

Agentic LLM for Information Access among Refugees: A Field Report from a Resource-Constrained Organization

Nari Yoo

University of Michigan
School of Social Work
nariyoo@umich.edu

Youngah Kim

Migration to Asia Peace
Seoul, Republic of Korea
map@mapcast.org

Abstract

Machine translation can make a government notice readable in a refugee’s language, but a faithful translation of bureaucratic Korean is still bureaucratic, and newcomers often cannot act on it. We report, with the operating organization as a co-author, what happened when a small refugee-serving organization in South Korea put a large language model in front of the translation step at the service it has operated since 2018. Scoped with the frontline staff who field refugees’ questions, the model rewrites the Korean source into plain language, explains Korea-specific terms where they appear, checks whether figures and deadlines have gone out of date, and pins proper names to a fixed glossary before anything is translated into seven languages. It covers most of the catalog and shortens the sentences readers meet. It also fails in ways a translation-only system never does. The stages allowed to search the web made corrections no translation could make, and they also wrote every serious error we found, including policy updates that were simply wrong; the stages without that access wrote a handful, none of them critical and all of them confined to metadata. Value and failure come from one capability, so the lesson is not to withhold it but to put the human check where it acts. For an organization with nobody on staff who reads the seven target languages, that check must be placed on the Korean source before publication, which makes Korean-language staff review the safeguard the whole design depends on.

1 Introduction

Claiming asylum, meeting a filing deadline, and applying for a benefit all depend on receiving the right information in time, in a language they read (Quadri et al., 2023; Mannan et al., 2025), which makes in-language access a structural determinant of protection (World Health Organization, 2008; Flores et al., 2003; Karliner et al., 2007). For a refugee-serving organization in South Korea, the

languages that matter most are the ones the country can least staff and that language technology serves worst (Joshi et al., 2020; Blasi et al., 2022; Robinson et al., 2023), so machine translation (MT) has become the practical way to publish across them (Valdez et al., 2023; Todorova and Liu, 2024).

A second barrier remains after translation. When the authoritative source is written in the language of the host state rather than in English, it is dense administrative prose: a notice about residence status, a benefit threshold expressed as a fraction of median income, an appeal deadline in bureaucratic register. A faithful translation of such a notice is a faithful translation of bureaucracy, and the South Korean government’s own public COVID-19 portal was measured at a median professional reading level, with 90% of its documents too hard for the general public (Moon et al., 2022). The simplification, therefore, has to happen at the source, in Korean, before translation into seven languages.

Generative LLMs make this possible as a translation engine cannot. Beyond translating, an LLM can rewrite a dense notice into plain language, add an explanation of a Korea-specific term (Alva-Manchego et al., 2020; Joseph et al., 2023; Kim et al., 2020), and, given a web-search tool, check whether a figure or a deadline has changed and refresh it (Schick et al., 2023; Yao et al., 2023; Nakano et al., 2021; Mialon et al., 2023). This is attractive where readers cannot verify what they receive. It also introduces a generation-specific risk, since an LLM can produce fluent output that has silently added, dropped, or invented content, with no marker to signal that it should be checked (Guerreiro et al., 2023; Ji et al., 2023; Maynez et al., 2020). These are two descriptions of one capability, and a deployment can separate them empirically when its components differ in what they are allowed to do, and this deployment’s components do. The enhancement consisted of a series of generative stages over the Korean catalog, some given a

web-search tool and the freedom to act over several steps and some given neither, and commit history records which stage wrote every sentence readers see.

This paper is a field report written from inside the deployment, co-authored by a researcher in migrant and refugee welfare with working knowledge of the AI and the executive director of Migration to Asia Peace (MAP), who has operated its service, Refugee News, since 2018. Our research question is whether dense administrative Korean can be rewritten into plain language for newcomers without losing legal, procedural, or factual meaning, and where an LLM enhancement fails in its attempt to do so. Translation reliability across the seven languages is out of scope here.

We frame the study using the taxonomy of implementation outcomes (Proctor et al., 2023). A production audit addresses penetration, feasibility, fidelity, and cost; effectiveness lies outside what it can establish (Glasgow et al., 1999). Our contributions are: (1) a deployed, low-cost, agentic enhancement design, scoped with practitioners, that a small organization applies to every edit; (2) a production audit of coverage, readability, and generation-specific failure modes over a 246-article record, with an adjudicated faithfulness review of 17 articles, automatic flagging over all 79 guides the agentic currency update touched, and the organization’s own staff catalog review; (3) worked examples, including a policy-currency correction that agentic web search made possible; and (4) an account of the value and cost of the model’s agency, in which every confirmed error is attributed to the stage that wrote it, holding the model and execution context constant and leaving tool access as the property the failures follow.

2 Background and related work

Text simplification and elaboration. Simplification is an established NLP task (Siddharthan, 2014; Alva-Manchego et al., 2020) now dominated by LLMs (Kew et al., 2023; Agrawal and Carpuat, 2023). Ours is only partly deletion-style, since it also adds explanatory content, which the literature terms “elaborative simplification” (Srikanth and Li, 2021), and the closest designs pair plain-language definitions with in-situ summaries for lay readers (August et al., 2023; Devaraj et al., 2021; Manor and Li, 2019; Stodden et al., 2023; Färber et al., 2025). Evaluation is a known difficulty, because

reference-based metrics require gold-standard simplifications that the deployment lacks (Sulem et al., 2018; Alva-Manchego et al., 2021; Xu et al., 2016), formula-based readability scores are weak targets (Tanprasert and Kauchak, 2021; Redish, 2000), and simplification can silently insert, delete, or substitute facts (Devaraj et al., 2022). We therefore report structural statistics computed on a morphological analysis of the Korean, deterministic audits, and an adjudicated review.

Plain language for limited-proficiency readers. Simplifying official text improves access for limited-proficiency newcomers (Kim et al., 2020; Berkman et al., 2011; Yameogo et al., 2025), and inline term explanation for lay readers is an established design (Kloehn et al., 2018; Leroy et al., 2013).

Agentic LLMs and faithfulness risks. Tool-augmented and retrieval-augmented LLMs act over several steps and consult external sources (Schick et al., 2023; Yao et al., 2023; Nakano et al., 2021; Lewis et al., 2020; Mialon et al., 2023). Our policy-currency function is a search-grounded revision of existing text, studied as attributed revision and factual freshness (Gao et al., 2023; Vu et al., 2024), and the attribution literature sets the standard such edits should meet: each change traceable to an identified source (Rashkin et al., 2023; Liu et al., 2023). The capability that lets the model refresh a stale figure is the capability that lets it hallucinate one (Ji et al., 2023; Maynez et al., 2020; Dale et al., 2023; Min et al., 2023). Keeping entities and numbers correct is terminology and critical-error control (Dinu et al., 2019; Post and Vilar, 2018; Hokamp and Liu, 2017; Alam et al., 2021; Al Sharou and Specia, 2022), which motivates our glossary.

Humanitarian language technology and refugee-serving deployments. MT for displaced and migrant populations is studied through crisis engines and field deployments (Anastasopoulos et al., 2020; Öktem et al., 2020; Way et al., 2020; Cadwell et al., 2019; Lewis, 2010) and human-in-the-loop platforms built with and for refugees (Brown and Grinter, 2016). Editing the source before translation continues the pre-editing tradition in crisis translation (O’Brien and Federici, 2019); the source edit here is generative and agentic, and audited from inside a deployment. In the Korean context, refugees report high yet unmet information needs (Kwon and Lee, 2022), language barriers obstruct

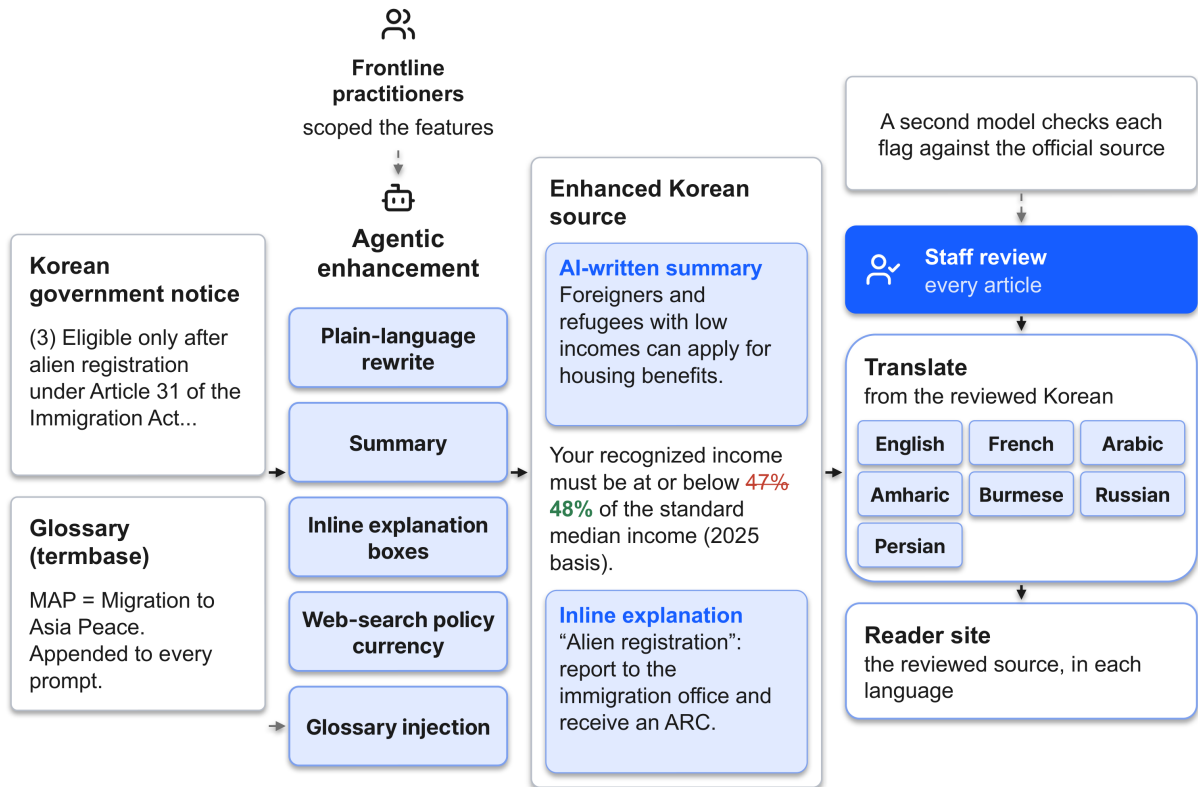


Figure 1: The deployed pipeline of Refugee News, shown on the housing-benefit guide. The correction it displays, 47% to 48% of the standard median income, follows MOLIT notice No. 2024-445, in force from January 2025. The guide is an extreme rewrite case, 227.0 to 24.3 morphemes per sentence; Table 1 gives the stages.

service access for refugees and migrants (Kim et al., 2017; Lee, 2021), and North Korean refugees use government digital services more than native-born Koreans (Yoo et al., 2026b).

Implementation outcomes and participatory NLP. We use Proctor et al. (2011)’s outcome taxonomy (extended in Proctor et al., 2023) with process frameworks (Moullin et al., 2019; Damschroder et al., 2022), RE-AIM (Glasgow et al., 1999), and implementation science for AI in nonprofit human services (Trinkley et al., 2024; Wutz et al., 2023; Goldkind et al., 2025); because the tool was scoped with practitioners we also draw on participatory NLP (Cooper et al., 2022; Bird, 2020; Delgado et al., 2023; Birhane et al., 2022; Sambasivan et al., 2021) and documentation norms (Bender and Friedman, 2018; Mitchell et al., 2019). Most AI-for-good work reports model building; this paper joins the smaller strand that reports deployment lessons (Diddee et al., 2022; Nekoto et al., 2020; Lin et al., 2024; Ramjee et al., 2025; Mate et al., 2022; Nair et al., 2023; Abilov et al., 2025).

3 Setting and source-language problem

Refugee News is operated by MAP, a refugee and migrant support organization in South Korea, and this report is co-authored with the organization. It publishes news and practical guidance for asylum seekers and refugees, with Korean as the authoritative source, and translates it into English, French, Arabic, Amharic, Burmese, Russian, and Persian. The outer context (Moullin et al., 2019) is a funding stream contracting after the 2025 suspension of United States foreign aid (OHCHR, 2025; UNHCR, 2025), an immigration regime that decides which information is high-stakes, and almost no professional translators for the target languages. The inner context is a small nonprofit with no staff who read those languages well enough to verify a translation and no budget to hire them; its readers already face information precarity (Wall et al., 2017).

The source-language problem is specific to a non-English-source deployment. The corpus is 246 non-draft articles, 104 guides, and 142 news items (topics and comprehension stakes in Table 2). Much of it originates as Korean government no-

tices: eligibility rules as cross-references to statute, benefit thresholds as a percentage of median income, appeal deadlines in bureaucratic register. The catalog is also uneven, written over years amid staff turnover and reliance on volunteers, and a faithful translation inherited that inconsistency. The organization had already moved from scarce paid human translation to Google Translate, which broadened coverage while rendering the source faithfully; this enhancement is the step meant to close the remaining gap.

Participatory scoping. The feature set was scoped with the practitioners who operate the service. The executive director brought the editorial policy and operating constraints; the frontline staff and volunteers who field refugees' questions were included by role, as the people who see where a faithful translation still fails readers, and were not sampled for representativeness. Across three working discussions, practitioners identified recurring points at which readers could not act on a translated notice, the researcher proposed a feature for each, and the executive director validated it against what the organization could sustain and its policy of publishing advocacy statements verbatim. On the participation ladder of [Delgado et al. \(2023\)](#) this is consultation on problem definition and collaboration on scoping: practitioners held the problem list and veto, and the researcher held proposal power. They named four needs, and §4 answers them one for one (Appendix M). They were sharpest, and unprompted, about proper names never varying. This is practitioner-informed scoping. Refugee end-users were informed by the staff who serve them and were not direct co-designers ([Sambasivan et al., 2021](#); [Birhane et al., 2022](#)).

Grounds for publication. General-purpose assistants with web search can answer much of what this catalog covers, and the staff use them daily. The pipeline exists because publishing and answering are different acts. Newcomers have to know a question exists before asking it, and nobody asks about a deadline that moved or a threshold that was raised; a published guide is delivered to readers who were not looking for it. The organization is answerable for what it publishes as no assistant is for what it says, and that accountability makes a Korean-reading staff review possible. An assistant querying the open web also lands on the same superseded government pages the currency stage had to correct, with nobody between the reader

and the answer, and readers in information precarity do not uniformly hold the device, the data or the second-language confidence a query interface assumes ([Wall et al., 2017](#); [Kwon and Lee, 2022](#)).

4 Enhancement system

The enhancement is applied to the Korean source before translation, at the head of an automated pipeline (Figure 1). The stages behind it differ in what the model was allowed to do, and Table 1 gives each stage its model, tool access, edit latitude, and execution context. §8 uses those four properties to attribute what the enhancement added and what it broke.

A one-time catalog-modernization rewrite converted the accumulated back catalog into plain, warm wiki prose: first over 116 articles with claude-opus-4-8, then over the remaining catalog with claude-sonnet-4-6, both given a web-search tool (at most four searches per article) to repair links and refresh institutional facts. This is the stage that restructures the administrative prose, and the readability contrast of §5 measures it. A tool-less stage then wrote the summary for every article, and a policy-currency update refreshed the 79 dated guides with web search (at most five searches per article). Since launch, every edit to a Korean article triggers an automatic step that adds a summary and explanation boxes, is constrained to leave the article body unchanged, and is followed by staff review and then translation. Every prompt is issued as a single request with the glossary appended; no stage chains model calls. Across the stages the enhancement performs five functions.

Plain-language rewriting. The catalog rewrite converts each article into an accurate, easy-to-understand wiki document in a single warm register, flattening accumulated quality and tone differences, removing internal management tags and editing artifacts, repairing broken links via web search, and appending a standing note to confirm figures and deadlines with the agency.

Plain-language summary. A one-to-two-sentence Korean summary (mean 125 characters) states what the article says and what the reader should do, and is shown at the top of the page with text-to-speech. The summary stage wrote it for the catalog at build; the per-edit step writes it for any article edited since. Advocacy statements, which the organization publishes verbatim, receive only

Stage	Model	Art.	Web	Latitude	Execution
Catalog rewrite	opus-4-8, then sonnet-4-6	116; 244	yes	full rewrite	at build
Summary	sonnet-4-6	203	no	one added field	at build
Policy currency	sonnet-4-6	79	yes	revise in place	on commit
Per-edit step	sonnet-4-6	every edit	no	additive only	per edit, before review

Table 1: The generative stages, with the four properties that distinguish them. “Web” is access to a web-search tool. The version-controlled glossary is appended to every prompt in every stage. The first three stages covered the whole catalog; only the last precedes staff review.

this summary from the per-edit step.

Inline plain-language explanation boxes. An explanation box is inserted immediately after a Korea-specific institution, term, or procedure newcomers may not know, only where genuinely helpful and not already explained. In the frozen record, the boxes come from the catalog rewrite, which processed every article; the per-edit step adds them on later edits under a prompt that keeps all existing text, markdown, tables, and links unchanged and forbids restructuring, summarizing, or expanding the main text, so its boxes are additive and the article body is preserved. §5 verifies that the body survived in every recoverable case. Figure 1 shows one deployed article’s enhancement; the rendered reader view appears in Appendix K.

Policy-currency and link-freshness check (agentic). For dated guides the model is given a web-search tool and asked to check whether figures, amounts, deadlines, program names, responsible agencies, and URLs have changed, to refresh them, to replace broken links with current official pages, and to leave unverifiable facts unchanged.

Glossary injection (required). A version-controlled glossary of proper nouns with their standard Korean and English forms, including explicit corrections for known wrong expansions, is injected into every enhancement prompt with the instruction to use these exact names and never invent or vary an expansion; the requirement practitioners set in scoping (§3).

The per-edit step can be repeated safely, and the deployed prompts are in Appendix P. Generation adds no authoring cost, since it happens inside the pipeline that already translates the article, and API cost is small (Appendix I). The remaining cost is human review.

Human review and information loss. Simplification and factual editing both risk information loss, because a rewrite can drop a condition or blur

a legal distinction, the currency update can assert a wrong change, and a summary is by construction selective. The design bounds one of these mechanically, since the per-edit prompt preserves the article body verbatim, and the glossary fixes the entity errors the model is otherwise prone to. The rewrite, the summary, and the currency edits can genuinely lose meaning, and staff review exists to catch them.

The organization reviews every enhanced article, and the practice is mandatory. In the per-edit flow the review precedes release; the earlier stages deployed on commit, and §7.2 reports the exposure window that opened. The reviewer sees the enhanced Korean, and the translations inherit that reviewed source without being reviewed themselves, a bound we return to in the ethics section. The review costs a shift of staff time from writing to reading, and §8 measures the reading it creates and the repairs it produced.

5 Coverage, readability, and currency

Catalog coverage. The enhancement covers most of the catalog and concentrates on the high-stakes topics (Table 2). Across the 246 non-draft articles, 230 (93.5%) have at least one plain-language explanation box (722 boxes, mean 2.9 per article; Appendix H) and 167 (67.9%) have a summary. The gap is a timing artifact, since the boxes come from the catalog rewrite, which processed every article, while the summary came from a single stage, and all 79 articles without one entered the catalog after that stage had finished. In implementation terms, coverage is penetration, integration across the service’s units (Proctor et al., 2011).

Fidelity to each stage’s prescription holds. Body prose is unchanged in all 174 recoverable events where a summary first appears, but 149 of those are the summary stage, whose script rewrites the file with the body it already had, so preservation there is structural. Only the 25 events from the deployed per-edit step write back the body the model returned, and those are audited compliance with

the preserve-body instruction (Appendix C). The 722 boxes are individually written, with no exact duplicate.

Structural readability. Screening 238 recoverable originals leaves 83 clean pairs against the frozen enhanced versions (Appendix A). Because formula-based readability scores are poor targets for simplification (Tanprasert and Kauchak, 2021; Redish, 2000), all statistics come from the kiwipiepy morphological analyzer and its sentence splitter, computed on body prose with the added summaries and boxes stripped (Table 3; full statistics in Appendix B). The rewritten text is substantially less dense, with the article-mean of morphemes per sentence falling from 40.8 to 25.0, and the mean within-article change is -22.3% (bootstrap 95% CI -28.1 to -16.6 ; Wilcoxon $p < .001$), and sentences shorten in 70 of the 83 articles. The mechanism is sentence splitting by the catalog rewrite, since pooled body length is nearly unchanged ($\times 1.04$ in morphemes) across 1.41 times as many sentences, and the contrast is equally large inside and outside the later currency stage (-24% against -21%).

Vocabulary moves less than structure, with morphemes per word falling 7% and content-morpheme diversity is unchanged (MATTR 0.77 to 0.79, Covington and McFall, 2010), so the syntax was simplified and the Sino-Korean administrative vocabulary left in place. Explanation boxes are the design’s answer to vocabulary, the elaborative-simplification division of labor (Srikanth and Li, 2021). The contrast is negative in every topic (-17% to -39%), and the published article is about twice the length of its source (median $\times 1.9$ in characters), the increase being the added summary and boxes.

Policy-currency worked example and edit record. Agentic web search matters where a source has gone stale. In the housing-benefit guide of Figure 1 the income-eligibility threshold had been copied from a superseded circular; the update found it raised, corrected 47% to 48% of median income, and added the official income-cutoff table.

The record shows this was typical, since the update changed all 79 dated guides, altering figures or amounts in 75 and inserting or replacing 154 distinct URLs, none on a fabricated domain (Appendix D). Verifying ten inserted claims against official sources (Appendix G) found no wrong values, but four stated a wrong attribute (effective year,

statute article, frequency cap, change attribution), the residual class only source-checking review can catch (Rashkin et al., 2023).

6 Evaluation design

We evaluate the enhancement as a set of implementation outcomes (Proctor et al., 2011; Glasgow et al., 1999), assigning each empirical quantity to one construct: coverage is penetration; completing the work with contained failure modes is feasibility; conformance to each stage’s prescription is fidelity; the scoping of §3 is evidence of acceptability and appropriateness; and API spend and review time are cost. We reserve *fidelity* for that implementation construct throughout, and call the preservation of legal, procedural, and factual meaning *faithfulness*, which the Korean-language review tests. The readability contrast is the intervention’s proximal output, and reader comprehension, an effectiveness outcome, lies outside a production audit. Because the readers are a vulnerable population, the risks that most need evidence, lost or distorted meaning, invented entities, and leaked process text, get the failure audit and adjudicated review, while the strengths are audited automatically. Both automatic strands are reference-free, matching a setting with no reference rewrites and no in-house reviewers.

Automatic coverage, readability, and failure audit. Over the frozen record we count summary and explanation-box coverage (Table 2), compute the structural readability contrast (Table 3), and characterize the generation-specific failures: leaked process text, detected by a planning-phrase scan, and invented name expansions, checked against the glossary. The counts and scans are deterministic and reproducible from the corpus and its git history; §7.2 sets the keyword scan against an independent full reading of the stage’s output.

Korean-language faithfulness review. Because the enhancement is Korean-to-Korean, reviewers read both sides directly. The faithfulness review has three recorded parts. First, a flagging stage read each article against its pre-enhancement source and recorded candidate errors with verbatim quotes (changed legal or procedural meaning, changed or dropped numbers or deadlines, dropped steps, hallucinated additions, incorrect policy updates, leaked process text, summary errors), classified by the insertion, deletion, and substitution factuality taxonomy (Devaraj et al., 2022) with sever-

Topic	Art.	Summary	Any box	Box/art.	Comprehension stakes
Mental health	13	100%	100%	4.00	crisis content; boxes explain clinical terms
Asylum & status	53	91%	92%	3.47	legal register; procedures to follow exactly
Rights	76	87%	91%	3.46	rights and penalties; terminology-heavy
Health	44	86%	100%	3.45	safety-critical guidance read at low literacy
Work	20	85%	90%	3.60	wage claims; amounts must survive simplification
Daily living	41	93%	90%	3.56	practical how-to; gains most from plain structure
Education	32	69%	94%	2.81	enrollment, read by parents in a second language
Support/programs	65	62%	92%	2.58	deadlines and eligibility a summary must include
Advocacy	61	61%	89%	2.39	published verbatim by policy; summary only
Korean language	15	47%	100%	2.87	learning resources, already plain
All	246	67.9%	93.5%	2.93	

Table 2: Enhancement coverage and comprehension stakes by topic (frozen record, 246 non-draft articles; multi-label, so topic rows overlap). “Summary” is the share with a model-written summary; “Any box” the share with at least one explanation box. Advocacy articles have boxes because the catalog rewrite processed all articles uniformly; the per-edit advocacy exception (summary only) applies from its deployment onward.

Measure	Src	Enh	Paired Δ [95% CI]
Morphemes / sent.	40.8	25.0	−22% [−28, −17]
Words / sent.	18.7	12.2	−16% [−22, −10]
Characters / sent.	64.5	40.9	−18% [−24, −12]
Morphemes / word	2.21	2.05	−0.16 [−0.18, −0.14]
MATTR-50 (content)	0.77	0.79	+0.02 [+0.02, +0.03]
Pooled body: morphemes $\times$ 1.04, sentences $\times$ 1.41			

Table 3: Readability of source vs. enhanced Korean body prose (83 clean monolingual pairs; kiwipiepy analyzer and sentence splitter; boxes and summaries stripped). Src/Enh are means over articles; the paired column is the mean within-article change with a bootstrap 95% CI (absolute for the two lexical rows). Sentences shorten sharply while lexical measures barely move, so the simplification is structural.

ities taken from the critical-error categories of [Al Sharou and Specia \(2022\)](#). Second, an independent model verified every checkable claim against official sources ([Devaraj et al., 2021](#); [Min et al., 2023](#)) and gave a full-text second judgment on the flags the web could not settle. Third, the authors adjudicated every flag on this evidence; the full packet (flags, quotes, evidence URLs, verdicts) accompanies this paper (Appendix L).

Adjudication covered a 17-article sample, hereafter the tier, stratified by high-stakes topic crossed with currency-edit status, reported with exact binomial intervals. Because 17 articles bound little on their own, we then applied the same flagging stage, unchanged, to all 79 currency-edited guides (Appendix N).

Staff catalog review. A fourth record comes from the organization’s ordinary editorial work. One staff member read the entire Korean catalog and listed each article with a factual defect, the

passage, the reason, and a replacement; the director corrected the reviewer on one. This is not third-party review, since the reviewer is the operating organization’s staff and its director is a co-author. It is independent of the LLM pipeline, of the flagging model, and of our adjudication, and it is the only strand in which a domain expert read every article rather than a sample.

7 Failure record

7.1 Faithfulness and factual accuracy

Across the tier, the flagging stage raised 173 candidates; adjudication accepted 49 (28%): 5 critical, 25 major, 19 minor (18 hallucinated additions, 12 wrong updates, 11 distortions, 5 summary errors, 3 formatting). Sixteen of 17 articles contain a confirmed error (95% CI 71 to 100%), twelve a major or critical one (44 to 90%), three a critical one (4 to 43%). Thirty-three of the 49 fall in the 11 currency-edited articles.

We match each confirmed error’s verbatim quote to the earliest committed version containing it and assign it to the stage behind that commit (Appendix O); all 49 matched. Four of the five critical errors fall in currency-edited articles, and three of them were written by the currency stage itself. The other two were written by the catalog rewrite, which also held a web-search tool, so the claim the record supports is about tool access.

At scale the picture holds (Appendix N). Over all 79 currency-edited guides the same stage raised 716 verifiable candidates, and every one of the 79 has at least one; an official source contradicted 59 across 37 guides. On the 11 the tier also covers it raised 106 against 114 the first time. These are

candidates, not confirmed errors.

The staff catalog review has a denominator the audit cannot. Reading the whole live catalog, one staff member found 34 factual defects in 28 of 233 non-draft Korean articles (12.0%): stale figures, wrong contact numbers, conflated legal categories, and the same reassurances beyond the source that the adjudicated review found. Four concern safety or a deadline. One item confirms our own, since the reviewer independently found the 1345 helpline described as a 24-hour service, and added what our review had missed, that after 18:00 it answers only in Korean, English and Chinese. All 34 are corrected on the live service, with the same helpline error in 32 further articles.

In the costliest pattern a superseded scheme is presented as current. The education-visa guide, which the currency stage never touched, misstates the fine-relief mechanism and conditions for cancelling a child’s stay. Among added explanations, the model repeatedly reassures the reader beyond what the source supports. An explanation box asserts psychiatric records have no effect on visa screening, an update guarantees departing readers no disadvantage when the official measure only defers the re-entry ban, and five articles call the 1345 helpline a 24-hour service (it operates weekdays 09:00 to 22:00). All confirmed errors were corrected on the live service.

7.2 Generation-specific failures

Every failure below resulted from a stage that held a web-search tool. First, leaked process text: when the tool was rate-limited, the model wrote planning and search narration into the article instead of the finished text (Ji et al., 2023). An independent full reading of the committed output finds leaked text in 27 of the 79 guides (34%): search narration and tool-limit notices in 20 guides, and staff-facing change summaries in 7 guides. All 27 were published on commit and removed within two to four hours that same day; the frozen record and current site are clean.

Against this ground truth, the production keyword scan flags 8 of 27 (precision 8/8, recall 30%; Appendix E): narration phrased outside the anticipated wording, or reproduced in the translations, evades a fixed phrase list. Attribution is exact, because all 39 leak-bearing versions in the repository were written by one of the two stages that held a web-search tool, the catalog rewrite (31 articles) and the currency stage (8), while the tool-less

stages and per-edit step introduced none in any committed version, a zero whose upper bound is set by the size of the sample (Hanley and Lippman-Hand, 1983).

Second, invented expansions: the model produced plausible but wrong expansions for names it could not recall, including a dozen for the organization’s own name and one substitution of a different agency’s. The glossary now supplies standard forms with corrections for the known wrong ones; a conformance scan still finds 37 legacy non-conforming expansions (Appendix F).

8 Value and cost of agency

The stages of Table 1 differ in four properties at once, and a field deployment varies none of them deliberately, but three can be held fixed while the failures are read off the fourth. Every one of the 49 confirmed errors was written by the same model in the same execution context, so neither the model nor the execution context separates the stages that failed from the stages that did not; tool access does. Forty-three of the 49 errors, all five critical ones, all twelve wrong policy updates, and all eighteen hallucinated additions were written by a stage holding a web-search tool. The six errors from tool-less stages are five summary errors and one distorted title: none critical, none a wrong policy update, and all six in metadata rather than in the article body.

The record cannot break one confound. Every tool-using stage was also licensed to change the body and every tool-less stage could only add a field, so tool access and edit latitude move together.

Fact maintenance. Only an agentic stage can keep a fact true, the use search augmentation is built for (Vu et al., 2024). A benefit threshold copied from a superseded circular is not a readability problem, and no amount of simplification repairs it: a tool-less pipeline renders the stale number faithfully in seven languages, and only makes it easier to read. The currency stage changed all 79 dated guides against the notices that superseded them (§5). Re-checking dozens of government sources against a catalog accumulated over five years recurs with every policy cycle and competes for the same hours as frontline casework. That is the case for agentic LLMs in a public-interest service.

Novel failures. The cost has the same source. Every leaked passage and every critical error came from a stage that held a web-search tool. These

failures are not incidental to the agency; models produce them once they are licensed to consult the world and revise the text, and then meet a rate limit or settle on a conclusion their sources do not support, the failure generative search engines are measured for (Liu et al., 2023).

Transfer of labor. Agency converts an open-ended task, keeping figures, deadlines, and links current across dozens of government sources, into a bounded one, reading the enhanced Korean before publication. That bounded task is 503,922 characters of Korean, 47% of it in the 79 currency-edited guides. The scarce resource was never compute or money but the hours of Korean-reading attention that agency makes indispensable at the moment it is most useful, the human work around AI that is routinely undercounted (Sambasivan et al., 2021). Organizations adopting this design should budget the review hours and concentrate them where the tools act.

9 Lessons for resource-constrained organizations

First, simplify at the source, as the pre-editing tradition in crisis translation does (O’Brien and Federici, 2019). One Korean edit shortens sentences within articles by 22% (§5), every target language inherits that, and the quality check remains in the one language the team reads closely.

Second, route review by tool access. Holding tool-using output for review before release is the highest-yield rule this record supports.

Third, make preservation structural rather than instructed. Only the per-edit events tested the instruction; the other events could not change the body, because the script rewrote each file with the body it already held (Appendix C). Put the guarantee in the script.

Fourth, low-cost checks are lower-bound detectors. The keyword scan caught under a third of the leaks at perfect precision, and the residual currency errors were correct values with the wrong attribute, which no phrase list detects. A per-change log naming the source of each edit would make auditing deterministic (Gao et al., 2023; Rashkin et al., 2023).

Fifth, instrument the assurance. The repairs are timestamped, 17 commits rewriting 2,758 lines of corrected article content (Appendix J); the reading behind them left no record, so log it from the start (Mitchell et al., 2019).

Limitations

The primary limitation is that we measure implementation outcomes rather than effectiveness. We do not observe whether refugees in the seven target languages understood or acted on the enhanced information. The readability contrast licenses a narrow claim, that the enhanced Korean has shorter and more numerous sentences, with vocabulary almost unchanged. Whether shorter Korean sentences yield a more usable Arabic or Amharic page for someone with limited literacy in an administrative system they are new to is a different question, and neither morphological statistics nor formula-based scores answer it (Tanprasert and Kauchak, 2021). A reader-side comprehension study would answer it, and we do not have one. The organization has begun collecting feedback from refugee readers, and such a study would build on it; until then the outcome is untested rather than merely unreported.

The faithfulness review is a small-sample design, and no rater has double-coded the tier itself. The staff catalog review covers every article but a later state of the catalog, so the two cannot be pooled, and its reviewer belongs to the organization whose service is under audit. Its three stages are independent of one another only in part, and unevenly. Where official sources settled a flag, adjudication agreed with the web verdict on 85% of 114 flags (Cohen’s $\kappa = 0.70$), and every disagreement went one way, rejecting a flag the evidence supported and never accepting one it contradicted. Where the web could not settle a flag and the second model gave a full-text judgment, adjudication matches that model on all 80, so those two are not independent judgments; severity is the flagging stage’s own on 48 of the 49 confirmed errors. None of the three stages is a human judgment, and the flagging stage bounds what adjudication can see at all. The scaled flagging over the 79 currency-edited guides inherits that recall and has no adjudication behind it, so its rates are candidate rates. The keyword scan’s operating point rests on a single independent full reading of the stage’s output, the ten-claim verification of currency edits is a sample, and cost figures are estimates from corpus size and list prices rather than billing records. We do not test the effect of enhancement on downstream MT quality in the seven languages.

Practitioners were the design partners, and refugee end-users were consulted through the or-

ganization rather than as direct co-designers, a bounded form of participation (Delgado et al., 2023; Birhane et al., 2022). Finally, this is one organization, one source language, and one model family; we pin the models, prompts, and frozen evaluation commit so the design transfers as a template (Mitchell et al., 2019; Bender and Friedman, 2018), but the rates are specific to this pipeline.

Ethical and societal implications

This report uses an operational humanitarian workflow as research data and is co-authored with the organization that operates it. The frozen record is analyzed under the project’s ethics protocol under a two-part consent that separates program use from research use, and the adjudicated review is covered by the same protocol. Because the content concerns asylum, residence status, and access to services, we report errors at the category and rate levels and use only already-corrected examples, so that nothing we publish exposes current readers to a live error. We do not release the underlying dataset, which is refugee-facing material owned by the organization and shared under a data agreement for internal research. Publishing the Korean sources with their machine rewrites and error labels would create avoidable risks for an active service. The deployed prompts are provided in Appendix P, and we describe the checks in sufficient detail to be reimplemented; the corpus remains with the organization.

The distributive concern is direct, and it is the shape of the social-good case together with its limit. Enhancement is the part of the pipeline that most helps newcomers act on official information, and it is also the part whose failures, an invented policy update or a leaked instruction, readers unable to check the Korean source cannot detect. Benefit and exposure fall on the same people and are produced by the same component, so an account that reports only the intervention’s coverage or only its error rate misdescribes what was deployed. The organization controls where the human check goes. Placing it on the Korean source, before publication, keeps a single generated error from becoming seven translated ones, so the review is worth the staff time it consumes.

The reader-facing notice currently discloses less than the intervention. Each translated page states “AI-assisted translation from the Korean original. It may contain errors; please check the original for

anything important,” which covers translation but not the AI-written summary, the AI-generated explanations, or the AI-refreshed figures; extending the notice to AI-assisted editing with staff review is a small site change the organization controls. The safeguards also operate only in Korean, so an enhancement error originates once and propagates into seven languages, and residual risk therefore concentrates on readers of the lowest-resourced languages, who have the fewest alternative channels and for whom the organization has the least review capacity.

One component of the deployment falls outside that safeguard. The published article includes a reader-facing assistant that answers at query time, so the pre-publication review that governs the enhancement does not stand between it and the reader. Its guardrails are accordingly design-level, and it is grounded in the retrieved article, instructed to cite and to decline to invent facts, numbers, or URLs, and rate-limited per user. The organization retains no transcripts, a privacy decision that also puts the assistant’s field behavior outside the frozen record this audit analyzes. When we say that staff review is the safeguard the design relies on, the claim concerns source enhancement and is bounded by that scope. Whether the assistant tells readers the truth is a reader-side question, answerable only with instrumentation and a study design aimed at readers.

Practitioners now use general-purpose AI daily, often ahead of any policy or training (Johnson et al., 2025), and are accountable for output whose failure modes are invisible to them (Reamer, 2023; Garrett, 2025). Guidance written for well-resourced deployers compounds this (Yoo et al., 2026a). Naming the enhancement-specific failure modes, measuring them, and giving organizations the low-cost checks that contain them is a precondition for using generative simplification responsibly in refugee support. Building this system needed a partner with AI expertise; maintaining the review and checks did not, so another organization can adopt them without one.

References

- Anton Abilov, Ke Zhang, Hemank Lamba, Elizabeth M Olson, Joel Tetreault, and Alex Jaimes. 2025. *Operationalizing AI for good: Spotlight on deployment and integration of AI models in humanitarian work*. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 189–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

- Sweta Agrawal and Marine Carpuat. 2023. [Controlling pre-trained language models for grade-specific text simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Khetam Al Sharou and Lucia Specia. 2022. [A taxonomy and study of critical errors in machine translation](#). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 171–180, Ghent, Belgium. European Association for Machine Translation.
- Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Philipp Koehn, Vassilina Nikoulina, and 1 others. 2021. [Findings of the WMT shared task on machine translation using terminologies](#). In *Proceedings of the Sixth Conference on Machine Translation (WMT)*. Association for Computational Linguistics.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. [Data-driven sentence simplification: Survey and benchmark](#). *Computational Linguistics*, 46(1):135–187.
- Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. [The \(un\)suitability of automatic evaluation metrics for text simplification](#). *Computational Linguistics*, 47(4):861–889.
- Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. [TICO-19: the translation initiative for COvid-19](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. [Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing](#). *ACM Transactions on Computer-Human Interaction*, 30(5):1–38.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Nancy D. Berkman, Stacey L. Sheridan, Katrina E. Donahue, David J. Halpern, and Karen Crotty. 2011. [Low health literacy and health outcomes: An updated systematic review](#). *Annals of Internal Medicine*, 155(2):97–107.
- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. [Power to the people? opportunities and challenges for participatory AI](#). In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. Association for Computing Machinery.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. [Systematic inequalities in language technology performance across the world’s languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Deana Brown and Rebecca E. Grinter. 2016. [Designing for transient use: A human-in-the-loop translation platform for refugees](#). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 321–330. Association for Computing Machinery.
- Patrick Cadwell, Sharon O’Brien, and Eric DeLuca. 2019. [More than tweets: A critical reflection on developing and testing crisis machine translation technology](#). *Translation Spaces*, 8(2):300–333.
- Ned Cooper, Tiffanie Horne, Gillian R. Hayes, Courtney Heldreth, Michal Lahav, Jess Holbrook, and Lauren Wilcox. 2022. [A systematic review and thematic analysis of community-collaborative approaches to computing research](#). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, New York, NY, USA. Association for Computing Machinery.
- Michael A. Covington and Joe D. McFall. 2010. [Cutting the Gordian knot: The moving-average type–token ratio \(MATTR\)](#). *Journal of Quantitative Linguistics*, 17(2):94–100.
- David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. [Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50. Association for Computational Linguistics.
- Laura J Damschroder, Caitlin M Reardon, Marilla A Opra Widerquist, and Julie Lowery. 2022. [The updated consolidated framework for implementation research based on user feedback](#). *Implement. Sci.*, 17(1):75.
- Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. [The participatory turn in AI design: Theoretical foundations and the current state of practice](#). In *Proceedings of the 3rd ACM Conference on*

- Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*. Association for Computing Machinery.
- Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. [Paragraph-level simplification of medical texts](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.
- Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. 2022. [Evaluating factuality in text simplification](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Harshita Diddee, Kalika Bali, Monojit Choudhury, and Namrata Mukhija. 2022. [The six conundrums of building and deploying language technologies for social good](#). In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, COMPASS '22, pages 12–19, New York, NY, USA. ACM.
- Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. [Training neural machine translation to apply terminology constraints](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.
- Michael Färber, Parisa Aghdam, Kyuri Im, Mario Tawfeelis, and Hardik Ghoshal. 2025. [SimplifyMyText: An LLM-based system for inclusive plain language text simplification](#). In *Advances in Information Retrieval (ECIR 2025)*, Lecture Notes in Computer Science, pages 418–424. Springer.
- Glenn Flores, M Barton Laws, Sandra J Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J Hardt. 2003. [Errors in medical interpretation and their potential clinical consequences in pediatric encounters](#). *Pediatrics*, 111(1):6–14.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. [RARR: Researching and revising what language models say, using language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.
- Paul Michael Garrett. 2025. [‘magic moments’: AI and the ‘disappearance’ of social work ethics?](#) *Br. J. Soc. Work*, 56(1):377–395.
- Russell E. Glasgow, Thomas M. Vogt, and Shawn M. Boles. 1999. [Evaluating the public health impact of health promotion interventions: The RE-AIM framework](#). *American Journal of Public Health*, 89(9):1322–1327.
- Lauri Goldkind, Joy Ming, and Alex Fink. 2025. [AI in the nonprofit human services: Distinguishing between hype, harm, and hope](#). *Hum. Serv. Organ. Manag. Leadersh. Gov.*, 49(3):225–236.
- Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F T Martins. 2023. [Hallucinations in large multilingual translation models](#). *Trans. Assoc. Comput. Linguist.*, 11:1500–1517.
- James A. Hanley and Abby Lippman-Hand. 1983. [If nothing goes wrong, is everything all right? Interpreting zero numerators](#). *JAMA*, 249(13):1743–1745.
- Chris Hokamp and Qun Liu. 2017. [Lexically constrained decoding for sequence generation using grid beam search](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12).
- Madigan Johnson, Ka Man Parkinson, and Lucy Hall. 2025. [Initial insights report: How are humanitarians using artificial intelligence in 2025?](#)
- Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Li. 2023. [Multilingual simplification of medical texts](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Leah S Karliner, Elizabeth A Jacobs, Alice Hm Chen, and Sunita Mutha. 2007. [Do professional interpreters improve clinical care for patients with limited English proficiency? a systematic review of the literature](#). *Health Serv. Res.*, 42(2):727–754.
- Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. [BLESS: Benchmarking large language models on sentence simplification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

- Min Sun Kim, In Gyu Song, Ah Reum An, Kyae Hyung Kim, Ji Hoon Sohn, and Sei Won Yang. 2017. [Health-care access challenges facing six African refugee mothers in South Korea: a qualitative multiple-case study](#). *Korean J. Pediatr.*, 60(5):138–144.
- Wooksoo Kim, Isok Kim, Krisztina Baltimore, Ahmed Salman Imtiaz, Biplab Sudhin Bhattacharya, and Li Lin. 2020. [Simple contents and good readability: Improving health literacy for LEP populations](#). *Int. J. Med. Inform.*, 141:104230.
- Nicholas Kloehn, Gondy Leroy, David Kauchak, Yang Gu, Sonia Colina, Nicole P. Yuan, and Debra Revere. 2018. [Improving consumer understanding of medical text: Development and validation of a new SubSimplify algorithm to automatically generate term explanations in English and Spanish](#). *Journal of Medical Internet Research*, 20(8):e10779.
- Nanju Kwon and Jee Yeon Lee. 2022. [Information behavior of foreign refugees staying in South Korea](#). *J. Korean Soc. Inf. Manag.*, 39(1):351–374.
- Jieun Lee. 2021. [Supporting the social integration of migrant women in south korea through language services: Roles of marriage migrant interpreters in multicultural family support centre counselling services](#). *Translator*, 27(1):57–74.
- Gondy Leroy, James E. Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. [User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention](#). *Journal of Medical Internet Research*, 15(7):e144.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 9459–9474.
- William Lewis. 2010. [Haitian Creole: How to build and ship an MT engine from scratch in 4 days, 17 hours, & 30 minutes](#). In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, Saint Raphaël, France. European Association for Machine Translation.
- Hongjin Lin, Naveena Karusala, Chinasa T Okolo, Catherine D’Ignazio, and Krzysztof Z Gajos. 2024. [“come to us first”: Centering community organizations in artificial intelligence for social good partnerships](#). *Proc. ACM Hum. Comput. Interact.*, 8(CSCW2):1–28.
- Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. [Evaluating verifiability in generative search engines](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.
- Nowshin Mannan, Afsana Monir, Shivani Mishra, Sarah Li, Nari Yoo, and Naheed Ahmed. 2025. [Community needs assessment for understanding the health and social needs for South Asian immigrant populations in Southeast Queens County, New York City, across levels of English proficiency](#). *J. Community Health*, 51(2):284–295.
- Laura Manor and Junyi Jessy Li. 2019. [Plain English summarization of contracts](#). In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics.
- Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe. 2022. [Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health](#). *Proc. Conf. AAAI Artif. Intell.*, 36(11):12017–12025.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919. Association for Computational Linguistics.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. [Augmented language models: a survey](#). *Transactions on Machine Learning Research*.
- Sewon Min, Kalpesh Krishna, Xinxin Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. [Model cards for model reporting](#). In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. Association for Computing Machinery.
- Hana Moon, Geon Ho Lee, and Yoon Jeong Cho. 2022. [Readability of Korean-language COVID-19 information from the South Korean national COVID-19 portal intended for the general public: Cross-sectional infodemiology study](#). *JMIR Formative Research*, 6(3):e30085.
- Joanna C Moullin, Kelsey S Dickson, Nicole A Stadnick, Borsika Rabin, and Gregory A Aarons. 2019. [Systematic review of the exploration, preparation, implementation, sustainment \(EPIS\) framework](#). *Implement. Sci.*, 14(1):1.

- Rahul Nair, Bo Madsen, and Alexander Kjærsum. 2023. [An explainable forecasting system for humanitarian needs assessment](#). *Proc. Conf. AAAI Artif. Intell.*, 37(13):15569–15575.
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, and 1 others. 2021. [WebGPT: Browser-assisted question-answering with human feedback](#). *Preprint*, arXiv:2112.09332.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, Solomon Oluwale Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. [Participatory research for low-resourced machine translation: A case study in African languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sharon O’Brien and Federico Marco Federici. 2019. [Crisis translation: considering language needs in multilingual disaster settings](#). *Disaster Prev. Manag.*, 29(2):129–143.
- OHCHR. 2025. [US government fuelling global humanitarian catastrophe: UN experts](#).
- Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, and Grace Tang. 2020. [Gamayun – language technology for humanitarian response](#). In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, Seattle, WA, USA. IEEE.
- Matt Post and David Vilar. 2018. [Fast lexically constrained decoding with dynamic beam allocation for neural machine translation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.
- Enola Proctor, Hiie Silmere, Ramesh Raghavan, Peter Hovmand, Greg Aarons, Alicia Bunger, Richard Grifey, and Melissa Hensley. 2011. [Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda](#). *Adm. Policy Ment. Health*, 38(2):65–76.
- Enola K Proctor, Alicia C Bunger, Rebecca Lengnick-Hall, Donald R Gerke, Jared K Martin, Rebecca J Phillips, and Julia C Swanson. 2023. [Ten years of implementation outcomes research: a scoping review](#). *Implement. Sci.*, 18(1):31.
- Nasreen S Quadri, Syreeta Wilkins, Kristina Krohn, Erin M Mann, William M Stauffer, and Patricia F Walker. 2023. [Language justice: Addressing linguistic disparities begins with language data collection](#). *Am. J. Trop. Med. Hyg.*, 109(1):1–3.
- Pragnya Ramjee, Mehak Chhokar, Bhuvan Sachdeva, Mahendra Meena, Hamid Abdullah, Aditya Vashistha, Ruchit Nagar, and Mohit Jain. 2025. [ASHABot: An LLM-powered chatbot to support the informational needs of community health workers](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, pages 1–22, New York, NY, USA. ACM.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. [Measuring attribution in natural language generation models](#). *Computational Linguistics*, 49(4):777–840.
- Frederic G Reamer. 2023. [Artificial intelligence in social work: Emerging ethical issues](#). *Int J SW Values and Ethics*, 20(2):52–71.
- Janice Redish. 2000. [Readability formulas have even more limitations than Klare discusses](#). *ACM Journal of Computer Documentation*, 24(3):132–137.
- Nathaniel Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. [ChatGPT MT: Competitive for high- \(but not low-\) resource languages](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. [“everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.
- Advait Siddharthan. 2014. [A survey of research on text simplification](#). *ITL – International Journal of Applied Linguistics*, 165(2):259–298.
- Neha Srikanth and Junyi Jessy Li. 2021. [Elaborative simplification: Content addition and explanation generation in text simplification](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics.
- Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. [DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

- Elior Sulem, Omri Abend, and Ari Rappoport. 2018. [BLEU is not suitable for the evaluation of text simplification](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.
- Teerapaun Tanprasert and David Kauchak. 2021. [Flesch-Kincaid is not a text simplification evaluation metric](#). In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Association for Computational Linguistics.
- M Todorova and Rachel Liu. 2024. [The use of MT by humanitarian NGOs in Hong Kong](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 51–52, Sheffield, UK. European Association for Machine Translation (EAMT).
- Katy E Trinkley, Ruopeng An, Anna M Maw, Russell E Glasgow, and Ross C Brownson. 2024. [Leveraging artificial intelligence to advance implementation science: potential opportunities and cautions](#). *Implementation Sci.*, 19(1):17.
- UNHCR. 2025. [On the brink: The devastating toll of aid cuts on people forced to flee](#).
- S Valdez, Ana Guerberof Arenas, and Kars Ligtenberg. 2023. [Migrant communities living in the Netherlands and their use of MT in healthcare settings](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 325–334, Tampere, Finland. European Association for Machine Translation.
- Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. [Fresh-LLMs: Refreshing large language models with search engine augmentation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.
- Melissa Wall, Madeline Otis Campbell, and Dana Janbek. 2017. [Syrian refugees and information precarity](#). *New Media & Society*, 19(2):240–254.
- Andy Way, Rejwanul Haque, Guodong Xie, Federico Gaspari, Maja Popović, and Alberto Poncelas. 2020. [Rapid development of competitive translation engines for access to multilingual COVID-19 information](#). *Informatics*, 7(2):19.
- World Health Organization. 2008. [Closing the gap in a generation: Health equity through action on the social determinants of health: Final report of the commission on social determinants of health](#). World Health Organization, Genève, Switzerland.
- Maximilian Wutz, Marius Hermes, Vera Winter, and Juliane Köberlein-Neu. 2023. [Factors influencing the acceptability, acceptance, and adoption of conversational agents in health care: Integrative review](#). *J. Med. Internet Res.*, 25:e46548.
- Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.
- Achille Roghemrazangba Yameogo, Carole Délétroz, Maxime Sasseville, Samira Amil, Sié Mathieu Aymar Romaric Da, James Plaisimond, Frédéric Bergeron, Sofia Tadlaoui, Patrick Bodenmann, and Marie-Pierre Gagnon. 2025. [Effectiveness of interventions to improve digital health literacy in forced migrant populations: Mixed methods systematic review](#). *Journal of Medical Internet Research*, 27:e69880.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations (ICLR 2023)*.
- Nari Yoo, Ashley Khor, Namrata Mukhija, Aminat Adebiyi, and Miri Zilka. 2026a. [Guidelines for whom? rethinking AI ethics in resource-constrained migration services](#). In *Proceedings of the Workshop on Evaluating Evaluations (EvalEval)*, pages 19–25, San Diego, CA. Association for Computational Linguistics.
- Nari Yoo, Jihye Lee, and Sou Hyun Jang. 2026b. [Beyond the digital divide: e-government services and subjective well-being among North Korean refugees and native-born Koreans during COVID-19](#). *Journal of Ethnic and Migration Studies*, pages 1–25.

A Pair selection funnel (readability corpus)

Of the recoverable pre-enhancement originals, we exclude already-enhanced copies, multilingual or contaminated scrapes, and very short notices, leaving 83 clean monolingual pairs (Table A1). The excluded-but-matched articles skew toward support, education, and Korean-language notices.

Stage	<i>n</i>
Recoverable pre-enhancement originals	238
Excluded: stored original already enhanced	11
Excluded: multilingual or contaminated scrape (under 80% Korean characters)	133
Excluded: body under 300 characters	11
Clean monolingual pairs analyzed	83

Table A1: Readability-corpus selection funnel.

B Readability, paired statistics

On the 83 clean pairs (kiwipiepy analyzer, body prose with summaries and boxes stripped), Wilcoxon signed-rank on morphemes per sentence gives $W = 275$, $p < .001$, Cohen’s $d_z = -0.42$ (Table A2). Pooled (micro) morphemes per sentence $33.2 \rightarrow 24.4$; sentence count pooled $\times 1.41$; body morphemes pooled $\times 1.04$ (per-pair mean $\times 1.38$, median $\times 1.24$, a small-article skew). Per-topic mean percent change (topics with $n \geq 5$): rights -19.5 , asylum -30.0 , advocacy -16.9 , health -31.2 , support -24.3 , daily living -21.3 , work -38.8 , education -30.1 , mental health -22.5 .

Measure	Src	Enh	Paired Δ [95% CI]
Morphemes / sent.	40.77	24.95	-22.3% [$-28.1, -16.6$]
Words / sent.	18.69	12.17	-15.9% [$-22.1, -9.9$]
Chars / sent.	64.47	40.87	-18.2% [$-24.3, -12.0$]
Morphemes / word	2.210	2.051	-0.160 [$-0.182, -0.136$]
Content morph./s.	18.6	11.5	-7.19 [$-12.0, -3.5$]
Content share	0.436	0.458	$+0.023$ [$+0.016, +0.030$]
MATTR-50	0.772	0.794	$+0.023$ [$+0.015, +0.030$]

Table A2: Readability of source vs. enhanced Korean body prose (83 pairs). The paired column is the mean within-article change with a bootstrap 95% CI (absolute for the two lexical rows). Content rows and MATTR-50 are computed on content morphemes.

C Stage decomposition (git-isolated)

Attributing the readability effect to a stage uses the site’s commit history (Table A3). The catalog rewrite produces the contrast; the stages that only add a summary leave the body unchanged.

Contrast	<i>n</i>	Morph./sent.
Catalog rewrite: raw export to first commit	83	$40.8 \rightarrow 24.8$
Summary added, clean-pair subset	72	$\pm 0.0\%$
Summary added, all events (149 summary stage, 25 per-edit)	174	$\pm 0.0\%$

Table A3: Stage decomposition. The catalog rewrite gives the readability contrast (paired -22.7% [$-28.7, -16.5$]) with containment averaging 1% (all $\leq 50\%$); the summary-adding events leave morphemes per sentence unchanged, with containment averaging 100% ($\geq 90\%$ for every event). Containment is the share of the pre-enhancement sentences that appear verbatim in the post-enhancement text, whitespace removed, over sentences of at least ten characters.

D Currency-stage edit typology and URL audit

The currency stage changed all 79 dated guides. Articles with: numeric or amount changes 75; year-reference updates 79; URL insertions or replacements 70 (154 distinct URLs); notes to confirm with the agency 69; step lists 69. URL liveness: 123 of 151 resolve (81%); 6 (4%) are dead, two education-office and four disease-control deep links; 22 could not be checked automatically but open in a browser; fabricated domains: 0.

E Leaked-process-text audit and scan validation

An independent full reading of all 79 committed outputs (three blind annotators, disjoint thirds) finds leaked process text in 27/79 (34.2%): search narration or tool-limit notices or editing promises in 20, staff-directed change summaries in 7. Against this ground truth the production keyword scan flags 8: precision 8/8 (100%), recall 8/27 (29.6%). The marker-detectable leaks were cleaned within 1.7 to 2.7 hours, and all 27 passages are absent from the freeze taken 4 hours later and from the current site. Attributing leak markers over every committed version of all 256 Korean articles gives 39 introductions, all of them by one of the two stages holding a web-search tool: the catalog rewrite (31 articles) and the currency stage (8). The per-edit step introduced none, and neither did any tool-less stage.

F Glossary conformance scan

Nonconforming “Expansion (ACRONYM)” instances against the glossary’s standard forms fall from 17.2% at the freeze commit (89/516) to 5.0% at the current site (68/1371). The organization’s own name was wrong in 56/248 instances at the freeze and 37/238 at the current site (for example “Migration for Asia Peace”, “Migrants for Asian Peace”). Strict exact matching also flags benign variants (the official UNHCR long form; an IDP sense collision), so all rates are upper bounds.

G Fact verification of currency edits (sample)

Ten discrete inserted claims were verified against official sources by two independent web-research agents: 6 fully correct, 4 partially correct (correct value, wrong or unsupported attribute), 0 wrong values. This sample motivated the fuller adjudicated review of Appendix L.

H Summary and box characteristics

Character counts include spaces and exclude HTML markup and the box’s own “in plain terms” label. The 167 summaries average 125 characters in one to two sentences, and are denser per sentence than the prose they head, 31.1 against 22.7 morphemes per sentence, which single-sentence compression produces. The 722 explanation boxes average 146 characters (median 137, p10 95, p90 209), with no exact duplicate among them.

I Cost estimate

Costs are estimated from corpus size and 2026-07 list prices (assumed token density 0.5–0.8 tokens/char for the Korean tokenizer, $\times 1.3$ for the larger model; a 50% discount; web search \$10 per 1,000 searches; 2,500 input tokens per search result). Catalog rewrite \$25–52; summary and per-edit steps \$13–20 ($\approx \$0.05$ – 0.08 per article marginal); policy currency \$7–9; total $\approx \$45$ – 80 .

J Review burden and repair record

API spend is the small half of the cost. The other half is reading, and the record measures how much there is to read but not how long it took.

The enhanced Korean a reviewer has to read is 503,922 characters across the 246 non-draft articles, a median of 2,016 per article (mean 2,048, p90 3,679). The 79 currency-edited guides account

for 236,464 of those characters, 47% of the catalog over 32% of the articles, so a review concentrated where the tools acted is a third of the articles and about half the text. The tool-less summary stage added 167 fields averaging 125 characters, 3.3% of the reading surface in total.

Review time is not in the history, because enhancement, translation, and review land in one commit. The repair is, because each correction after a generative stage is its own commit. Counting commits that name a content defect and change body prose rather than metadata alone: 17 such commits, 3 of them inside the frozen record and 14 after it, touching 273 Korean article-versions and rewriting 2,758 added and 3,655 deleted lines, with 113 of those article-edits on the 79 currency-edited guides and 32 on the 17 audited articles. The largest is the reversion and repair of 121 articles committed 1 h 42 m after the currency stage; the second rolls 16 advocacy statements back to their unenhanced text; and one of the 17, five weeks later, is the correction of the errors this paper’s own review confirmed. These are the visible half of the safeguard: the repairs it produced, and not the reading that found nothing to repair.

K Reader view

Figure A1 shows the enhancement as readers see it, in the English rendering of the housing-benefit guide (the same article used for the worked example in the main paper). The enhancement is authored on the Korean source and propagates to all seven target languages, including right-to-left Arabic. ❶ the AI-written plain-language summary with text-to-speech; ❷ an honesty notice that the text is AI-translated from Korean; ❸ an inline plain-language explanation of a Korea-specific administrative term. The policy-currency correction discussed in the main paper appears further down the same article.

Housing Benefit Application Eligibility and Process (Foreigners and Refugees) Guide

Published June 21, 2023 · Last updated June 19, 2026 · 27 min read

On this page

Edit

Suggest a translation fix

Listen

AI SUMMARY

Foreigners and refugees with low incomes can apply for **housing benefits**, which provide monthly rent support if they meet the required conditions. An alien registration card is required, and applications can be submitted at the nearest community center or through Bokjoro (bokjoro.go.kr).

AI-assisted translation from the Korean original. It may contain errors; please check the original for anything important.

On this page

The Housing Benefit is a government support program that either subsidizes monthly rent or repairs owner-occupied homes for households whose income falls below a set threshold. It is based on the Housing Benefit Act, and eligible foreign nationals and refugees can apply. You can apply in person at the Eup/Myeon/Dong Community Service Center (주민센터) that covers your registered address, or apply online at Bokjoro (bokjoro.go.kr).

NOTE

An "Eup/Myeon/Dong Community Service Center" (also called a resident center or 행정복지센터) is a local government office at the neighborhood or village level, the smallest unit of local administration in Korea. This is where most everyday welfare applications are processed.

Figure A1: The enhancement as readers see it (English rendering of the housing-benefit guide).

L Adjudicated faithfulness review packet

The three-stage review of §6 and §7.1 is: (i) an LLM flagging stage over 17 stratified articles that recorded 173 candidate errors with verbatim quotes; (ii) an independent model that web-verified every checkable claim against official sources and gave a full-text second judgment on the rest; and (iii) author adjudication of every flag (accept 49, reject 124; acceptance 28.3%). Accepted errors: 5 critical, 25 major, 19 minor; 16/17 articles have at least one (exact 95% CI 71–100%), 12/17 a major or critical one (44–90%), 3/17 a critical one (4–43%); 33 of the 49 fall in the 11 currency-edited articles, and four of the five critical errors do. Intervals are Clopper-Pearson exact binomial. The accompanying packet file lists every flag with its verbatim quote, category, severity, taxonomy label, web-verification verdict and evidence URL, and adjudication verdict, so each reported rate is traceable to the underlying judgment.

Agreement between the stages, computed from that packet: on the 114 flags official sources settled, the web verdict and the adjudication agree on 85.1% ($\kappa = 0.70$), and all 17 disagreements are flags the evidence supported and adjudication rejected. On the 80 flags the web could not settle, the second model's full-text judgment and the adjudication agree on all 80 ($\kappa = 1.00$), so those are not

independent judgments. The flagging stage's own confidence is calibrated against the outcome (accepted at 44.3% of its high-confidence flags, 22.9% medium, 6.1% low), and its severity label survived adjudication on 48 of the 49 confirmed errors.

M Practitioner-identified needs and matching features

The feature set of §4 was scoped against the needs practitioners raised in the working discussions of §3 (Table A4).

Practitioner-identified need	Enhancement feature
Readers cannot act on a faithful translation of dense administrative prose	Plain-language summary at the top of each article
Readers repeatedly ask what a Korea-specific institution or term means	Inline "in plain terms" explanation boxes
Figures, deadlines, agency names, and links fall out of date	Web-search policy-currency check (agentic)
Organization and program names are rendered inconsistently or wrongly	Version-controlled glossary injected into every prompt

Table A4: Feature scope, with each function matched to the practitioner-identified need raised in scoping meetings.

N Flagging at scale

The 17-article tier is small, so we repeated its flagging stage over all 79 currency-edited guides, under the same model, prompt, tool budget and inputs. Adjudication was not repeated, so these are candidate rates.

A recall-oriented flagger sometimes quotes text that is not in the article it is judging, so every quote is checked against the frozen article before it is counted: 716 of 816 flags at scale (87.7%) and 166 of 173 in the tier (96.0%), with 48 of the tier's 49 confirmed errors among its verified ones.

Over the 79 guides: 716 candidates, mean 9.1 and median 9 per guide (range 2 to 17), and 79 of 79 guides have at least one (95% Wilson 95.4 to 100%). As flagged, before any adjudication, 35 are critical, 161 major and 520 minor; by category, 347 hallucinated additions, 100 wrong updates, 64 number or amount changes, 58 formatting, 57 distortions, 56 summary errors, 27 dropped steps and 7 leaked passages. The flagger attempted web verification on 233 and an official source contradicted 59 of them, across 37 guides.

Eleven of the 79 are in the adjudicated tier, so the same articles were flagged twice, months apart, under the same protocol: 106 verified candidates the second time, 114 the first, a ratio of 0.93. The comparison bounds the stability of the rate the tier’s numbers rest on.

Projecting confirmed errors is possible but is a projection. The tier’s acceptance rates conditional on the flagger’s own confidence are 46.2% (high), 23.2% (medium) and 6.2% (low); applied to the confidence mix at scale they give roughly 207 confirmed errors over the 79 guides, or 2.6 per guide, against the 3.0 per guide observed in the tier’s 11 currency-edited articles. The assumption that those acceptance rates transfer is exactly what a larger adjudication would test and this one does not.

O Origin of each confirmed error

Every version of every Korean article is committed, so each confirmed error can be assigned to the stage that wrote it. We normalize the error’s verbatim quote and the article text to Korean characters and digits, dropping HTML markup, then take the earliest committed version containing the quote and map its commit to a stage using the generating scripts and their logs. All 49 matched, each present in the version its stage produced and absent from the version before it (Table A5).

Stage	Web	n	Crit	Maj	Min
Catalog rewrite	yes	34	2	17	15
Policy currency	yes	9	3	4	2
Summary	no	5	0	4	1
Title	no	1	0	0	1
All		49	5	25	19

Table A5: The stage that wrote each of the 49 confirmed errors. “Web” is whether that stage held a web-search tool. Every error came from the same model in the same execution context, so model and execution context do not separate the stages; tool access does. By category, the tool-using stages wrote all 12 wrong policy updates, all 18 hallucinated additions, 10 of the 11 distortions and all 3 formatting errors, and the tool-less stages wrote all 5 summary errors and 1 distorted title.

P Enhancement prompts

The deployed system prompts, lightly wrapped. The per-edit prompt (summary and explanation boxes) is issued in English; the catalog-rewrite and policy-currency prompts are issued in Korean and given here in English translation.

Catalog-modernization rewrite prompt (from Korean).

You are the editor of an encyclopedia-style wiki for refugees and migrants living in Korea. Rewrite the given source article as an accurate, easy-to-understand wiki document.

Style: consistent polite register, warm and friendly, as if speaking kindly to the reader; natural Korean; no em dashes, no AI-cliche phrasing, no exaggeration or filler.

Cleanup: silently drop internal management tags, CSS/HTML artifacts, images, and editing traces; never write about what was removed or changed.

Links: drop broken or unverifiable external file links but keep the institution name; use web_search to find the institution’s current official page and link it inline.

Reinforce: actively verify and update programs, agencies, and procedures via web_search; do not invent facts.

Structure: a one-paragraph definition and summary first, then logical sections with headings and lists.

Since figures, thresholds, and deadlines change with year and policy, add a note to confirm the latest information with 1345 or the competent immigration office.

Start directly with the ‘# title’ line and output Korean Markdown only, no preamble.

Summary and explanation-box prompt.

You are enhancing a Korean wiki article for refugees and migrants living in Korea.

Return TWO parts, split by the markers below.

- 1) A summary: 1-2 plain Korean sentences with the key takeaway and what to do.
- 2) The article body, KEEPING all existing text, markdown, tables and links unchanged, but adding a SIMPLE plain-language explanation box right after any Korea-specific institution, term, or procedure a newcomer might not know (only where genuinely helpful and not already explained). Do not restructure, summarize, or expand the main text; do not tailor to any nationality. No em dashes, no emojis.

[a version-controlled glossary of standard Korean/English names is appended here]

Policy-currency prompt (from Korean).

You are an editor of a wiki for refugees and migrants in Korea. Update the article below to a 2026 basis. Do not rewrite from scratch; keep the structure, tone, and title, and only revise.

- 1) Use web_search to check whether figures, amounts, deadlines, program names, responsible agencies, and URLs have changed, and correct them to current information; replace broken links with current official URLs. Do not change facts you cannot verify.
- 2) Strengthen inline links to official government pages.
- 3) Add relevant internal links.
- 4) After figures/deadlines, add a note to confirm the latest information with 1345 or the local immigration office.

Forbidden: em dashes, emojis, meta commentary about what you changed.