Closing the Welfare Outreach Gap: A Conversational Architecture and Cell-Level Eligibility Benchmark for Korean Welfare Recommendation

Anonymous Authors[1]

Abstract

Welfare benefits often fail to reach the people who need them most, including elderly citizens, digitally underserved users, and those unfamiliar with eligibility categories, because portal-based search requires users to know both *which programs to search for* and *how their personal attributes map to eligibility criteria*. This paper addresses the resulting outreach gap in three steps. (1) We propose a *conversational welfare-policy recommendation architecture* that elicits user attributes through natural-language dialogue. (2) To compare its core eligibility-matching component quantitatively, we construct *KWelfareBench*, a cell-level eligibility ground-truth table over 4,937 Korean welfare policies and 180 synthetic personas. (3) Using this benchmark we compare several recommendation architectures and identify an eligibility-matching structure suited to the Korean welfare domain. The resources, code, and personas are released as a shared foundation for subsequent Korean welfare outreach research.

1. Introduction

The Korean welfare access problem. Korean welfare operates on an opt-in basis: a citizen must already know that a relevant program exists, judge whether they qualify, and submit an application before any benefit is delivered. The Korean Ministry of Health and Welfare administers more than 4,937 programs across central government, 17 `sido`, and 226 `sigungu`, and citizens are expected to query a keyword portal under the assumption that they already know the program name, the responsible agency, and the eligibility vocabulary. This excludes precisely the populations whom welfare is intended to support: digitally underserved older adults, low-income households, and citizens combining several vulnerability axes.

Why current LLM/RAG systems are insufficient. Off-the-shelf RAG inherits four interacting weaknesses for welfare: quantitative evaluation is underdeveloped (Reddy & Bell, 2025; Alam et al., 2024); generic retrievers ignore eligibility (a top-ranked policy can be formally inapplicable); query pools leak persona attributes into the surface form, confounding eligibility with intent; and single-shot retrieval mismatches users who do not state every relevant attribute up front.

Approach. We address these weaknesses through measurement. We separate eligibility from intent into two independently labeled ground-truth tables (GT-1, GT-2) and study their intersection (GT-3), the metric a deployed assistant has to satisfy. The eligibility table is built from a deterministic rule over a 69-binary + 3-numeric tag schema unified across policies and personas; the intent table is graded $\{0, 1, 2\}$ by two independently prompted LLM judges with cross-LLM agreement reported. A five-stage conversational loop closes the back-end gap. The five stages are NL input (S1), tag extraction (S2), query rewrite (S3), eligibility-aware retrieval (S4), and follow-up question selection (S5); the loop is introduced fully in §5, and only the recommendation and follow-up stages (S4, S5) are evaluated end-to-end here, under oracle simulation of the upstream NL stages (S1–S3).

Contributions. **(C1) KWELFAREBENCH, a cell-level evaluation resource:** nine release files covering 4,937 policies, 180 synthetic personas, a 69-binary tag schema, a 66-query persona-orthogonal pool, and three ground-truth tables (GT-1 with 888,660 eligibility cells, GT-2 with 325,842 graded intent cells, GT-3 the intersection in six region/grading variants). **(C2) Persona-orthogonal query taxonomy:** 33 sub-topics, two phrasings each, curated under five rules R1–R5 so the query surface is linearly independent of persona attributes; a multi-prefix counterfactual rises from 8.7% to 11.4% non-zero topical labels when persona prefixes are re-attached. **(C3) Architecture comparison**

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and conversational evaluation: six retrievers (B1–B6) on GT-2 and GT-3 (B2 ko-SBERT NDCG@10 0.86 on GT-2; B6 Rule + Dense 0.706 on GT-3-strict; §4), and a conversational back-end where a RandomForest selector matches exhaustive information-gain within 4% recall at ∼1 ms per turn (§5). Reliability is validated by a two-LLM full pass on A3 (κ=0.822) and A8 (κ_w=0.69) with human adjudication of all disagreement cells (§3.7). All artifacts are released under KOGL Type 1.

2. Related Work

LLM-based recommenders and graded relevance. Modern recommender pipelines combine sparse retrievers (Robertson & Zaragoza, 2009), dense retrievers (Karpukhin et al., 2020; Lewis et al., 2020), and increasingly LLM rerankers or end-to-end generative recommenders (Gao et al., 2024). Evaluation typically follows the IR tradition of graded relevance (Thakur et al., 2021), expressing *topical* relevance to a query rather than whether a candidate is *actionable* for a specific user. In welfare recommendation the user is not implicit: a topically relevant policy for which the user is statutorily ineligible is a matching failure, yet existing benchmarks lack the user-conditioned ground truth needed to measure this.

Public-sector and welfare AI. Public-sector welfare recommendation is a recently emerging area: in 2024–2025 government-scheme matching tools (Reddy & Bell, 2025), small-government LLM applications (Huggins-Daines, 2024), homelessness and social-security benefits chatbots (Nelson et al., 2024; US Department of Health and Human Services, 2024), and trust-and-ethics studies (Kaun & Männiste, 2025) have appeared. Together they demonstrate feasibility but typically evaluate top-K accuracy on a small demonstration persona set, do not release a reusable ground-truth table, and are not directly comparable across architectures. Per-segment fairness diagnostics follow the intersectional framework of Wang et al. (2024).

Korean policy retrieval and LLM-as-judge. Empirical work on Korean welfare retrieval is minimal: the *Bokjiro* portal exposes only a keyword interface, Reddy & Bell (2025) prototypes RAG search but evaluates on small persona sets and does not release ground truth, and Ban et al. (2025) focuses on a single sub-population. The closest non-welfare analogue is TrialGPT (Jin et al., 2024) for clinical-trial eligibility, whose cell-level patient–trial structure is structurally identical to a persona–policy table but assumes users already know their attributes. Recent personalized LLM-assistant benchmarks (Zhao et al., 2025) share the cell-level philosophy. Constraint-based recommendation (Jannach et al., 2015; Le et al., 2023) provides a complementary lens, hard rules over user attributes, which we instantiate as our rule eligibility function. LLM-as-judge labeling is now common (Zheng et al., 2023; Han et al., 2025); we cross-validate two independently prompted judges and report quadratic-weighted κ (Cohen, 1968) appropriate for ordinal $\{0, 1, 2\}$ labels. Chi et al. (2024) learns a retrieval-aware clarification policy end-to-end; our selector is a simpler RandomForest whose advantage is sub-millisecond inference.

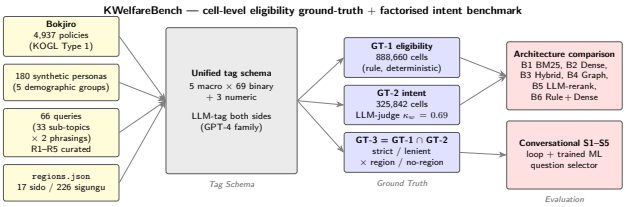

Figure 1. KWELFAREBENCH pipeline overview. Raw policies (A1) and the regional standard (A2) feed a unified tag schema (A3, A5) shared with the synthetic persona corpus (A4); the persona-orthogonal query pool (A6) and the three ground-truth tables (A7 GT-1 eligibility, A8 GT-2 intent, A9 GT-3 intersection) drive the architecture comparison (§4) and the conversational evaluation (§5).

ID	Artifact	Format	Size
A1	Raw policies	JSON	4,937 policies
A2	Regional standard	JSON	17 sido / 226 sigungu
A3	Policy tags	JSON	4,937 × 72
A4	Synthetic personas	JSON	180 personas
A5	Persona tags	JSON	180 × 72
A6	Query pool	JSON	33 topics, 66 queries
A7	GT-1 eligibility	NPZ	888,660 cells
A8	GT-2 intent (graded)	NPZ	325,842 cells
A9	GT-3 intersection	NPZ	11,880×3

Table 1. The nine KWELFAREBENCH release artifacts. All files are released under KOGL Type 1, the same licence as the Bokjiro source. A9 ships in three region variants (×3: strict, sido-only lenient, region-agnostic) crossed with two grading thresholds, yielding the six derived GT-3 matrices.

3. KWELFAREBENCH

3.1. Overview: Released Resource Catalogue

KWELFAREBENCH ships as nine release artifacts (A1–A9, Table 1) that fall into three groups: (i) raw policy data and regional standard labels (A1–A2, §3.2–§3.2); (ii) the unified tag schema and the persona corpus, with both sides labeled in the same schema (A3–A5, §3.3–§3.4); (iii) the persona-orthogonal query pool (A6, §3.5) and the three ground-truth tables with six derived matrices (A7–A9, §3.6). Reliability of the LLM labels underlying A3 and A8 is reported in §3.7. The pipeline is summarised in Figure 1.

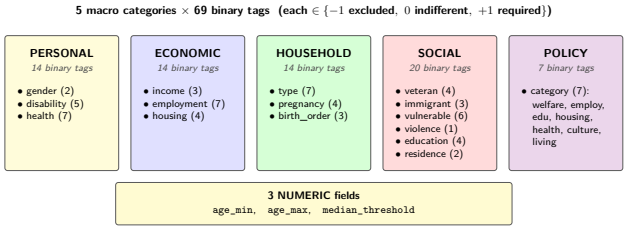

Figure 2. The KWELFAREBENCH tag schema: 5 macro-categories × 69 binary tags + 3 numeric fields. PERSONAL, ECONOMIC, HOUSEHOLD, and SOCIAL describe *eligibility* (shared between policies and personas); POLICY describes *program type* and inherits the seven Bokjiro categories.

3.2. A1–A2: Raw Snapshot and Regional Standard

A1. Policy metadata from *Bokjiro* (`https://www.bokjiro.go.kr`), the Korean Ministry of Health and Welfare's portal. As of an April 2026 snapshot, the collection contains 4,937 programs (name, summary, description, eligibility text, benefits, application procedure, region, responsible agency). The Bokjiro footer (accessed 2026-04) states "Gonggongnuri Type 1 (KOGL)"—attribution required, commercial use allowed, modification allowed; all A1–A9 inherit this licence. **A2.** A frozen list of 17 `sido` and 226 `sigungu`, released alongside the tag schema. The eligibility function in §3.6 consumes these standard codes; free-text neighborhood preferences live in a separate `region.nl` channel and surface only at the GT-2 retrieval layer.

3.3. A3: Policy Tag Labels (69 binary + 3 numeric)

We design a hierarchical schema of five macro-categories (PERSONAL, ECONOMIC, HOUSEHOLD, SOCIAL, POLICY) over 69 binary tags (Figure 2; full list in Appendix A). The first four follow the standard four-axis eligibility taxonomy used in Korean social-welfare textbooks; POLICY (welfare, employment, education, housing, health, culture, living) takes the seven category labels used in Bokjiro's policy-classification metadata field, which is distinct from the fifteen search-filter sub-topics exposed in Bokjiro's public UI. We verified that all 4,937 policies in our snapshot carry one of these seven labels. Each binary tag takes one of three values: $+1$ *required*, -1 *excluded* (disqualifies the applicant), or 0 *indifferent*. The 3-state representation distinguishes "not mentioned" from "explicitly excluded", which noticeably affects matching precision. Three numeric fields (`age_min`, `age_max`, `median_threshold`) complete the schema.

Construction. We call `gpt-5.4` (OpenAI chat-completions API; April 2026; temperature 0) in parallel: each prompt includes the policy's name, summary, eligibility text, benefits, application procedure, and region, and the model emits the $69 + 3 = 72$ values on a single comma-separated line rather than as JSON. The comma-separated form reduces output tokens by roughly $5.5\times$ (per-policy cost and total runtime details in Appendix D). Reliability is reported in §3.7.

3.4. A4–A5: Personas and Persona Tag Labels

The 180 personas (A4) are organized as **Base (50)**, ten personas in each of five reference groups (disability, single parent, senior, youth, general); **Extra (80)**, regional coverage spanning 17 provinces and 226 districts; and **Intersectional (50)**, 18 intersectional axes with 1–6 personas each. The five reference groups span the at-risk groupings most frequently cited in MOHW yearbooks and represented by Bokjiro's recommendation filter; the alignment is not one-to-one (Bokjiro's filter additionally distinguishes multi-cultural and multi-child families, which we fold into intersectional axes rather than top-level groups). Each persona is synthesized with stratified sampling (gender 50/50, group-specific age, 17-sido weighted distribution); the seed and the generator are released, and all personas are synthetic with no PII.

A5. Each persona's attributes are encoded as `tag_values` that mirror the policy-side 69-binary + 3-numeric schema, so the eligibility function runs on a single shared vocabulary across both sides of the cell-level table.

Sample-size design. We chose n=180 (5 groups × 36) to keep ≥ 30 per group for stratified bootstrap CIs while keeping the LLM tagging budget feasible; the L5 limitation discusses scale-up.

3.5. A6: Persona-Orthogonal Query Pool

To evaluate retrieval that respects *user intent* (GT-2)—i.e., topical relevance between a free-form query and a candidate—we construct a pool of **66 queries** drawn from **33 sub-topics**, paired with each 180 persona at test time. The central design requirement is *persona-orthogonality*: the sub-topic taxonomy is linearly independent of the 9-dimensional persona attribute space used by the eligibility function. If a sub-topic embeds a token like "youth" or "low-income", GT-1 and GT-2 confound and retrieval failures cannot be attributed to mismatched eligibility vs. intent.

Stage 1. We adopt the seven Bokjiro macro-categories (welfare, employment, education, housing, health, culture, living) as the top of the taxonomy. From each we randomly sample 30 policies and prompt `gpt-5.4-mini` (temperature 0) to extract candidate sub-topics under the *interest unit* criterion; 98 raw candidates result. **Stage 2 (curation, five rules).** Two authors independently apply: **(R1) Persona-orthogonality**, reject sub-topics whose surface form mirrors a persona attribute ("youth rent assis-

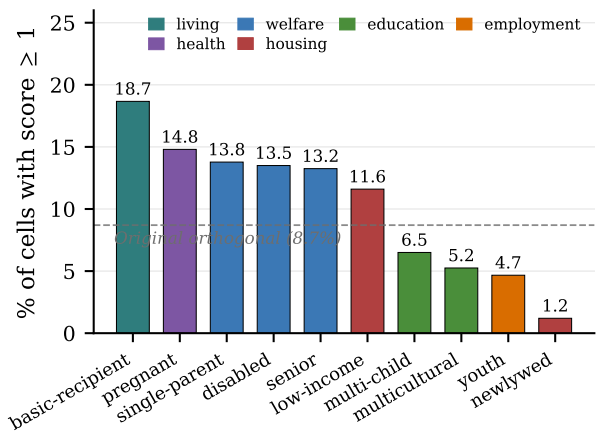

Figure 3. R1-violation multi-prefix ablation: % of cells scoring ≥ 1 when the persona-orthogonal query is prefixed by 10 demographic identifiers. Pooled mean 11.4% vs. orthogonal baseline 8.7% (+2.6pp). The variance across prefixes (1.2% newlywed → 18.7% basic-recipient) shows R1 violations have heterogeneous semantic spillover.

tance" → "rent assistance"); **(R2) Topical consolidation**, merge same-concern sub-topics ({school-uniform, childcare, loans, scholarships} → *tuition and education-cost support*); **(R3) Coverage granularity**, each sub-topic must match ≥ 30 policies (mean > 100); **(R4) Cross-category deduplication**, one best-fit category each; and **(R5) No specific values** (no region names, exact ages, income amounts). The 98 raw sub-topics reduce to 33 (welfare 5, employment 4, education 4, housing 5, health 5, culture 4, living 6). **Stage 3 (phrasing).** Two query variants per sub-topic, a *procedural* ("how do I apply for X support?") and a *situational* ("X is becoming financially hard for me…"), yielding $33 \times 2 = 66$ queries, each 10–25 Korean characters.

R1-violation ablation. To check that R1 actually matters, we re-prepend ten representative demographic prefixes (senior, disabled, single-parent, youth, multi-child, multicultural, newlywed, low-income, pregnant, basic-recipient) to the sub-topics and re-grade a stratified 4,950-cell sample with the independent judge `gemini-2.5-flash-lite`.[1] Pooled across the ten prefixes, the fraction of non-zero (graded ≥ 1) labels rises from 8.7% to 11.4% (Δ=+2.6 pp); 8 of 10 prefixes show positive shifts, full distribution in Figure 3. The direction supports R1: re-attaching demographics inflates topical relevance, which is exactly the leak factoring eligibility from intent is meant to prevent.

[1]We use an LLM judge rather than a semantic-similarity classifier because the question is whether a re-prefixed query *changes the recommendation-relevant intent* of the cell, which requires welfare-policy domain knowledge a similarity classifier lacks; on this subset the OpenAI judge agrees at κ_w=0.71.

Persona group	n	Mean eligible policies
Disability	26	**122.2**
Senior	28	108.4
Intersectional	50	103.5
Single parent	24	89.2
Youth	26	78.2
General	26	73.0
Total	**180**	**97.0**

Table 2. KWELFAREBENCH mean eligible-policy count by persona group.

3.6. A7–A9: Three Ground-Truth Tables

GT-1 (A7) rule-based eligibility. A deterministic function $\text{eligible}(\pi, p) \in \{0, 1\}$ requires *all* nine dimensions (logical AND): income, special, age, gender, employment, household, education, sigungu, sido. Separating the dimensions allows attribution of failures during diagnostic analysis. Region matching uses only A2 codes; free-text `region.nl` preferences are deferred to GT-2. A citizen is rule-eligible for $\approx 1.96\%$ of all programs (mean 97.0 per persona; group breakdown in Table 2). **GT-2 (A8) graded intent.** For each (query, policy) pair we elicit a 3-point label (0 unrelated, 1 adjacent, 2 direct) from `gpt-5.4-nano` (primary) and `gemini-2.5-flash-lite` (cross-vendor), each prompted independently with a shared few-shot anchor set, over $66 \times 4{,}937 = 325{,}842$ cells per judge. **GT-3 (A9) intersection.** A policy is in GT-3 when both GT-1-eligible *and* GT-2-relevant. We release six matrices: {strict ($= 2$), lenient (≥ 1)} grading $\times$ {strict, lenient, no-region} region. **GT-3-strict** (mean $\bar{n}_{\text{el}}$=4 targets per pair, 20.86% empty cells) is the operating point a deployed assistant has to clear.

Group-level eligibility counts. Vulnerable groups are eligible for more policies than the general population (Table 2: disability 122 vs. general 73, +67%), reflecting the program-side volume targeted at these groups—distinct from the *application rate* behind the opt-in outreach gap.

3.7. Reliability: Cross-LLM and Human Audit

Every LLM-judged release artifact (A3, A8) follows the same two-LLM-full-pass plus human-on-disagreement protocol. Reviewers have welfare-policy backgrounds (graduate-student researchers and domain practitioners; affiliations omitted for review), treated as a strong reliability tier above pure LLM labels but below a credentialed-expert audit (§6, L6).

A3 (policy tags). Both `gpt-5.4` and `gpt-5.5` independently emitted the full $4{,}937 \times 72$ tag matrix at temperature 0. The two LLMs agree on **96.8**% of the $\sim$ 356,000 binary cells (κ=0.822, "almost perfect"). Numeric-field agreement is high (`age_min` 98%, `age_max` 96%,

median_threshold 100%). The LLM-disagreement set (n≈11,400 cells, 3.2%) was split four ways and fully adjudicated by the reviewer team in a 30-minute consensus pass per shard. The released A3 uses human adjudication on the disagreement set and the LLM-majority label on the agreement set.

A8 (query–policy intent). Both `gpt-5.4-nano` and `gemini-2.5-flash-lite` graded all 325,842 cells independently with a shared few-shot anchor set. Cross-LLM agreement is **91.03**% ($\kappa_{\text{unweighted}}$=0.50, κ_w=0.69 (Cohen, 1968)); disagreements concentrate on adjacent grades. The agreement set (n=296,575) was sanity-checked with a 100-cell uniform random sample reviewed independently by four humans, who concurred with the LLM majority on **96/100** cells (the four disagreements were single-step $1 \leftrightarrow 2$). The disagreement set (n=29,267, 8.97%) was split four ways and fully adjudicated. The released A8 uses human adjudication on the disagreement set and the LLM-majority label on the agreement set; `gpt-5.4-nano` is the primary alias for downstream metrics.

4. Architecture Comparison

4.1. Setup

Task and metrics. Given a persona's natural-language query, return the top-K candidate policies among the 4,937 KWELFAREBENCH programs. Evaluation metrics, computed against KWELFAREBENCH ground truth, are Precision@K, Recall@K, and NDCG@10 for $K \in \{5, 10, 20\}$, with K=10 used as the primary cutoff throughout. Means are reported with bootstrap 95% CIs (N_{boot}=10,000, query-level resampling); pairwise comparisons use the paired Wilcoxon signed-rank test (Wilcoxon, 1945) with Holm–Bonferroni correction (Holm, 1979).

Components. BM25Okapi (Robertson & Zaragoza, 2009; Trotman et al., 2014) (k_1=1.5, b=0.75); dense retriever `jhgan/ko-sroberta-multitask` (jhgan, 2021; Reimers & Gurevych, 2019) (768-d Korean SBERT, KorNLI/KorSTS contrastive); RRF (k_{RRF}=60) (Cormack et al., 2009); LLM rerank judges `gpt-5.4-nano` (OpenAI, 2026) and `gemini-2.5-flash-lite` (Google DeepMind, 2025). Policy text concatenates `name‖summary‖ eligibility‖benefits` (mean 247 Korean characters); no chunking. BM25 uses a hybrid Hangul tokenizer (whitespace + 2-gram syllable backoff).

Baselines. **B1 BM25**; **B2 Dense** (ko-SBERT) with multilingual variant **B2′ BGE-M3** (BAAI/bge-m3 single-vector dense); **B3 Hybrid** (RRF of B1 & B2); **B4 Graph** (cosine of persona-satisfied vs. policy-required tag sets, Appendix B.4); **B5 LLM-rerank** (BM25 top-50 reranked by Gemini with structured-output enforcement) with cross-encoder variant **B5′ BGE-reranker-v2-m3** (BAAI multilingual cross-encoder reranking BM25 top-200); **B6 Rule + Dense rerank** (rule prefilter then ko-SBERT rerank). Matched-prefilter controls **B2+rule** and **B5+rule** apply the rule prefilter then rerank by Dense or Gemini; B6 = B2+rule by construction. We omit a rule-only competitor on GT-3: it shares the predicate with the ground truth and would be circular.

Baseline	NDCG@10	P@10$_{\geq 2}$	R@10$_{\geq 2}$
Sparse / lexical			
B1 BM25-Okapi	0.642	0.544	0.096
Dense (single-vector cosine)			
B2 ko-SRoBERTa	**0.858***	**0.752***	**0.142***
B2′ BGE-M3	0.815	0.698	0.126
Cross-encoder rerank (BM25 top-200)			
B5′ BGE-reranker-v2-m3	0.853	0.752	0.139
Hybrid / LLM-rerank			
B3 Hybrid (RRF, B1+B2)	0.849	0.745	0.134
B5 Gemini-rerank (BM25 top-50)	0.789	0.683	0.126

Table 3. GT-2 (intent-only) retrieval over the 66 queries × 4,937 policies. Bootstrap 95% CIs (query-level resampling, N_{boot}=10,000): B2 NDCG@10 $[0.81, 0.90]$, B1 NDCG@10 $[0.57, 0.71]$. *Dense (B2 ko-SRoBERTa) > BM25 paired Wilcoxon $p<10^{-8}$; Hybrid vs. Dense is not statistically distinguishable on NDCG@10 (p=0.76). The cross-encoder reranker BGE-reranker-v2-m3 (BAAI multilingual) reranks BM25 top-200; the multilingual BGE-M3 embedding (BAAI/bge-m3) replaces ko-SRoBERTa as a single-vector dense retriever. An auxiliary Korean cross-encoder result (`ko-reranker`) is reported in Appendix D.

4.2. GT-2 Results: Intent-Only Retrieval

GT-2 is computed on the 66 × 4,937=325,842 query–policy cells with graded LLM-judge labels and is persona-independent. Table 3 reports NDCG@10, P@10 (graded ≥ 2), and R@10 (graded ≥ 2) for the four *rule-free* retrievers; B4 (Graph) and B6 (Rule+Dense) are inherently persona-conditioned and therefore have no natural query-only formulation, so the GT-2 column for them is omitted by design.

Finding 1. Korean-text dense retrieval (B2 ko-SRoBERTa) is the strongest rule-free retriever on GT-2 by a small but consistent margin (Figure 4). We were initially surprised that the multilingual BGE-M3 underperforms the Korean ko-SRoBERTa given that domain mismatch usually runs the other way; we attribute this to the persona-orthogonal queries staying close to Bokjiro's own phrasing, where Korean-specific contrastive pretraining pays off. Two cross-encoder rerankers further support the finding: (i) the multilingual BGE-M3 single-vector embedding scores −0.04 NDCG@10 below ko-SRoBERTa (0.815 vs. 0.858), suggesting that a Korean-specific embedding remains preferable to multilingual SOTA on this corpus; (ii) the multilingual cross-encoder BGE-reranker-v2-m3,

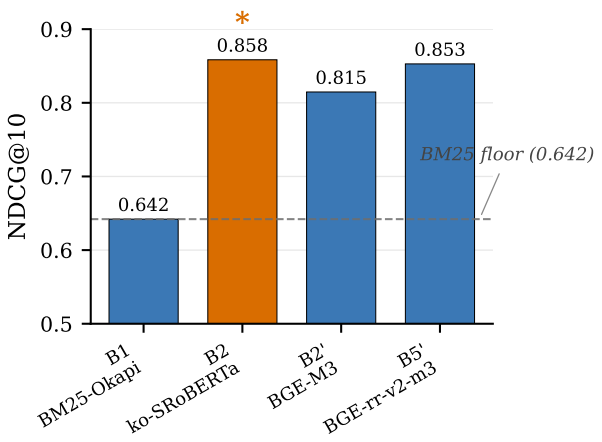

Figure 4. GT-2 (graded) NDCG@10 across baselines. ko-SRoBERTa (B2) and BGE-reranker-v2-m3 (B5′) tie at the top; the multilingual BGE-M3 underperforms on the persona-orthogonal query distribution. Dashed line: BM25 floor.

Baseline	NDCG@10	P@10	R@10
Rule-free			
B1 BM25	0.071	0.027	0.092
B2 Dense	0.119	0.051	0.133
B3 Hybrid	0.102	0.042	0.135
B4 Graph	0.122	0.047	0.126
Rule-augmented (matched prefilter)			
B5+rule	0.249	0.073	0.208
B6 (B2+rule)	**0.706**	**0.341**	**0.742**

Table 4. GT-3 strict (eligibility ∩ intent, strict region match). 11,880 (persona, query) pairs. **B5+rule** denotes B5 (Gemini-rerank) applied after the rule prefilter, the matched-control counterpart of B6 (Dense rerank after the same prefilter). B5+rule and B6 share the *same* rule prefilter; the only difference is the rerank function (Gemini vs. Dense). Holding the prefilter constant, Dense rerank dominates Gemini rerank by +0.46 NDCG@10. Lenient-grading counterpart: B6 NDCG@10 0.697, B5+rule 0.162.

applied as a rerank over BM25 top-200, attains 0.853 NDCG@10, within 0.005 of the Korean dense backbone but roughly 190× slower per query (§4.4). The hybrid (B3) does not improve over ko-SRoBERTa alone, so BM25 noise appears to hurt rather than help when fused. The Gemini LLM-rerank (B5) lifts BM25 from 0.642 to 0.789 NDCG@10 yet remains below B2.

4.3. GT-3 Results: Eligibility ∩ Intent

GT-3 is the metric a deployed assistant has to satisfy: a returned policy must be *both* eligible for the persona *and* topically relevant to the query. Table 4 reports the strict variant (graded = 2, strict region match; $\bar{n}_{\text{eligible}}$=4) over 11,880 (persona, query) pairs; the lenient variant follows the same ordering. Figure 5 visualises both panels jointly.

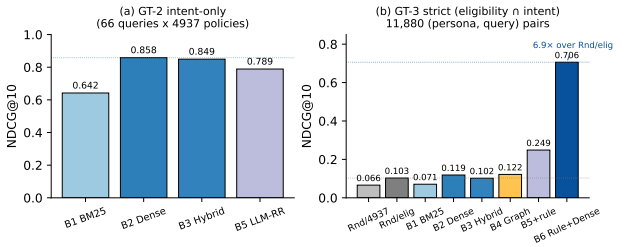

Figure 5. NDCG@10 across architectures on (a) GT-2 intent-only retrieval over 325,842 cells and (b) GT-3 strict (eligibility∩intent) over 11,880 (persona, query) pairs. B4 (Graph) and B6 (Rule+Dense) are persona-conditioned and therefore omitted from (a). In (b), the two leftmost gray bars are random baselines — uniform random over all 4,937 policies (0.066) and over the rule-eligible subset (0.103). B6 attains a ≈ 6.9× lift over the rule-conditioned chance level; B5+rule and B6 share the same rule prefilter, so the gap between them directly isolates the rerank-method contribution.

Finding 2. Without a rule prefilter, every rule-free retriever collapses to NDCG@10 around 0.12 on GT-3-strict (B1–B4 in $[0.07, 0.13]$), an order of magnitude below GT-2. Topical relevance alone is insufficient when eligibility is the gate.

Finding 3 (matched-prefilter diagnostic). B5+rule and B6 differ only in the rerank function (Gemini vs. Dense) atop an identical rule prefilter. Dense rerank dominates by +0.46 NDCG@10, so the rerank *ordering* contributes substantively over and above the prefilter; the operationally best recipe on GT-3 is rule-eligibility prefilter + Korean-domain dense rerank.

Sparsity caveat and random baselines. GT-3-strict is intrinsically sparse: $\bar{n}_{\text{el}}$=4 targets per (persona, query) pair and 20.86% of cells are empty (no policy is both rule-eligible and graded = 2). To confirm the absolute level of B6's NDCG@10= 0.706 is meaningful rather than an artifact of sparse target sets, we compare against two random baselines under the same metric. (i) Uniform random over all 4,937 policies yields NDCG@10≈ 0.066, the chance level when no signal is used at all. (ii) Uniform random over the rule-eligible subset (mean 97 candidates per persona) yields NDCG@10≈ 0.103, the chance level conditional on the eligibility prefilter alone. B6's 0.706 is ≈ 6.9× above this rule-conditioned chance level, confirming that the dense rerank ordering carries substantive signal beyond what the eligibility prefilter contributes by itself, and that the rule+dense gap over B5+rule (Finding 3) is not a sparsity artifact (Figure 5(b), where the two random baselines are shown as gray bars to the left of the architecture comparison).

4.4. Latency and Statistical Tests

Per-query latency on a single CPU to retrieve the top-10 over the 4,937-policy index is BM25 9.82 ms, Dense 5.99 ms, Hybrid (RRF) 14.90 ms, with one-time setup costs of 0.41 s, 9.93 s, and 10.34 s respectively. Dense is fastest at query time and Hybrid pays the sum.

The 95% bootstrap CIs in Tables 3–4 use 10,000 resamples (query-level for GT-2, (persona, query)-level for GT-3). Pairwise comparisons report paired Wilcoxon p-values with Holm–Bonferroni correction. On GT-2, B2 Dense beats B1 BM25 at $p<10^{-8}$ (NDCG@10); B3 Hybrid vs. B2 Dense is not distinguishable (p=0.76). On GT-3 strict, B6 dominates every other architecture at $p<10^{-3}$ (Holm-corrected).

5. Conversational Architecture

Every retrieval baseline in §4 presupposes that the user knows and articulates every relevant eligibility attribute. The Korean welfare access problem violates this assumption: elderly, digitally underserved, and multiply-vulnerable users cannot promptly answer "are you below 75% of median income" nor author a single query expressing both need and eligibility profile. The architecture in this section elicits attributes across turns.

5.1. Five-Stage Closed Loop

The deployed conversational pipeline (Figure 6) is a closed loop of five stages that consumes the entire prior dialogue history at each turn and emits both an updated ranked policy list and the next clarification question, terminating when the user accepts a recommendation or no informative attribute remains.

S1 NL input. Free-form utterance (voice or text). **S2 NL→tag extraction.** An LLM extractor maps the running dialogue history to a partial 9-dimensional persona attribute vector plus the free-text intent; each slot is *revealed*, *negated*, or *unknown*. **S3 Query rewriting.** The user's utterance is rewritten conditioned on the partial attribute vector, expanding domain-elided context (e.g., adding "rent assistance" when the user said "*my son's housing*") and discarding tokens that duplicate already-revealed attributes. **S4 Eligibility-aware recommendation.** The current attribute vector restricts the GT-1 candidate set, the rewritten query restricts the GT-2 candidate set, and the rule + dense rerank of §4 produces a ranked top-K. **S5 Follow-up question.** The system selects the next attribute to elicit from among the still-unknown slots and surfaces it as a natural-language question.

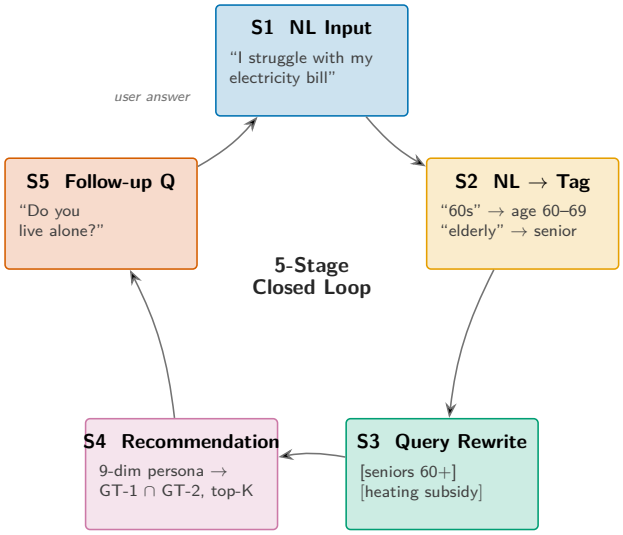

Figure 6. Five-stage conversational loop. S1 receives a free-form utterance (voice through STT or text); S2 updates a partial 9-dimensional persona vector (revealed/negated/unknown); S3 rewrites the query conditioned on the partial vector; S4 retrieves eligibility-aware candidates (rule prefilter + dense rerank); S5 selects the next attribute to elicit. The user's reply re-enters S1.

5.2. Evaluation Scope, Soft Eligibility, Selector

Scope. Stages S2–S3 are deployed but not benchmarked here: KWELFAREBENCH governs S4 (eligibility-aware retrieval) and S5 (follow-up selection). We evaluate the architecture as *progressive attribute revelation*: at each turn the system invokes S5 to select an attribute, and an oracle reveals its true value, simulating a successful S1–S3 loop. **Soft eligibility.** We replace the hard eligibility predicate with a *soft* score: revealed attributes contribute hard $\{0, 1\}$, unrevealed attributes contribute the corpus marginal $\hat{P}_t$. Full equations in Appendix C.

Trained selector positioning. As a simpler counterpart to CLARINET (Chi et al., 2024), we train a `RandomForest` (n=200, depth 10) on 29-dim features over an 80/20 persona-level split (144/36); both $\hat{P}_t$ and the selector are estimated only on the 144 training personas. Hyperparameter detail in Appendix D.

LOPO robustness check. Leave-one-persona-out CV on the 144 training personas yields turn-5 Recall@10 = 0.090 $[0.085, 0.094]$ for trained vs. 0.074 $[0.069, 0.078]$ heuristic (paired bootstrap +0.016, significant), tracking the leak-free 0.095 closely; per-fold detail in Appendix D.

Simpler clarification baselines. The most informative comparator is ϵ-greedy (ϵ=0.1) over attribute information-gain (IG), which requires no learned model. On the 36 leak-free test personas at turn 5 the trained selector reaches 0.095

Turn	Heuristic	IG dyn.	**Trained ML**	Oracle
0	0.003 [.002,.005]	0.003 [.002,.005]	0.003 [.002,.005]	0.003 [.002,.005]
1	0.001 [.000,.003]	0.059 [.050,.067]	**0.060** [.051,.069]	0.063 [.055,.071]
2	0.009 [.006,.012]	0.087 [.081,.094]	0.084 [.077,.091]	0.090 [.083,.097]
3	0.073 [.066,.080]	0.091 [.085,.098]	0.090 [.083,.097]	0.098 [.091,.106]
4	0.071 [.063,.080]	0.090 [.084,.096]	0.096 [.089,.103]	0.102 [.095,.110]
5	0.071 [.062,.080]	0.095 [.087,.102]	0.095 [.088,.103]	0.102 [.095,.111]

Table 5. Recall@10 over turns on n=36 leak-free test personas (bootstrap mean and 95% CI in brackets, N_{boot}=10,000). Trained ML and IG both beat the Heuristic from turn 1 onward (paired Wilcoxon $p<10^{-4}$). Trained ML and IG are not statistically distinguishable in accuracy at any turn ($p \in [0.22, 0.88]$); the differential is inference cost.

[0.088, 0.103] and ϵ-greedy(IG) reaches 0.092 [0.084, 0.100] (paired diff $+0.003$ [-0.004, $+0.011$], n.s.); heuristic 0.071, random 0.050, oracle 0.103. We therefore position the trained selector not as a methodological contribution but as a deployment-feasibility demonstration: when ϵ-greedy(IG) already matches accuracy, the operational question is inference cost, where the RandomForest's $\sim$1 ms predict beats IG's $\sim$60 ms (10 dense reranks) per turn—more than an order of magnitude. A Thompson-sampling baseline and the CLARINET head-to-head are deferred to Appendix D and future work, respectively.

5.3. Results: Trained ML vs. IG vs. Heuristic vs. Oracle

We compare four selectors over turns 0–5 with identical retrieval (soft eligibility plus dense rerank, §5.2): **Heuristic** (fixed priority list, no learning), **IG dynamic** (exhaustive per-turn IG trial, no learning), **Trained ML** (the RandomForest above), and **Oracle** (lookup of the actually best attribute at training time).

Inference-cost win. IG dynamic must simulate soft scoring and dense reranking for every candidate attribute (up to ten) at every turn—roughly ten retrieval calls per persona-turn, each over 4,937 policies. The Trained ML selector instead requires a single `predict_proba` call ($\sim$ 1 ms). When IG and ML are statistically tied on accuracy (Table 5), ML wins on user-facing latency, the operational constraint of any deployed dialogue system. At turn 5 both reach roughly 93% of the oracle ceiling (0.0953/0.0950 vs. 0.1025).

Narrowing the access gap. Trained ML at turn 5 (0.0953) is roughly an order of magnitude above the single-shot BM25 query-only baseline (0.0071, 180 personas, eligibility-blind); the two settings differ on multi-turn vs. single-shot, eligibility-aware vs. blind, and n=36 vs. 180, so the figure quantifies the joint architecture benefit, not the marginal selector contribution. The matched-n=36 selector gap is trained$-$heuristic $+0.025$ at turn 5 (Table 5, paired bootstrap significant).

6. Limitations

L1–L6 bound the *benchmark*; L7–L12 bound the *architectures*. **(L1)** Eligibility conditions outside the 69-binary + 3-numeric schema are absent by construction. **(L2)** April 2026 snapshot; policy drift and seasonal programs are not measured. **(L3)** Sub-regional preferences in PDF/HWP attachments are partially recovered by the GT-2 free-text channel but absent from GT-1. **(L4)** Sub-municipal regions (`dong`, `eup`, `myeon`) are not represented. **(L5)** 180 personas cover five reference groups and 18 compound axes; some intersectional cells (e.g. rural single-parent foreign-born) have single-digit counts. The generator is shipped seed-pinned for community-extensible expansion targeting 1,000+ personas. **(L6)** Reviewers behind every disagreement-adjudication are graduate students and domain practitioners, not licensed social workers; their output is a strong tier above pure LLM labels but below a formal expert audit.

(L7) LLM-as-judge at temperature 0 with cross-LLM κ_w=0.69; provider model rotation behind aliases requires dated snapshot identifiers. **(L8)** The query pool assumes the user is the prospective beneficiary; family-proxy search is out of scope. **(L9)** Two query phrasings per sub-topic; paraphrase robustness and spoken disfluencies are absorbed upstream by the conversational architecture, not stress-tested at the pool level. **(L10)** The 80/20 conversational split is in-distribution; the 1,000+-persona L5 expansion is the planned OOD route. **(L11)** STT error and upstream NL-stage mismatch are not quantified here; a controlled user study with eligibility-recovery scores is in progress. **(L12)** Calibrated for the Korean Bokjiro taxonomy; cross-jurisdiction transfer requires a sub-topic discovery pass.

7. Conclusion and Social Impact

KWELFAREBENCH is a cell-level Korean welfare evaluation resource (4,937 policies, 180 personas, 69-binary schema, 66-query pool, three ground-truth tables). Korean-domain dense retrieval is the strongest rule-free baseline on GT-2 (NDCG@10 0.86, $p<10^{-8}$ over BM25); no rule-free retriever exceeds NDCG@10 0.13 on GT-3-strict; under a matched rule prefilter, dense rerank beats Gemini LLM rerank by $+0.46$. A 1 ms RandomForest selector matches exhaustive IG within 4% recall at 60$\times$ less inference. Risks are mitigated by surfacing matches as *candidates* (agency-confirmed) and by two-LLM full pass plus human adjudication on every LLM-judged layer; personas are synthetic (no PII). Future work: 1,000+-persona expansion, credentialed-expert audit, trained NL$\rightarrow$tag at S2, and real-user evaluation of S1–S3.

References

Alam, S. M., Zou, H., Vir, R., and Salehi, N. SAGE: System for accessible guided exploration of health information. In *AAAI Workshop on Public Sector LLMs (PubLLM)*, 2024. URL https://publlm.github.io/.

Ban, S., Kim, Y., Yoon, Y., Kim, J. H., and Lee, J. Design and implementation of a mobile system to provide personalised welfare information for people with disabilities: A pilot study. *Digital Health*, 2025.

Chi, Y., Lin, J., Lin, K., and Klein, D. CLARINET: Augmenting language models to ask clarification questions for retrieval. *arXiv preprint arXiv:2405.15784*, 2024. Rank-aware clarification question generation; cited as prior conversational architecture.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. Foundational weighted-kappa definition; quadratic-weighted variant used for our 0/1/2 ordinal labels.

Cormack, G. V., Clarke, C. L. A., and Büttcher, S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 758–759. ACM, 2009. doi: 10.1145/1571941.1572114.

Gao, Y. et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2024.

Google DeepMind. Gemini 2.5 Flash-Lite: A fast, multimodal LLM for throughput-optimized inference. https://ai.google.dev/gemini-api/docs/models/gemini-2-5-flash-lite, 2025. Used as the cross-vendor judge for GT-2 LLM-as-judge agreement check.

Han, S., Titericz Jr., G., Balough, T., and Zhou, W. Judge's Verdict: A comprehensive analysis of LLM judge capability through human agreement. *arXiv preprint arXiv:2510.09738*, 2025. NVIDIA; 54-LLM Cohen's kappa cross-judge study; cited as concurrent work on cross-LLM agreement reporting.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. Holm–Bonferroni correction across the seven baseline comparisons.

Huggins-Daines, D. Try that in a small town: Large language models for modest municipalities. In *AAAI Workshop on Public Sector LLMs (PubLLM)*, 2024. URL https://publlm.github.io/.

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. Constraint-based recommender systems. In *Recommender Systems Handbook*. Springer, 2015.

jhgan. ko-sroberta-multitask. https://huggingface.co/jhgan/ko-sroberta-multitask, 2021. Korean Sentence-BERT, multitask-trained on KorNLI/KorSTS.

Jin, Q., Wang, Z., Floudas, C. S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., and Lu, Z. Matching patients to clinical trials with large language models (TrialGPT). *Nature Communications*, 2024. URL https://arxiv.org/abs/2307.15051.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Kaun, A. and Männiste, M. Public sector chatbots: AI frictions and data infrastructures at the interface of the digital welfare state. *New Media & Society*, 2025. URL https://journals.sagepub.com/doi/10.1177/14614448251314394.

Le, N. L., Abel, M.-H., and Gouspillou, P. A constraint-based recommender system via RDF knowledge graphs. *arXiv preprint arXiv:2307.10702*, 2023. Vehicle-purchase domain.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Nelson, W., Lee, M. K., Choi, E., and Wang, V. Designing LLM-based support for homelessness caseworkers. In *AAAI 2024 Workshop on Public Sector Large Language Models (PubLLM)*, 2024.

OpenAI. GPT-5.4-nano: Cost-efficient frontier reasoning at production latency. https://platform.openai.com/docs/models/gpt-5.4-nano, 2026. Model card; structured-output (JSON Schema) enforcement with sub-200 ms p50 latency.

Reddy, V. and Bell, J. Government schemes recommendation API with multi-language chatbot using RAG vector search and conversational AI. In *2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2025. URL https://ieeexplore.ieee.org/document/11200536/. IEEE Xplore document 11200536.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. URL https://aclanthology.org/D19-1410/.

Robertson, S. and Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 2009.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*, 2021. Foundational zero-shot IR benchmark; graded relevance qrels.

Trotman, A., Puurula, A., and Burgess, B. Improvements to BM25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium (ADCS)*, 2014. Empirical guidance on k_1 and b for the Okapi parameterisation.

US Department of Health and Human Services. Public benefits and AI. Technical report, HHS, 2024. URL https://www.hhs.gov/sites/default/files/public-benefits-and-ai.pdf.

Wang, Y. et al. Intersectional two-sided fairness in recommendation. In *Proceedings of the ACM Web Conference (WWW)*, 2024.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. Foundational; signed-rank test used for paired persona-level significance.

Zhao, Z., Vania, C., Kayal, S., Khan, N., Cohen, S. B., and Yilmaz, E. PersonaLens: Personalized recommendation benchmark with rich user profiles. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 2025. URL https://aclanthology.org/2025.findings-acl.927/. Cell-level personalization benchmark with rich user profiles; concept-isomorphic to KWB but task-oriented assistant rather than welfare.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023. LLM-as-judge evaluation framework; cited for the methodology our cross-LLM judge agreement check borrows from.

A. Full Tag Schema

The KWELFAREBENCH labeling schema groups 69 binary tags under five top-level macro-categories together with three numeric fields. Each binary tag takes one of three values: -1 *excluded* (disqualifies the applicant), 0 *indifferent* (not mentioned), $+1$ *required*. The same vocabulary is used on both the policy side (A3) and the persona side (A5), so the GT-1 eligibility function operates on a single shared schema.

A.1 Binary Tag Inventory

- **PERSONAL (14 binary).**
  - `personal.gender (2)` — female_only, male_only.
  - `personal.disability (5)` — required_any, severe, developmental, visual, hearing.
  - `personal.health (7)` — chronic_or_severe, rare_or_cancer, mental_health, dementia, diabetes, hypertension, depression.
- **ECONOMIC (14 binary).**
  - `economic.income (3)` — basic_recipient, secondary, medical_aid.
  - `economic.employment (7)` — unemployed, self_employed, startup, agriculture_fishery, small_business_sme, public_servant_military, industrial_accident.
  - `economic.housing (4)` — no_house, jeonse, monthly_rent, rental.
- **HOUSEHOLD (14 binary).**
  - `household.type (7)` — single, single_parent, multi_child, newlywed, multicultural, grandparent, youth_minor.
  - `household.pregnancy (4)` — pregnant_postpartum, infertility, unmarried_mother, birth.
  - `household.birth_order (3)` — first, second, third_plus.
- **SOCIAL (20 binary).**
  - `social.veteran (4)` — war_veteran, national_merit, independence, veteran_family.
  - `social.immigrant (3)` — foreigner_general, north_korean_defector, late_arrival_child.
  - `social.vulnerable (6)` — vulnerable_general, crisis_household, homeless, solo_elderly, foster_or_protected_child, adoption.
  - `social.violence_victim (1)` — violence.
  - `social.education (4)` — elementary_to_high, university, out_of_school, ged.
  - `social.residence (2)` — resident_required, long_term_resident.
- **POLICY (7 binary).**
  - `policy.category (7)` — welfare, employment, education, housing, health, culture, living. Inherits the seven Bokjiro operational categories.

A.2 Numeric Fields (3)

- `age_min` (`int|null`) — lower bound; `null` when no lower bound is stated.
- `age_max` (`int|null`) — upper bound; `null` when no upper bound is stated.
- `median_threshold` (`int|null`) — percentage of national median income required for eligibility; `null` when no income threshold is stated.

All three numeric fields are populated only when explicitly stated by the policy text.

A.3 Schema Example: Policy and Persona Tagging

To illustrate how a single schema is shared between the policy and persona sides of GT-1, we walk through one (policy, persona) cell.

Policy: "Seoul Youth Monthly Rent Support". The encoded `tag_values` (showing only nonzero slots) are:

- `personal.gender`: all 0 (gender-indifferent).

- `economic.income.basic_recipient`=+1 (required: basic-recipient or secondary status).
- `economic.income.secondary`=+1 (required).
- `economic.housing.no_house`=+1 (required: no ownership).
- `household.type.single`=+1 (required: single-person household).
- `social.residence.resident_required`=+1 (required: registered Seoul resident).
- `policy.category.housing`=+1.
- `age_min`=19, `age_max`=39, `median_threshold`=150 (i.e. $\leq 150\%$ median).

Persona P_YOU_03 (Youth group). Synthetic `tag_values`: age=24, gender=female, sido="Seoul", sigungu="Gwanak-gu", `economic.income.secondary`=+1 (satisfied), `economic.housing.no_house`=+1 (satisfied), `household.type.single`=+1 (satisfied), `social.residence.resident_required`=+1 (satisfied), median_income_pct=120.

GT-1 cell evaluation. The deterministic eligibility function checks all nine dimensions (§3.6): age $19 \leq 24 \leq 39$ ✓; income secondary required and persona has secondary status ✓; `no_house` required and persona satisfies ✓; `single` household required and persona satisfies ✓; Seoul residency required and persona is in Gwanak-gu (Seoul) ✓; median threshold 150 and persona at 120 ✓. Result: `eligible(P_YOU_03, this policy)`=1.

GT-2 cell evaluation. Independently, query Q_{17} ("*rent is becoming hard to afford*", a persona-orthogonal situational phrasing under R1) is graded by the LLM judge against this policy. Both `gpt-5.4-nano` and `gemini-2.5-flash-lite` return grade $= 2$ (direct topical match), so this (query, policy) cell is in GT-2 at the strict threshold.

GT-3 cell. Since GT-1$= 1$ *and* GT-2≥ 2 *and* the strict region predicate matches (persona Seoul, policy Seoul), this (persona, query, policy) triple lies in GT-3-strict.

B. Baseline Pseudocode

B.1. B1 BM25

```
def tokenize(text):
    # Korean word tokens + 2-gram syllable backoff
    words = re.split(KOREAN_WORD_BREAK, text)
    tokens = []
    for w in words if len(w) >= 2:
        tokens.append(w)
        for i in range(len(w) - 1):
            tokens.append(w[i:i+2])
    return tokens

def fit(policies):
    return BM25Okapi([tokenize(policy_text(p)) for p in policies])

def retrieve(persona, k):
    q = tokenize(persona.query + persona.attribute_keywords)
    return top_k_by_score(policy_ids, bm25.get_scores(q), k)
```

B.2. B2 Dense Embedding

Model: `jhgan/ko-sroberta-multitask` (768-d Korean SBERT). Policy text and persona query are encoded and ranked by L2-normalized cosine similarity.

B.3. B3 Hybrid (Reciprocal Rank Fusion)

$$\text{score}_{\text{hybrid}}(p) = \frac{1}{k_{\text{rrf}} + r_{\text{BM25}}(p)} + \frac{1}{k_{\text{rrf}} + r_{\text{dense}}(p)}, \quad k_{\text{rrf}}{=}60.$$

Each retriever contributes its top $\max(K{\times}5, 100)$ candidates.

B.4. B4 Graph (Tag Overlap)

Let $R(p) = \{t : \text{label}(p,t) = +1\}$ and $S(\pi)$ be persona π's satisfied-tag set; the score is $|R(p) \cap S(\pi)|/\sqrt{|R(p)| \cdot |S(\pi)|}$.

B.5. B5 LLM-rerank

BM25 prefilters 50 candidates; `gemini-2.5-flash-lite` receives the persona context and the candidate list and returns top-K identifiers in comma-separated form with structured-output enforcement.

B.6. B6 Rule + Dense rerank

1. Apply the rule eligibility function to obtain the candidate set (mean 97 policies).
2. Score each candidate's dense embedding against the persona query embedding (cosine).
3. Return the top-K.

B2+rule is by construction identical to B6 (different naming emphasizes the matched-prefilter perspective in Table 4).

C. Soft Eligibility Scoring

For persona π, policy q, revealed attribute set $R \subseteq \mathcal{A}$:

$$\text{score}(\pi, q; R) = \sum_{t \in T_q} \log \tilde{P}_t(\pi, q; R),$$

where $T_q = \{t : \text{label}(q,t) \neq 0\}$ and

$$\tilde{P}_t = \begin{cases} P_t(\pi) & \text{if } a(t) \in R \text{ (hard match, } \in \{0,1\}) \\ \hat{P}_t & \text{if } a(t) \notin R \text{ (marginal estimate)} \end{cases}$$

$a(t)$ is the persona attribute on which tag t depends. When $\text{label}(q,t) = -1$, we substitute $1 - \tilde{P}_t$. To avoid $\log 0$ we clip to $[10^{-6}, 1.0]$. The marginal $\hat{P}_t$ is estimated on the 144 training personas: $\hat{P}_t = N_{\text{train}}^{-1} \sum_\pi \mathbb{1}[\pi \text{ satisfies } t]$. For numeric and region tags: when revealed, $\log(10^{-6})$ is added on mismatch (effective disqualification); when unrevealed, the marginal is used.

D. Persona Examples, Reproducibility, and Auxiliary Results

Persona examples. **P_DIS_01 (Disability group):** age 65, male, Seoul Seodaemun-gu, secondary income, present disability, employed; query "*Can my disabled child receive education-fee support?*" **P_INT_15 (Intersectional):** age 73, female, Busan Haeundae-gu, basic-recipient income, present disability, single-person household, special target solo elderly. **P_GEN_22 (General):** age 35, male, Gyeonggi Suwon, general income, no disability, newlywed household, employed.

A3 tagging cost. We ran the A3 tag-extraction prompt against `gpt-5.4` (April 2026, temperature 0) with 16 concurrent workers; the full 4,937-policy pass completed in roughly 13 minutes. The comma-separated output format reduces output tokens $790 \rightarrow 144$ per policy on average ($5.5\times$); total cost is approximately $23 for the full corpus.

Index build. BM25 index build is 0.41 s on a single CPU; dense and hybrid indices build in 9.93 s and 10.34 s respectively.

Auxiliary results. **ko-reranker:** `Dongjin-kr/ko-reranker` reranks BM25 top-200, NDCG@10 0.803 (below BGE-reranker-v2-m3 0.853 and BGE-M3 0.815). **LOPO per-fold:** held-out per-persona Recall@10 trained 0.090 [0.085, 0.094]

vs. heuristic 0.074 $[0.069, 0.078]$, paired trained–heuristic $+0.016$ $[+0.013, +0.020]$; per-fold dispersion released with result tables. **Thompson sampling:** Beta(1,1) prior per attribute, updated with observed IG. At the 5-turn budget the policy under-explores (0.049 $[0.037, 0.061]$, numerically below random 0.050).