshold calibration by linking wiki characteristics and domain risk to empirical safety–function trade-offs. Future work will address limited test-set sizes and domain-generic safety judging through human annotation and domain-specific scoring.

Keywords: LLM agents, constraint calibration, wiki, Pareto frontier, cost-sensitive classification, knowledge freshness.

Submitted: 07 May 2026
Revised: 08 June 2026
Accepted: 05 August 2026
Published: 05 September 2026

Vol. 2, **No.** 3, 2026.
10.62762/NGCST.2026.503855

***Corresponding author:**
✉ Bailing Zhang
bailing.zhang1961@gmail.com

Citation
Zhang, B. (2026). WikiConstraintCalibration: Content-Aware Cost Preference Estimation for Constraint Threshold Selection in Wiki-Grounded LLM Agents. *Next-Generation Computing Systems and Technologies*, 2(3), 99–108.

1 Introduction

Augmenting large language models with structured knowledge bases is an established strategy for constraining agent behaviour in high-stakes domains such as medical question answering and legal advisory services [1]. In this architecture the knowledge base — here called a *wiki* — encodes domain-specific facts, safety rules, and scope constraints; at runtime a constraint scorer $s(\cdot)$ evaluates each candidate output and the agent applies a threshold θ to decide whether to block, warn, or pass.

Because the blocking condition is $s(o) \geq \theta$, a low θ flags most outputs as risky (restrictive), while a high θ permits most outputs through (permissive). Finding the right operating point requires balancing two competing objectives: *safety* — the fraction of genuinely dangerous queries that are correctly blocked — and *function* — the fraction of genuinely safe queries that are correctly passed. This trade-off is domain-specific. A GP medical agent handling patient consultations demands an aggressive safety weight, warranting a low θ even at some cost to function. A language training agent, by contrast, can tolerate a higher θ without meaningful safety risk. Yet current practice sets θ manually without a principled connection to the wiki that defines the constraint rules.

This paper addresses that gap. The central observation is that the appropriate operating region can be inferred from quantifiable properties of the wiki itself. A wiki with high dangerous assertion density, poor citation support, and elevated staleness warrants a more restrictive operating point; a wiki covering predominantly educational content with stable, well-cited knowledge can afford a more permissive setting. We formalise this as a cost-sensitive optimisation problem and propose WikiConstraintCalibration, a framework that derives a cost ratio from wiki features, uses it to weight a Safety–Function Pareto frontier analysis, and recommends a calibration region rather than a single threshold.

The key contribution is estimating the domain-specific cost preference from wiki content. The frontier is measured on a small validation set, but the weight applied to navigate it — and hence the recommended region — is driven by the wiki's content features, not by further labelled data.

Three contributions are claimed. First, a structured 8-dimensional wiki feature vector is defined, grouping features into hazard indicators and reliability indicators; both groups influence the cost ratio, though through different mechanisms. Second, a constrained utility maximisation over the Pareto frontier yields a calibration region with explicit safety and function properties. Third, a dynamic module integrates WikiMonitor-style staleness signals to update class-conditional response policies when knowledge freshness changes.

2 Related Work

2.1 Wiki-Grounded and Guardrail-Augmented Agents

The idea of grounding LLM outputs in structured knowledge bases has been explored in retrieval-augmented generation [2] and more recently in wiki-style agent architectures [1]. Parallel work has studied safety filtering at the output layer through purpose-built guard models. Inan et al. [3] demonstrated that a fine-tuned LLM can classify input-output pairs as safe or unsafe with reasonable accuracy, but noted that such classifiers require domain-appropriate calibration to avoid systematic over-rejection or under-rejection. Bai et al. [4] showed that reinforcement learning from human feedback can improve the trade-off between helpfulness and harmlessness, demonstrating that the tension between safety and utility is a persistent structural property of aligned LLM systems rather than an artefact of any particular training objective. This observation motivates the need for principled operating-point selection at deployment time, which the present work addresses in the specific context of wiki-grounded agents, where the wiki's content encodes the domain's risk profile.

2.2 Threshold Calibration and Selective Prediction

Calibrating a classifier's decision threshold to minimise expected cost under a known cost matrix is well-established [5]. Post-hoc probability calibration methods such as Platt scaling [6] and isotonic regression produce well-calibrated scores that facilitate threshold selection but still require the practitioner to specify the cost ratio externally. Conformal risk control [7] provides distribution-free guarantees on user-specified loss functions, but presupposes a calibration set drawn from the deployment distribution. Neither approach provides a principled way to estimate the cost ratio from the knowledge base. Our framework fills this gap: the cost ratio is derived from wiki features, while the operating region is selected on a measured frontier,

combining both perspectives.

2.3 Safety Scoring and Cost-Sensitive Framing

The cost-sensitive classification framework of Elkan [5] underpins our cost ratio formulation. Recent work on LLM output safety [8] embeds harmlessness objectives directly into the training process through AI-generated feedback, reducing reliance on human annotation for safety supervision. However, training-time alignment does not eliminate the need for deployed systems to reason about the asymmetric costs of false positives (over-restriction) and false negatives (missed harmful outputs) at inference time — a gap that threshold calibration addresses. Our work provides this reasoning layer for wiki-constrained agents specifically.

3 Method

3.1 Problem Formulation

Let $\mathcal{W}$ be a wiki consisting of N entries and an optional formal rule file $\mathcal{R}$. A constraint scorer $s : \mathcal{O} \to [0, 1]$ assigns a risk score to each agent output $o \in \mathcal{O}$, where higher scores indicate greater risk. The agent blocks o when $s(o) \geq \theta$. Because the block condition is a lower-bounded exceedance, a smaller θ is more restrictive and a larger θ is more permissive.

Given a test set with n_B block-labelled (dangerous) outputs and n_P pass-labelled (safe) outputs, Safety and Function at threshold θ are:

$$\text{Safety}(\theta) = \frac{TP(\theta)}{n_B}, \quad \text{Function}(\theta) = \frac{TN(\theta)}{n_P}. \tag{1}$$

Ambiguous (warn-labelled) outputs are excluded from both metrics. As θ decreases, Safety approaches 1 and Function falls; as θ increases, Function approaches 1 and Safety falls. The Pareto frontier traces this trade-off.

The calibration task is: given $\mathcal{W}$ and a small validation set on which the frontier is measured, estimate the domain's cost preference and select an operating region.

3.2 Wiki Feature Extraction

We extract eight features from $\mathcal{W}$, grouped into two conceptually distinct sets.

Hazard features ($\mathcal{H}$) capture the intrinsic danger level of the wiki's content:

- f_1: proportion of assertions containing domain red-flag keywords or high-risk Prolog predicates.
- f_2: proportion of warning-level assertions requiring immediate clinical action. Computed with Laplace smoothing to handle low red-flag counts:

$$f_2 = \frac{k + \alpha}{n + 2\alpha}, \tag{2}$$

  where k is the count of immediate assertions, n the total red-flag assertion count, and $\alpha = 2$ a smoothing constant.
- f_7: domain risk prior, expert-assigned on $[0, 1]$ (e.g., 0.92 for GP medical, 0.20 for an AI course).
- f_8: proportion of constraint entries in high-stakes categories (diagnosis, dosage, red-flag) as opposed to educational content.

Reliability features ($\mathcal{U}$) capture how trustworthy the wiki's knowledge is:

- f_3: proportion of entries with traceable evidence citations.
- f_4: proportion of entries satisfying required structural fields (case ID, title, source documents, presentation).
- f_5: mean staleness score from a freshness monitor, or estimated from last-reviewed dates.
- f_6: proportion of entries labelled stale.

3.3 Cost Ratio Formula

The hazard component of the cost ratio captures the raw danger level of the wiki domain:

$$\rho_{\text{haz}}(\mathbf{f}) = \rho_0 + w_1 f_1 + w_2 f_2 + w_7 f_7 + w_8 f_8, \tag{3}$$

with $\rho_0 = 1.0$ and weights $(w_1, w_2, w_7, w_8) = (2.0, 1.5, 3.0, 1.0)$.

The reliability component acts as an epistemic uncertainty multiplier: a wiki with low citation support, poor completeness, or high staleness warrants greater caution beyond what the hazard level alone implies:

$$u(\mathbf{f}) = v_3(1 - f_3) + v_4(1 - f_4) + v_5 f_5 + v_6 f_6, \tag{4}$$

with reliability weights (v_3, v_4, v_5, v_6). The full cost ratio combines both:

$$\rho(\mathbf{f}) = \rho_{\text{haz}}(\mathbf{f}) \cdot \big(1 + \lambda\, u(\mathbf{f})\big). \tag{5}$$

In the current experiments we set $\lambda = 0$ and incorporate staleness directly into the hazard component by assigning $w_5 = 1.0$ to f_5 in place of the reliability multiplier, yielding:

$$\rho(\mathbf{f}) = \rho_0 + w_1 f_1 + w_2 f_2 + w_5 f_5 + w_7 f_7 + w_8 f_8. \tag{6}$$

This simplification is reported for transparency; learning λ from data is identified as a priority in future work.

Monotonicity. Equation (6) satisfies the natural assumption that false-negative cost should not decrease when any hazard or staleness indicator increases: all five weights are strictly positive. The plateau stability observed in Sections 4.4 and 4.5 shows that removing or perturbing individual features does not alter the recommended region for the four experimental domains.

From ρ the safety and function weights follow the cost-sensitive classification framework [5]:

$$w_s = \frac{\rho}{\rho + 1}, \qquad w_f = 1 - w_s. \tag{7}$$

3.4 Pareto Frontier and Recommended Region

For each domain, the Safety–Function frontier is measured by sweeping θ over 45 uniformly spaced points in $[0.10, 0.98]$. For each threshold, we define the weighted utility

$$U(\theta) = w_s \,\mathrm{Safety}(\theta) + w_f \,\mathrm{Function}(\theta). \tag{8}$$

Candidate operating points must satisfy the minimum-function constraint

$$\mathrm{Function}(\theta) \geq \epsilon, \qquad \epsilon = 0.30. \tag{9}$$

This constraint prevents the optimiser from selecting thresholds that block an excessive proportion of genuinely safe outputs.

We first identify the set of feasible thresholds attaining the maximum utility:

$$\Theta^* = \underset{\theta:\,\mathrm{Function}(\theta)\geq\epsilon}{\arg\max} \; U(\theta). \tag{10}$$

When multiple feasible thresholds attain the same maximum utility, we apply a conservative tie-breaking rule and select the smallest threshold:

$$\theta^* = \min \Theta^*. \tag{11}$$

Because lower thresholds are more restrictive under the blocking rule $s(o) \geq \theta$, this tie-break favours the more restrictive operating point only when multiple feasible thresholds attain the same maximum utility.

To characterise near-optimal operating points, we define the recommended threshold set as

$$\Theta_{\mathrm{rec}} = \left\{\theta_j : \mathrm{Function}(\theta_j) \geq \epsilon,\, U(\theta_j) \geq U(\theta^*) - \delta\right\}, \tag{12}$$

where θ_j denotes one of the evaluated threshold grid points and $\delta = 0.03$ is the utility tolerance. The reported recommended operating region is the envelope of this discrete set:

$$\theta_l = \min \Theta_{\mathrm{rec}}, \qquad \theta_h = \max \Theta_{\mathrm{rec}}. \tag{13}$$

Accordingly, the recommended operating region is reported as $[\theta_l, \theta_h]$. Because the frontier is evaluated on a discrete threshold grid, this interval should be interpreted as the envelope of the evaluated near-optimal operating points rather than as a guarantee that every unevaluated threshold within the interval has identical Safety, Function, or utility.

It is important to clarify the relationship between the safety weight and the selected θ^*. A high w_s places greater emphasis on Safety in the utility function but does not directly determine the numerical value of θ. The selected θ^* is therefore not necessarily the globally lowest feasible threshold; rather, it is the smallest threshold among the feasible operating points that attain the maximum weighted utility. The minimum-function constraint determines which thresholds are admissible, while the empirical score distributions determine the shape of the Safety–Function frontier. If PASS queries cluster near a particular score s_P, thresholds below this region may over-block safe outputs and become infeasible. This can create a domain-specific score-distribution cliff that interacts with, but is distinct from, the safety weight.

3.5 Dynamic Calibration

WikiMonitor-style freshness monitoring periodically reports staleness scores for wiki entries. When staleness increases, the appropriate response differs by constraint class. For high-risk constraint classes (red-flag, diagnosis, dosage), a moderate increase in staleness triggers a request for stricter handling. This may result in a lower threshold when a more restrictive feasible operating point exists on the measured Safety–Function frontier. When no such point is available, the numerical threshold remains unchanged and the escalation is implemented through the response policy instead. For lower-risk classes, increased staleness primarily triggers stronger evidence requirements rather than threshold reduction. Algorithm 1 formalises this class-conditional escalation policy.

When the static optimisation identifies a different feasible threshold under updated staleness conditions, the numerical threshold is smoothed using an

exponential moving average with coefficient $\beta = 0.3$: $\tilde{\theta}_t = \beta\theta_t^* + (1-\beta)\tilde{\theta}_{t-1}$. If the utility-optimal feasible plateau does not change, then θ_t^* remains constant and the dynamic response is expressed through policy escalation rather than threshold movement. In the current four-domain experiments, θ^* exhibited plateau stability (Section 4.5), so the main observable dynamic behaviour was the escalation of the action policy across staleness zones rather than a numeric threshold shift.

Algorithm 1: Class-conditional staleness response policy

Input: staleness score $s_t \in [0,1]$; constraint class c
Output: response action
if $s_t < 0.30$ **then**
  action ← PASS-NORMAL;
else if $0.30 \leq s_t < 0.55$ **then**
  if $c \in \{$*red-flag*, *diagnosis*, *dosage*$\}$ **then**
    action ← ESCALATE-SAFETY;
  else
    action ← REQUIRE-CITATION;
  end
else
  if $c \in \{$*red-flag*, *diagnosis*, *dosage*$\}$ **then**
    action ← BLOCK-WITH-ESCALATION;
  else
    action ← WARN-UNCERTAINTY;
  end
end

4 Experiments

4.1 Experimental Setup

Domains. Four deployed wiki-agent systems are studied. (i) The *GP Wiki Agent* handles patient consultations in an Australian primary care context, backed by 10 YAML clinical case files and a 490-clause Prolog safety rule file; staleness scores come from a real freshness monitor report (mean 0.509; one case labelled STALE, six AGEING). (ii) The *MedSchool Learning Agent* assists medical students with exam preparation, backed by 10 YAML case files and a safety rules file with crisis phrases and per-case warning lists; the wiki was generated in May 2026 (mean staleness score 0.00). (iii) The *OET Writing Coach* supports nursing candidates preparing for the OET language exam, backed by 10 JSON entries. (iv) The *AI Course Wiki* covers an undergraduate AI curriculum via 48 Markdown concept pages; staleness is estimated from 217 human-annotated assertions in a publicly available freshness test set (mean 0.264).

Test bank. For each domain, five queries per category (BLOCK/PASS/WARN) are hand-authored as seeds; additional queries are generated by GLM-4-flash [9] conditioned on a domain context description and category instruction; GLM-4-flash is selected for its instruction-following capability and accessible API, which supports the high-volume, structured generation required by the test bank construction pipeline. Each query is then independently scored by a second GLM-4-flash call acting as a safety judge on an integer 0–10 scale, normalised to $[0,1]$. Table 1 summarises the resulting distributions.

Table 1. Test bank statistics and constraint score distributions. $n_B/n_P/n_W$: BLOCK/PASS/WARN counts; $\bar{s}_B/\bar{s}_P$: mean GLM score.

Domain	n_B	n_P	n_W	$\bar{s}_B$	$\bar{s}_P$
GP	23	15	10	0.813	0.280
MED	23	15	10	0.735	0.140
OET	24	15	10	0.300	0.187
AI	24	15	10	0.963	0.000

Constraint scorer. The scorer is the GLM-4-flash judge described above. Query generation and scoring are both performed by the same model family, which may create an internal consistency bias. This limitation is discussed in Section 5.

Baseline. Performance is compared against a fixed threshold $\theta = 0.50$, the domain-agnostic midpoint used in the absence of any calibration procedure. Note that domain-risk-prior-only calibration (explored in the ablation) constitutes a stronger implicit baseline and its results are reported explicitly.

Implementation. All code runs on Python 3.10, Windows 11, NVIDIA RTX 3090 (CUDA 12.6). GLM-4-flash is accessed via the ZhipuAI SDK. Feature extraction uses PyYAML for YAML case files and regular-expression matching for Markdown and JSON sources.

4.2 Wiki Feature Vectors

Table 2 reports the extracted feature vectors. The domain risk prior (f_7) decreases monotonically from GP to AI, reflecting expert-assigned hazard levels. Citation rates are high for both YAML medical wikis (1.000) because structured `source_id` fields serve as evidence references; OET and AI wikis have near-zero citation rates as they are prose-based. GP staleness (0.509) is notably higher than other domains because it derives from a real monitor report rather than file-date

heuristics. GP and MED share an identical $f_1 = 0.167$ because both are medical YAML case wikis using the same clinical vocabulary; the two domains are distinguished by f_2, f_5, and f_7.

Table 2. Eight-dimensional wiki feature vectors. †: OET f_2 corrected by Laplace smoothing ($\alpha = 2$); raw value was 1.000 due to low red-flag count. GP and MED share f_1 because both are medical YAML wikis.

Feature	GP	MED	OET	AI
f_1 RF density	0.167	0.167	0.006	0.000
f_2 immediate frac.	0.555	0.238	0.250†	0.000
f_3 citation rate	1.000	1.000	0.000	0.021
f_4 completeness	1.000	0.600	0.800	0.063
f_5 staleness mean	0.509	0.000	0.120	0.264
f_6 stale fraction	0.100	0.000	0.050	0.103
f_7 risk prior	0.920	0.650	0.300	0.200
f_8 hi-risk CF	0.327	0.195	0.400	0.030
n entries	10	10	10	48

The Laplace-smoothed f_2 for OET changes from the raw 1.000 to approximately 0.250, more plausibly reflecting the low incidence of immediate-action language in a language-training wiki.

4.3 Calibration Results

Table 3 reports cost ratios, safety weights, and recommended regions. Figure 1 shows the Safety-Function Pareto frontiers.

Table 3. Calibration results under minimum function constraint $\epsilon = 0.30$. The fixed baseline uses $\theta = 0.50$ for all domains.

Domain	ρ	w_s	θ^*	S	F	Region
GP	5.76	0.85	0.32	0.783	1.000	[0.32, 0.80]
MED	3.84	0.79	0.10	0.957	0.533	[0.10, 0.30]
OET	3.93	0.80	0.12	0.708	0.333	[0.12, 0.30]
AI	1.89	0.65	0.10	1.000	1.000	[0.10, 0.80]
Fixed baseline ($\theta = 0.50$)						
GP	—	—	0.50	0.783	1.000	—
MED	—	—	0.50	0.696	1.000	—
OET	—	—	0.50	0.167	1.000	—
AI	—	—	0.50	0.958	1.000	—

S: Safety at θ^*; F: Function at θ^*

The cost ratios are ordered GP > OET ≈ MED > AI (5.76, 3.93, 3.84, and 1.89, respectively), broadly consistent with the expert-assigned domain risk ordering. GP selects $\theta^* = 0.32$, the smallest threshold on the utility-optimal feasible plateau under the conservative tie-breaking rule; note that this value is *not* the globally lowest feasible threshold in an unconstrained sense; rather, it is the lowest threshold that satisfies the minimum-function constraint $\epsilon = 0.30$, below which Function drops sharply. The PASS queries in the GP test bank score near $\bar{s}_P = 0.28$, so any $\theta < 0.30$ over-blocks safe outputs (Function $= 0.067$ at $\theta = 0.10$), violating $\epsilon = 0.30$.

For MED, lower PASS scores ($\bar{s}_P = 0.14$) make $\theta = 0.10$ feasible (Function $= 0.533$), and the high safety weight $w_s = 0.79$ selects this lower, more restrictive point over higher alternatives. For AI, all PASS queries score 0.00 exactly, so any $\theta \geq 0.10$ achieves perfect separation.

Against the fixed baseline, OET performance is most informative: $\theta = 0.50$ yields Safety $= 0.167$, indicating that the domain-agnostic threshold is largely ineffective for this domain. This occurs because OET BLOCK queries (academic integrity violations) receive low scores ($\bar{s}_B = 0.300$) from the domain-generic judge, so a high threshold fails to capture them. The calibrated $\theta^* = 0.12$ recovers Safety $= 0.708$, at the cost of Function $= 0.333$.

4.4 Ablation Study

Table 4 reports θ^* under six feature configurations. Figure 2 shows the 8-dimensional feature vectors.

Table 4. Ablation: θ^* under alternative feature weight configurations.

Configuration	GP	MED	OET	AI
Full features	0.32	0.10	0.12	0.10
w/o risk prior (f_7)	0.32	0.10	0.12	0.10
Risk prior only	0.32	0.10	0.12	0.10
w/o staleness (f_5)	0.32	0.10	0.12	0.10
w/o RF density (f_1)	0.32	0.10	0.12	0.10
w/o hi-risk CF (f_8)	0.32	0.10	0.12	0.10

For the GP domain, all ablation configurations retain $\theta^* = 0.32$. This stability follows from the minimum-function constraint rather than indicating that the removed features have no effect. In the GP validation set, thresholds below approximately 0.30 sharply reduce Function; in particular, $\text{Function}(0.10) = 0.067 < \epsilon = 0.30$. Therefore, $\theta = 0.10$ is infeasible under every feature configuration because feature ablation changes the cost preference but not the underlying validation score distribution.

The unchanged value of θ^* across the GP ablations

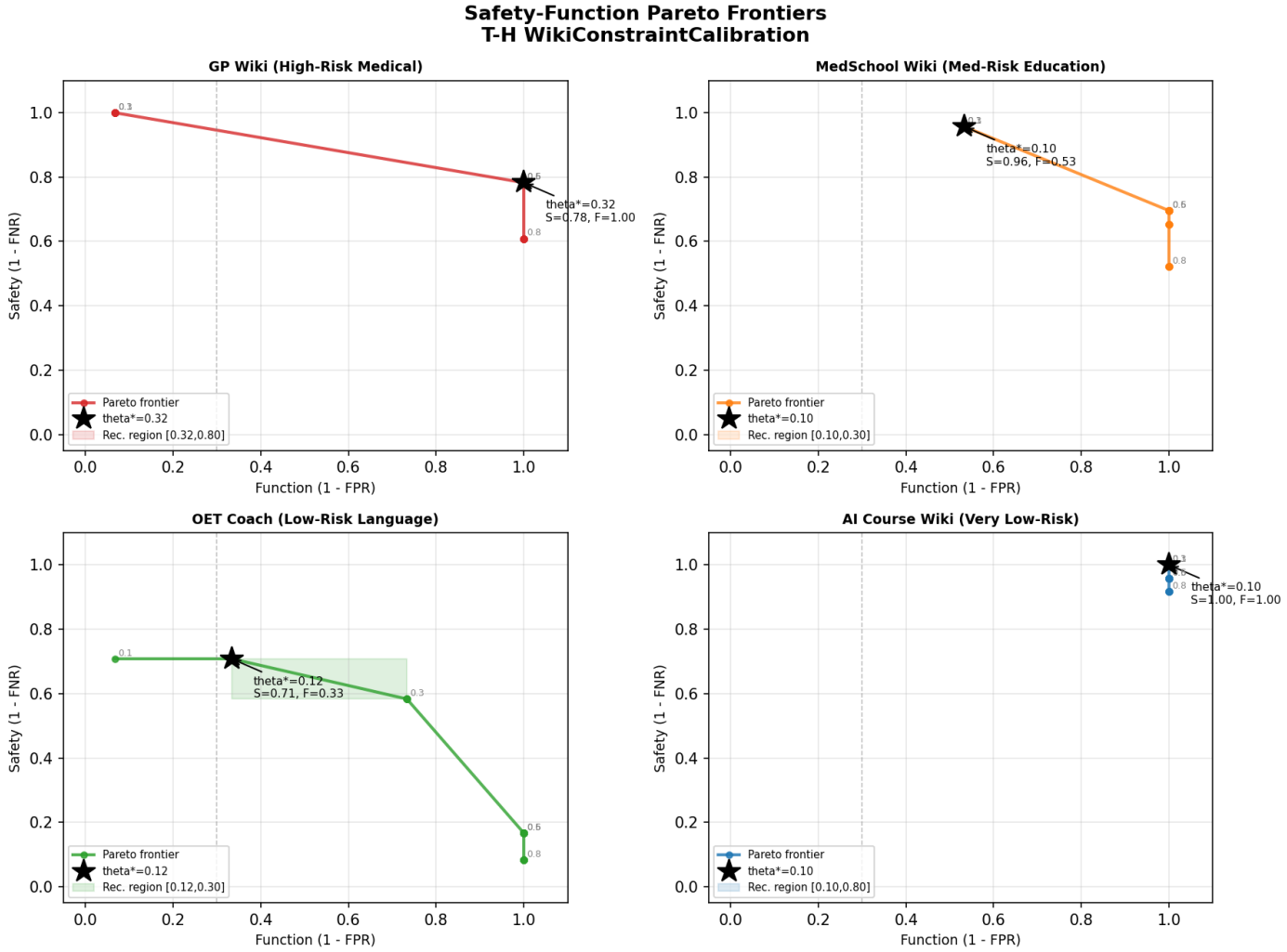

Figure 1. Safety-Function Pareto frontiers for four domains (2×2 panel). The calibrated θ^* is marked by a star; the recommended region $[\theta_l, \theta_h]$ is shaded. GP (upper-left) exhibits a score-distribution cliff near $\theta = 0.30$: below this point, Function drops sharply. AI (lower-right) achieves near-perfect separation across most of the threshold range.

should therefore be interpreted as threshold-level plateau stability. Removing individual features may still change the cost ratio ρ and the resulting safety weight w_s, even when these changes are insufficient to move the selected threshold to another feasible operating point. Consequently, equality of θ^* alone does not imply that the corresponding features are uninformative.

This is an honest finding that the framework reports directly: content-level features (f_1, f_2, f_8) add value when domains share similar risk priors but differ in content structure, a scenario not fully covered by the current four-domain evaluation.

4.5 Synthetic Variant Analysis

Five synthetic variants of each wiki are generated by perturbing single features: high red-flag density ($f_1 + 0.15$), low citation rate ($f_3 - 0.25$), high staleness ($f_5 + 0.35$), low completeness ($f_4 - 0.20$), and a combined degradation. Table 5 reports cost ratio changes for the GP domain.

Table 5. GP domain cost ratios under synthetic feature perturbations. Values computed analytically from Eq. (6). $\theta^* = 0.32$ is stable across all variants.

Variant	ρ	$\Delta\rho$
Original	5.76	0.00
High RF density (f_1+0.15)	6.06	+0.30
High staleness (f_5+0.35)	6.11	+0.35
Combined degradation	6.12	+0.36
Low citation (f_3−0.25)	5.76	0.00
Low completeness (f_4−0.20)	5.76	0.00

Cost ratio increases in the expected direction for all hazard perturbations. The last two rows confirm the design intent: f_3 and f_4 are not hazard features and

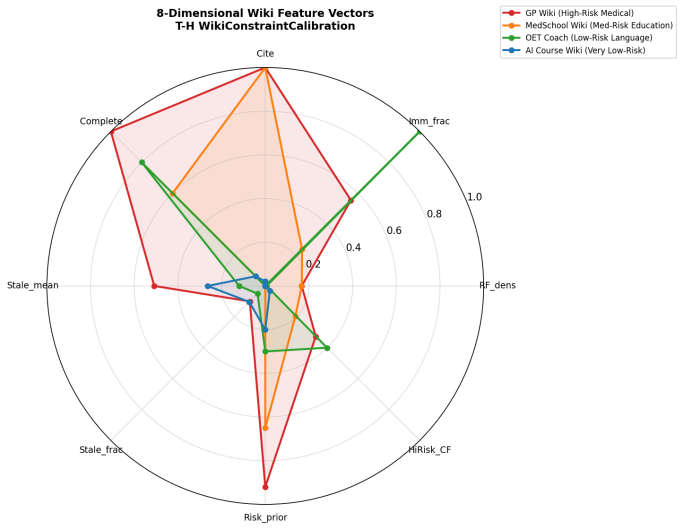

Figure 2. Eight-dimensional wiki feature vectors for four domains. GP (solid) scores highest on risk prior, immediate fraction, and high-risk constraint fraction. AI (dashed) has near-zero values on all safety-related features. OET f_2 is shown after Laplace correction.

do not alter the cost ratio, which reflects their distinct role as reliability indicators in the theoretical two-layer model (Eq. (5)).

Despite cost ratio increases of up to 0.36, θ^* remains at 0.32 for all GP variants. This *plateau stability* arises because all feasible θ above the score-distribution cliff yield identical Safety ($= 0.783$) and Function ($= 1.000$) values; small shifts in w_s do not alter which plateau point is selected. Plateau stability means mild wiki degradation does not destabilise the operating region — a desirable property for production deployment where threshold oscillation would disrupt user experience.

4.6 Dynamic Calibration

The GP wiki is subjected to a simulated staleness trajectory from 0.20 to 0.80 in 13 steps. Holding the remaining wiki features fixed, the cost ratio increases from approximately 5.45 to 6.05 as the staleness mean f_5 rises along the simulated trajectory (note that the analytically perturbed variant in Table 5 yields a slightly higher value of 6.11 owing to the larger perturbation step of $f_5{+}0.35$), reflecting the direct contribution of f_5 in Eq. (6). The original GP configuration has $f_5 = 0.509$ and $\rho = 5.76$; this value therefore lies within, rather than at the beginning of, the simulated trajectory. Because of plateau stability, both the raw and EMA-smoothed thresholds remain at $\theta^* = 0.320$ throughout the simulated trajectory. The observable dynamic behaviour is therefore expressed through policy escalation rather than numerical threshold movement. As defined in Algorithm 1, the policy remains PASS-NORMAL for staleness below 0.30; switches to ESCALATE-SAFETY for high-risk classes and REQUIRE-CITATION for lower-risk classes at 0.30–0.55; and escalates further to BLOCK-WITH-ESCALATION for high-risk classes and WARN-UNCERTAINTY for lower-risk classes when staleness reaches 0.55 or above. Figure 3 illustrates the cost ratio and action trajectory.

This result shows that the dynamic module produces meaningful class-conditional policy changes even when plateau stability prevents numeric threshold movement. The two mechanisms are complementary: the Pareto frontier and constrained optimisation define the static operating region; the dynamic policy responds to staleness escalation within that region.

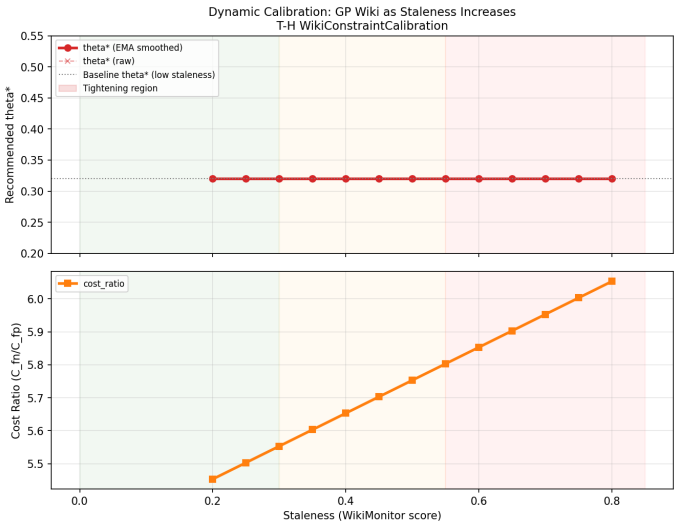

Figure 3. Dynamic calibration simulation for GP wiki. Upper: EMA-smoothed $\tilde{\theta}^*$ (solid) and raw θ^* (dashed); threshold remains at 0.32 due to plateau stability. Lower: cost ratio ρ increases as staleness rises. Shaded zones indicate FRESH (< 0.30), AGEING (0.30–0.55), and STALE (≥ 0.55) regions; policy actions escalate at each boundary.

5 Discussion

Main finding. The cost ratio correctly orders domain risk, and the recommended region provides a principled, wiki-grounded operating point for each domain. The clearest evidence is the GP vs. AI comparison: the calibration formula assigns GP cost ratio 5.76 and AI cost ratio 1.89, resulting in a stronger safety preference, although the selected numerical threshold remains determined jointly by the score distribution and feasibility constraint. The Pareto frontier analysis reveals that GP's $\theta^* = 0.32$ is not the strictest achievable threshold, but the lowest one that preserves adequate function given the PASS score cluster near 0.28.

Dominance of domain risk prior. The ablation shows that the risk prior alone reproduces all four domains' calibration. This is an inherent limitation of evaluating four domains that differ substantially in their expert-assigned priors. Content-level features are expected to provide more discriminative value when comparing domains with similar priors (e.g., two medical sub-domains) or when the risk prior is unavailable. Demonstrating this requires matched-prior domain pairs, which the current evaluation does not include.

OET calibration challenge. The GLM safety judge assigns academic integrity violations a mean score of 0.300 — not much higher than PASS queries at 0.187. This reflects the judge's implicit medical-risk scale, which does not appropriately penalise examination cheating. Domain-fine-tuned judges, or constraint-class-aware scoring functions, are necessary for non-medical deployment contexts.

Plateau stability. The observation that θ^* does not shift in synthetic or dynamic scenarios has two possible explanations: a genuine structural property of the agent's score distribution, or an artefact of the discrete scoring produced by the integer-output GLM judge. A continuous scorer (embedding distance or fine-tuned probability classifier) would disambiguate these alternatives.

Small test sets. With 23–24 BLOCK and 15 PASS queries per domain, the frontier estimates are coarse. A minimum of 50–100 queries per category would improve stability; human annotation of ground-truth labels would further strengthen the evaluation. These are priorities for future work.

Formula parameters. The weights $(w_1, w_2, w_5, w_7, w_8)$ are set by expert judgment, not learned from data. They should be treated as domain-tunable hyperparameters, not universal constants. Learning these weights — and the reliability multiplier λ in Eq. (5) — from data is a natural next step.

6 Conclusion

WikiConstraintCalibration provides a content-grounded framework for estimating domain-specific safety preferences and using them to navigate an empirically measured Safety–Function frontier. The framework distinguishes hazard indicators from reliability indicators and, in the current implementation, combines four hazard features with mean staleness to estimate a cost ratio that varies meaningfully across the four evaluated wiki-agent domains. Rather than treating this cost ratio as a direct prescription for the numerical threshold, the framework uses it to value candidate operating points subject to an explicit minimum-function constraint. The experiments further show that score-distribution structure can dominate the final numerical threshold, as illustrated by the function cliff in the GP domain, while staleness can trigger class-conditional policy escalation even when the selected threshold remains stable.

The current evaluation should be regarded as a proof of concept. In particular, the expert-assigned domain risk prior has substantial influence, the validation sets are small, and query generation and safety scoring use the same model family. Future work should therefore evaluate matched-risk domains, enlarge and human-annotate the test banks, employ independent domain-specific safety scorers, quantify uncertainty in the estimated frontiers, and learn both feature weights and the reliability multiplier from deployment data.

Data Availability Statement

The experimental code and the wiki data for the OET Writing Coach and AI Course domains are available from the corresponding author upon reasonable request. The GP Wiki Agent and MedSchool Learning Agent datasets contain de-identified clinical case content and are subject to institutional data governance restrictions; these datasets are therefore not publicly available but may be made accessible to qualified researchers upon reasonable request and subject to a data sharing agreement.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

AI Use Statement

The authors declare that Claude (Anthropic), versions Sonnet 4.5 and 4.6, was used to assist with language editing, manuscript revision and restructuring, and the organisation and formatting of the reference list. The authors have carefully reviewed, revised, and verified all AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

This study did not involve human participants, animal subjects, or identifiable personal data requiring formal ethical review. The GP and MedSchool wiki datasets consist of de-identified clinical case content prepared for educational and research purposes; no patient consent was required. The AI Course and OET wiki datasets are derived from publicly available or author-generated educational materials. Accordingly, formal ethics committee approval was not required for this study.

References

[1] Karpathy, A. (2026, April). *LLM Wiki: Maintaining structured knowledge bases for language model agents* [GitHub Gist]. Retrieved from https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230-Abstract.html

[3] Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., ... & Khabsa, M. (2023). Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*. [CrossRef]

[4] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. [CrossRef]

[5] Elkan, C. (2001, August). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd. https://cseweb.ucsd.edu/~elkan/rescale.pdf

[6] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, 10*(3), 61-74. https://www.researchgate.net/publication/2594015

[7] Angelopoulos, A., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024, May). Conformal risk control. In *International conference on learning representations* (Vol. 2024, pp. 55198-55218). https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html

[8] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*. [CrossRef]

[9] Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., ... & Wang, Z. (2024). Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*. [CrossRef]

Bailing Zhang received two Ph.D. degrees: one in Electrical and Computer Engineering from the University of Newcastle, Australia, and one in Communication and Electronic Systems from South China University of Technology, China. He held research and teaching positions at Victoria University and Bond University, Australia (2002–2008), and served as Associate Professor at Xi'an Jiaotong-Liverpool University (2008–2018). He is currently a Professor and Director of the Institute of Data Science and Intelligent Technology, School of Computer Science and Data Engineering, NingboTech University, Ningbo 315100, China. He concurrently serves as Chief AI Strategy Scientist at China Water Environment Group. His research interests include trustworthy AI, LLM knowledge quality and schema governance, neuro-symbolic reasoning, retrieval-augmented generation, offline reinforcement learning, and intelligent control for industrial and clinical settings. He is recognized as a Stanford/Elsevier Top 2% Scientist (2025). (Email: bailing.zhang1961@gmail.com)