Mitigating the "Triple Burden": Grade-Level Translation of Science Texts Using Multi-Agent Systems and Prompt Optimization

Joel Walsh[1,†], Oscar Arenas[2,†], Ashley Marquez[3], Karina Méndez Pérez[4] and Benjamin Nye[5]

[1]*Occidental College, United States*
[2]*California State University, Long Beach, United States*
[3]*Tanque Verde Unified School District, United States*
[4]*St. Elmo Elementary, United States*
[5]*University of Southern California Institute for Creative Technologies, United States*

Abstract

Immigrant students around the world face the "Double Burden", the challenge of learning a new language and domain-specific content simultaneously. In the United States many Spanish-speaking learners face barriers to understanding academic content from textbooks while simultaneously learning English. Existing machine translation benchmarks and engineering approaches have led to models that can competently produce target-language outputs that often default to the same grade level as the source (e.g., 10th-grade English translates to 10th-grade Spanish). This can impose a "Triple Burden" for students who are not yet reading at grade level in their heritage language or academic first language. While simplification via Large Language Models (LLMs) is a potential solution, it often compromises curriculum alignment. This study evaluates the effectiveness of three large language models (Aya, Gemini 2.5 Pro, and GPT-4o) at producing grade-appropriate Spanish translation of English science textbook excerpts, with and without an agent-based optimization scheme. While agent-based refinement nudges grade-level alignment closer to the target grade level for all models, human evaluations reveal that the quality of the translations can vary. These human labels and feedback were then used for Genetic Pareto (GEPA)-based prompt optimization, a method that creates highly tuned prompts that outperform post-training methods like Supervised Finetuning or Reinforcement Learning. Our initial human evaluation of the optimized translation prompt showed promise, but further testing is needed. This pipeline can serve as a model for combining agent-based translations with human feedback, creating improvements that are portable across models in ways that other post-training methods are not.

Keywords

Machine Translation, Multilingual Education, Large Language Models, Prompt Optimization, Science Education, Multi-Agent Systems

1. Introduction

In the United States, 25 states reported shortages of bilingual or English-as-a-Second-Language teachers in the 2023–24 school year, per federal Teacher Shortage Area filings [1], and bilingual/ESL positions were among the most difficult to fill across school levels in 2024–25 [2]. While multilingual LLMs are not substitutes for multilingual teachers, these shortages imply that there is great demand for robust machine translation (MT) tools in school settings. This development is seemingly not lost on textbook manufacturers such as McGraw Hill, who have recently (April 2026) built MT tools into their McGraw Hill Connect online learning platform for higher education [3].

While such textbook-based MT tools are likely helpful for some students, both our conversations with U.S.-based teachers and our own exploratory research surfaced a recurring limitation. LLM translations are generally accurate, but they tend to mirror the readability level of the source text rather than adjust

iTextbooks'26: Seventh Workshop on Intelligent Textbooks, June 28, 2026, Seoul, Republic of Korea
[†]These authors contributed equally.
jwalsh2@oxy.edu (J. Walsh); oscararenas625@gmail.com (O. Arenas); ashleymarquezteach@gmail.com (A. Marquez); Karina.mendez4413@gmail.com (K. M. Pérez); nye@ict.usc.edu (B. Nye)
0000-0002-7013-1888 (J. Walsh); 0000-0003-2818-3390 (K. M. Pérez); 0000-0002-5902-9196 (B. Nye)


it to a target grade level. This is a poor fit for many U.S. classrooms, where "Newcomer" students have often missed significant schooling in their home countries or intermediate host countries [4] and may therefore read below grade level even in their heritage language.

Educators in multilingual settings often require a higher level of granularity and content preservation for LLM-based translations. When educators require a translation of a text to a target language, they must be able to specify a numeric grade level of a target language (e.g., "Take this passage and translate it into Spanish at a 6th grade reading level") while preserving the content standards of the original text.

To this end, we sought to answer three research questions related to this problem: **(RQ1)** Can a multi-agent system using text complexity metrics nudge textbook translations towards a target grade better than zero-shot prompting? **(RQ2)** How do human subject matter experts rate these agent-based translations, using both open and closed multilingual LLMs? **(RQ3)** Can Genetic Pareto (GEPA) prompt optimization using human ratings create prompts that improve local model performance?

We first sought to test whether or not the LLMs could meet a target grade level translation through zero-shot prompting, which they could not. We then constructed a multi-agent system that continuously measures and retries translation until the result registers at a specific grade-level readability index. We tested this agent-based system with open and closed LLMs using automated metrics and human annotation. Lastly, we used human annotations and feedback to optimize a prompt for this particular type of translation. The technique we chose was Genetic Pareto (GEPA), which leverages self-reflection reasoning chains and gold standard examples. GEPA prompt optimization has been shown to surpass Supervised Fine Tuning at targeted tasks [5], without requiring expensive training regimes or compute. Between this agent-based method and prompt optimization, we demonstrate a pathway for deploying translation systems on local hardware, reducing the privacy and cost barriers that currently limit the scale of AI adoption in schools.

2. Related Work

Before the advent of neural models for Machine Translation (MT), statistical machine translation [6] and rule-based systems [7] dominated the field. In the mid-2010s, neural attention [8] and LSTM-based models [9] began to dominate the academic and production landscape. As neural machine translation (NMT) tools pushed into educational use cases, there was considerable debate over the quality of NMT translations [10], and whether or not the tools of that era could be trusted to translate domain specific texts in educational settings [11].

Eventually, Large Language Models began to show promise for translation [12], and have now met or exceeded state of the art machine translation systems in accuracy [13]. Commonly recognized metrics for these translation tasks include the word-overlap-based BLEU [14] or the more recent COMET [15], which leverages cross-lingual neural embeddings to assess semantic similarity between translations and gold-standard references.

Makers of large multilingual models often report translation performance using BLEU and COMET, but a system that succeeds according to these metrics does not necessarily succeed in K–12 settings. Educational translations can have multiple valid translations or pedagogical framings for the same source text depending on target grade level. For this reason, Spanish readability metrics are a passable automated metric to assess the grade-level readability of machine translated passages. Although many metrics use similar features [16, 17] such as the average number of syllables per word or average word length, the Fernández–Huerta (FH) index [18] also aligns with U.S. grade levels. This allows for some standardization and alignment between prompts and datasets.

This approach also builds on Text Simplification literature [19], as the task of grade-level translation also involves simplifying the target language output while still retaining meaning of the original passage. Language-specific datasets are needed to finetune and evaluate performance with multilingual simplification [20]. In recent years, researchers have developed specific datasets for Spanish simplification [21].

As agent-based workflows have begun to infiltrate both enterprise software and academic research,

several new agent-based approaches to translation have appeared. These approaches include using agents to: simulate a translation agency [22], individually evaluate each domain of the Multidimensional Quality Metrics [23], debate a translation's quality [24], and iteratively call Google Translate [25].

Our work takes inspiration from agent-based approaches to translation, with the added wrinkle of translating English to a specific grade level of Spanish while preserving the standards and spirit of the original document. We also diverge from the existing literature by using human-annotated data and Genetic Pareto prompt optimization (GEPA) [5] to form specialized prompts that address this particular task. This method is a promising alternative to Supervised Finetuning (SFT), and requires no model weight updating or serving. This method holds promise for learning designers, providing a means to effectively leverage the rich feedback of domain experts in a cost-effective manner.

3. Methodology

3.1. Readability Metric

To evaluate whether Spanish translations matched their intended grade levels, we used the Fernández–Huerta (FH) readability formula [18], a Spanish adaptation of the Flesch Reading Ease score [26]. The formula calculates readability as:

$$\text{FH} = 206.84 - 60 \times \frac{\text{syllables}}{\text{words}} - 102 \times \frac{\text{sentences}}{\text{words}} \tag{1}$$

where higher scores indicate easier text. Table 1 shows how FH index maps to U.S. grade levels and corresponding age ranges. We defined a translation as "successful" when its FH index matches the target age.

Table 1
Fernández–Huerta readability scale

FH Index	Grade Level	Age Range
>100	2–3	7–8
90–100	4th	9–10
80–89	5th	10–11
70–79	6th	11–12
60–69	7–8	12–14
50–59	9–10	14–16
30–49	11–12	16–18
<30	Coll.	18+

Our dataset initially consisted of 715 English science passages drawn from openly licensed textbooks, including OpenStax [27] and CK-12 Foundation [28], as well as supplementary materials shared by practicing K–12 teachers. The passages spanned grades 3 through high school and covered all four Next Generation Science Standards (NGSS) [29] aligned science domains: Life Science ($n = 245$), Physical Science ($n = 203$), Earth and Space Science ($n = 251$), and Engineering Design ($n = 16$). Each passage ranged from 3 to 12 sentences, preserving complete conceptual units rather than isolated statements. Each passage was manually curated and annotated with four structured attributes:

- `english_text`: the original English passage
- `textbook_grade`: the instructional grade level (Grades 3–5, Middle School, or High School)
- `domain`: the corresponding NGSS science discipline (Life Science, Physical Science, Earth and Space Science, or Engineering Design)
- `subject`: a fine-grained topical label within the domain (e.g., "Discovering Cells" within Life Science).

Analysis of the full corpus indicated that textual complexity was primarily driven by textbook grade-level rather than science domain. Based on this observation, we selected a balanced random sample of 125 passages, ensuring representation across grade bands and science domains, for use in translation and evaluation experiments.

3.2. Translation Pipeline and Experiments

The translation process was conducted in three phases using Gemini 2.5 Pro, GPT-4o, and Aya 23 8B [30], an open-weight multilingual model run locally on a laptop equipped with an NVIDIA RTX 4070 GPU. First, we established baseline performance by translating all 125 passages without specifying target grade levels to assess whether the models could naturally preserve grade-appropriate linguistic complexity when translating educational science content.

Second, we established a target-grade baseline by explicitly specifying target grade levels in the translation prompts, without applying any retry or refinement mechanism. Each source passage was translated into multiple target grade bands using a "top-down" approach: higher-grade content (e.g., High School) was progressively translated to simpler grade levels (11th–12th, 9th–10th, 7th–8th, etc.), while lower-grade passages were translated to fewer target grades. This process yielded 608 translations per model and evaluated whether explicit grade-level instructions alone improved alignment prior to iterative refinement.

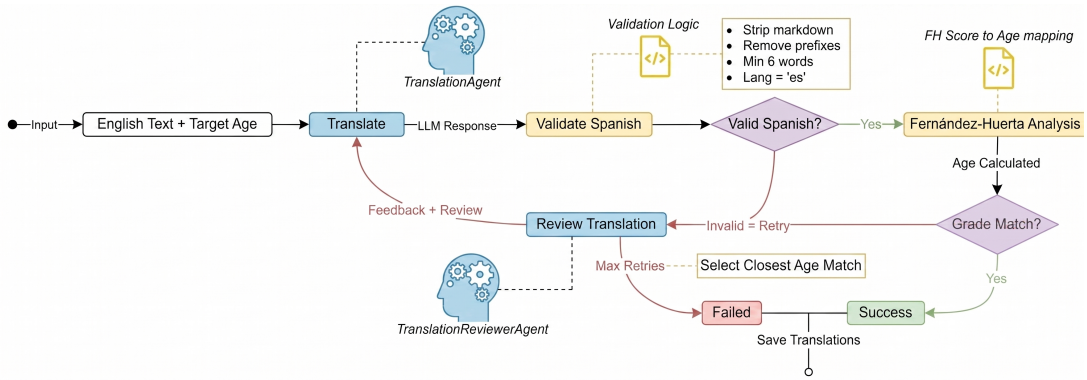

Figure 1: Multi-agent retry workflow for grade-level translation.

Third, we implemented a multi-agent retry system, shown in Figure 1, using AutoGen [31] to iteratively refine translations. The system comprises two agents: a *TranslationAgent*, which generates Spanish translations targeting specific grade levels, and a *TranslationReviewerAgent*, which evaluates translations using Fernández-Huerta readability metric. The *TranslationAgent* uses a seed prompt developed iteratively with input from practicing educators and refined through early experimentation (see Supplementary Materials A). Each target grade level is associated with a standardized age range (e.g., 5th grade → ages 10–11), and the FH index of the Spanish translation is used to determine its achieved age bracket. When the achieved age range does not align with the target grade level, the reviewer provides targeted feedback (e.g., "simplify vocabulary," "shorten sentences"), which is incorporated into subsequent translation attempts. This loop continues until the translation satisfies the target grade-level criterion or a maximum of five attempts is reached. For standards alignment evaluation, we first aggregated all successful translations (those achieving exact grade-to-age matches) across the three models.

We initially evaluated the corresponding English source passages against NGSS standards using the Chan Zuckerberg Initiative (CZI) Learning Standards Knowledge Graph [32]. Following CZI's recommended approach for standards-based retrieval, we restricted the search to Science standards with "Multi-State" jurisdiction (i.e., NGSS standards) and applied grade-based filtering to ensure developmentally appropriate matches. Semantic similarities between passages and standards were computed using sentence-transformer embeddings (`all-MiniLM-L6-v2`), and the top five candidate standards

were retrieved for each passage based on cosine similarity. After an initial pass of human annotation, this method proved to be unusable as the reviewer rejected 95% of the tagged standards. This would have rendered any downstream translation-preservation evaluation uninterpretable, as a rater cannot assess whether or not a translation preserves a standard that the source passage was never aligned with.

This standards tagging failure inspired a pivot to a different method of NGSS standard tagging using Anthropic's Claude Opus 4.6. The full NGSS standards corpus was partitioned into four grade bands: Kindergarten through second grade, grades 3–5, middle school, and high school. Each passage was matched to the grade band, and the model was prompted to return the set of standards where a student could possibly learn some part of the standard from the given passage. The total cost of three iterations of this pipeline was approximately six dollars. Of the 125 source passages, the multi-agent system successfully translated 98 passages for all three models. Of those 98, all but 16 were tagged with at least one standard (see Figure 2). Human annotators were still asked to rate those translations for quality, but not standard alignment. They were given the option of writing in a specific NGSS standard if they felt that it was applicable.

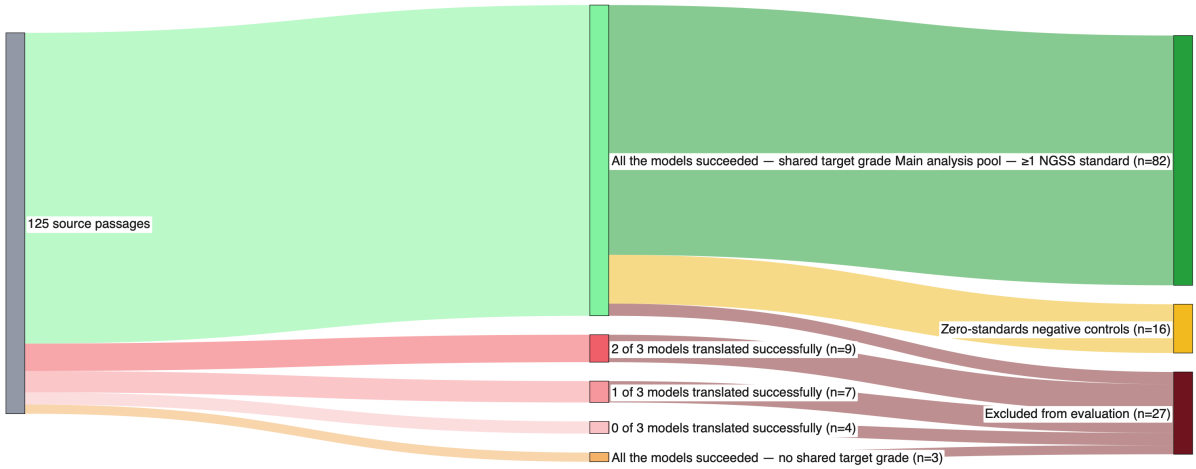

Figure 2: Filtering pipeline from the 125 source-passage sample to the 98-passage human-evaluation pool. Twenty-three passages lacked a clean A/B/C translation triple at a shared target grade, and four were not yet processed by the standards tagger; the remaining 98 split into 82 with at least one NGSS standard from the tagger (main analysis pool) and 16 with zero standards (retained as negative controls).

3.3. Human Evaluation

Human evaluation was conducted by one secondary public school teacher and one primary public school teacher, from two different U.S. Southwestern cities near the border with Mexico. Both raters are primary and secondary Science specialists. Annotation took place via a custom static HTML annotation tool deployed via GitHub Pages. The raters were asked to rank three translations Worst/Middle/Best (one from each model), and to evaluate how well each translation retained NGSS standards from the original passage (the tagging was automated). Each translation also had a field where the teacher could comment on their rating, for use with prompt optimization.

3.4. GEPA Prompt Optimization

The two processes described in this paper are separate but sequential: the second (GEPA optimization) builds on the data produced by the first (the multi-agent retry system). The multi-agent retry system revises the *translation*, re-translating a single passage until it hits its target FH index. GEPA optimization then leverages the translations and human feedback created by the multi-agent system to revise the *system prompt*, producing a reusable artifact tuned for domain-specific translations.

For GEPA optimization we constructed 76 training examples (the subset of the 98-passage evaluation pool with at least one rater-confirmed NGSS standard), split 50/26 between training and validation. Each example paired an English source passage with the Aya 23 8B agent-retry Spanish translation produced for that passage, along with the corresponding Set 1 human-evaluation data: per-standard preservation ratings, Best/Middle/Worst forced-rank quality labels, and rater free-text comments (e.g., "Loses all context in the original source"). The rater free-text comments served as the textual reflection feedback that GEPA's reflection model read when proposing each prompt revision; this is the channel through which qualitative failure modes that raters flagged — preamble artifacts (*"Aquí está la traducción del pasaje..."*), oversimplification of scientific terms — were folded into the evolving prompt across iterations.

We took the zero-shot prompt (originally designed with educator input) as the starting point. Optimization ran on an NVIDIA GB10 with 128 GB unified memory, with Cohere's Aya 23 8B serving translations and Alibaba's Qwen 3.6-27B [33, 34] serving as the reflection model that rewrote the prompt at each iteration; both ran locally via Ollama. We used DSPy's "light" optimization setting, which produced 14 prompt-revision iterations and 403 total evaluations over 14.1 minutes of wall-clock time (mean 60.3s per iteration).

4. Results

4.1. Phase 1: Baseline (No Target Grade)

In the baseline condition, where no target grade level was specified, all three models struggled to produce translations that aligned with the instructional grade of the original English passages. Aya achieved the highest match count (20/125), followed by GPT-4o (15/125) and Gemini 2.5 Pro (13/125). These results indicate that unguided translation alone is insufficient for educational settings where a retry system could be leveraged.

Across all three models, Spanish translations clustered within a narrow Fernández–Huerta range corresponding to upper elementary readability (approximately FH 82–99), regardless of the source textbook grade. Consequently, lower-grade passages aligned more often by chance due to overlap with this default readability range, while middle school and high school passages were systematically under-leveled.

Phase 1 establishes a clear baseline limitation: without explicit grade-level guidance the large language models converge toward a generic middle-grade readability level, making them unreliable for direct use in grade-sensitive educational translation tasks.

4.2. Phase 2: Target Grade Specified (No Retry)

When target grade levels were explicitly specified in the prompt, grade-level alignment improved relative to the baseline but remained limited overall. Across all models, explicit grade-level prompting increased successful matches, but overall alignment rates remained low. Gemini 2.5 Pro produced the highest number of successful translations (124/608), followed by GPT-4o (95/608) and Aya (92/608), indicating that models can respond to grade-level instructions to some extent even without iterative refinement.

Performance varied substantially by target grade. Translations targeting middle-grade levels were most likely to align, particularly for grades 5–8. In contrast, early elementary targets (grades 2–3) were rarely achieved, and upper high-school–level translations (grades 11–12) were similarly difficult to produce, indicating that explicit grade-level prompting struggles to reliably generate outputs at the extremes of the readability spectrum.

Phase 2 demonstrates that while explicit grade-level prompts improve translation control relative to baseline, single-pass prompting remains inadequate for consistently achieving precise grade-level readability without iterative refinement.

4.3. Phase 3: AI Multi-Agent Retry System

The multi-agent retry system, shown in Figure 1, substantially improved grade-level alignment relative to Phase 2. Gemini 2.5 Pro achieved the highest number of successful translations (495/608) while requiring the fewest attempts on average. GPT-4o followed with 408/608 successful translations, and Aya achieved 233/608. These results indicate that all models benefited from retries and iterative feedback.

Performance continued to vary by target grade, but the retry mechanism substantially narrowed the gap between middle grades and more challenging extremes. Translations targeting grades 5–8 aligned most consistently, while upper high-school targets (grades 9–12) also showed improvement relative to Phase 2. Early elementary targets (grades 2–3) remained the most difficult, though alignment increased dramatically compared to the single-attempt prompt, indicating that even the most challenging grade levels benefited from iterative feedback. Table 2 shows that multiple retries were typically required to achieve grade-level alignment in Phase 3.

In contrast to earlier phases, the Fernández–Huerta scores in the retry system were more consistently distributed around their intended grade ranges, with less overlap between adjacent targets.

4.4. Token Usage

Token usage in Phase 1 to 3 is summarized in Table 2. Although Gemini 2.5 Pro consumed substantially more tokens per translation due to its reasoning design and achieved somewhat higher match rates, GPT-4o and Aya 23 8B attained comparable alignment with far fewer tokens, indicating that reasoning token usage alone does not ensure improved outcomes.

Allowing retries substantially increased token usage for all models in Phase 3. Gemini 2.5 Pro and GPT-4o incurred comparable overall token usage, while Aya required markedly more tokens while achieving fewer successful translations. The increase in token usage primarily reflects the cumulative cost of multiple translation attempts during retries, rather than higher-quality output from any single generation.

These results show that token usage scales with the number of allowed retries, and that additional tokens are most effective when retries lead to successful grade alignment rather than accumulating cost without improving alignment. The total API cost for all three phases was approximately $30 for GPT-4o and $37 for Gemini 2.5 Pro, while Aya ran at zero cost on local hardware.

Table 2
Average token usage and retries per translation across phases, with API spend by model. Translation costs cover all three phases combined. Additional pipeline costs: standards tagging using `all-MiniLM-L6-v2` (local, $0) initial pass and Claude Opus 4.6 (Anthropic, $6) production pass; GEPA optimization using Qwen 3.6-27B (local, $0).

	Average Tokens Used					
Model	**Phase 1**	**Phase 2**	**Phase 3**	**Avg Retries**	**Provider**	**Cost (USD)**
Aya 23 8B	483	548	8,612	3.94	Local	$0
GPT-4o	474	535	6,332	3.35	OpenAI	$30
Gemini 2.5 Pro	2,314	2,248	6,582	2.84	Google	$37

4.5. Retry Agent performance by model

The two human evaluators independently rated the multi-agent retry translations on the 98-passage pool. For each passage they produced three judgments: a binary check of whether the English source actually teaches each Opus-tagged NGSS standard, a 1–4 preservation rating per (standard, translation) pair, and a forced-rank Worst/Middle/Best label across the three model translations. Inter-rater reliability was slight-to-fair across all three tasks: Cohen's κ was 0.185 for source-confirmation (n = 240), 0.365 quadratic-weighted for forced-rank quality (n = 243 over 81 pages), and 0.305 quadratic-weighted for

preservation (n = 144). Agreement was strongly asymmetric: raters picked the same worst translation on 68% of pages (chance: 33%) but picked the same best on only 37%. We therefore frame per-model claims through the IRR-defensible signal — which model is bad — rather than headline marginals on which is best. Across passages where both raters used forced-rank, Aya 23 8B was the consensus worst translation on 22/55 passages, more than either larger model (GPT-4o: 18, Gemini 2.5 Pro: 15). On the smaller set of passages where both raters agreed on a best, Gemini led with 14, then Aya 10 passages and GPT-4o 6 passages. GPT-4o had the highest Middle count for both raters, sitting as the consensus middle.

As described in §3.3, the initial CZI semantic-similarity tagger results (sentence-transformer cosine similarity, grade-filtered, top-5 per passage) were rejected by the first-round rater on 95% of tagged (passage, standard) pairs, which would have rendered downstream preservation analysis impossible. The Claude Opus 4.6 method, prompted to return only standards a student could plausibly learn "some part of" from the passage, reduced the human rejection rate to 66% (rater 1) and 55% (rater 2), a ~30 percentage-point reduction in rejection at a marginal cost of approximately $6 (Table 2). Acceptance varied by grade band (rater 2, the more permissive rater): 52% at Middle School, 48% at Grade 5, 42% at High School, and 33% at Grades 3–4, suggesting Opus's "some part of" phrasing tags more loosely at the elementary extremes where NGSS performance expectations are broader. The mean preservation ratings (Table 3) gave a consistent ordering — Aya preserved standards least, with Gemini 2.5 Pro and GPT-4o nearly deadlocked.

Table 3
Standards preservation rating distributions (Set 1, multi-agent retry pipeline). Rater 1's distribution is right-truncated at 3; they never assigned the value 4.

Rater	Model	1	2	3	4	n
Rater 1	Aya 23 8B	31	33	18	0	82
	Gemini 2.5 Pro	32	16	34	0	82
	GPT-4o	17	34	31	0	82
Rater 2	Aya 23 8B	12	34	45	18	109
	Gemini 2.5 Pro	5	19	53	32	109
	GPT-4o	4	26	55	24	109

4.6. GEPA Optimization vs. Seed prompt

We chose Aya 23 8B for optimization because it was the smallest model, with the worst results from human evaluations. We hoped that GEPA optimization with a larger more sophisticated reasoning model could produce some reasoning steps for the smaller model to follow.

As optimization iterations progressed, several prompt features emerged. After the first reflection iteration, markdown bullet points and bold emphasis appeared and persisted through subsequent alterations, while an explicit no-preamble directive present from iteration 1 was eventually formalized as a "Zero-Addition Policy" in the final selected prompt. Across the 14 reflection iterations the prompt fluctuated in the 1,500–3,500 character range, peaked at 6,101 characters in iteration 14, and was then selected at 2,720 characters from the Pareto frontier — roughly a 4× growth from the 642-character seed (Figure 3; see Supplementary Materials B). Mean reflection time per iteration was 60.3 seconds and scaled roughly linearly with output length, for a total optimization budget of 14.1 minutes.

To test whether the GEPA-optimized prompt actually translates better than the seed prompt on the local model, we generated 30 paired Aya 23 8B translations on passages that the original retry pipeline had previously failed on (27 from the excluded pool plus 3 validation passages not used for optimization), and blinded each pair into A/B positions. This set of passages was chosen to avoid using passages that were part of the optimization process. We then asked the same two raters which translation better preserved English source content, and which had better Spanish fluency. Pooled across both raters, the optimized prompt won **62% of decisive content-preservation comparisons** (28 of 45; 95% CI 48–76%;

Table 4
GEPA pairwise preferences (Set 2, $n = 30$ passages). Counts shown as Optimized / Seed / Tie; sign tests are computed on decisive votes (ties excluded). The Both-raters-agreed rows restrict to pages where both raters picked the same option (excluding the 15 / 16 mixed-disagreement pages for content preservation / Spanish fluency). CIs given in percent.

Dimension	Subset	O / S / T	Dec. n	Opt. rate (95% CI)	p
Content pres.	Rater 1	18 / 9 / 3	27	67% [49, 84]	0.083
	Rater 2	10 / 8 / 12	18	56% [33, 79]	0.637
	Combined	**28 / 17 / 15**	**45**	**62% [48, 76]**	**0.101**
	Both agreed	8 / 5 / 2	13	62% [32, 86]	0.581
Sp. fluency	Rater 1	16 / 10 / 4	26	62% [43, 80]	0.239
	Rater 2	13 / 13 / 4	26	50% [31, 69]	1.000
	Combined	**29 / 23 / 8**	**52**	**56% [42, 69]**	**0.405**
	Both agreed	9 / 5 / 0	14	64% [35, 87]	0.424

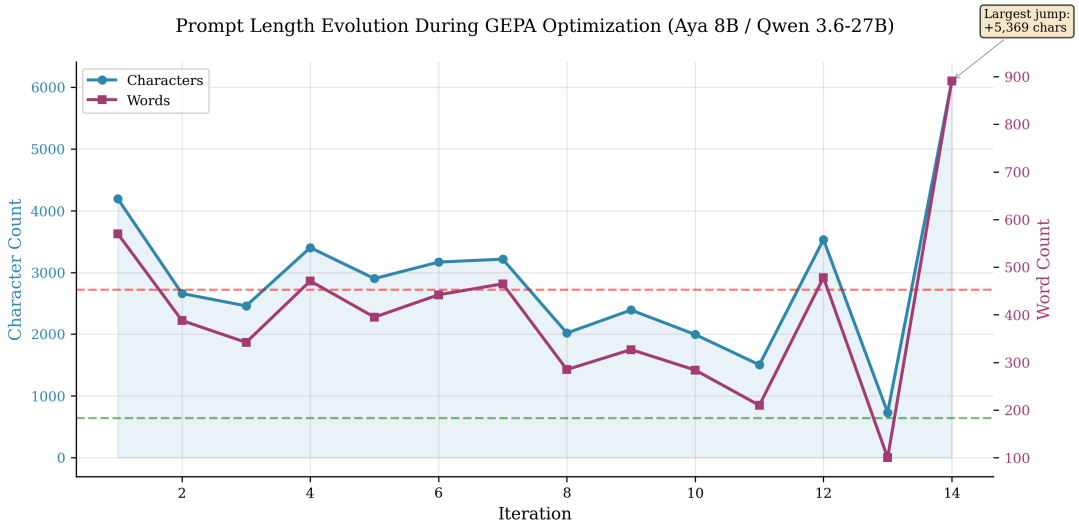

Figure 3: Prompt length evolution across the 14 GEPA reflection iterations on Aya 23 8B.

sign-test $p = 0.10$) and **56% of fluency comparisons** (29 of 52; 95% CI 42–69%; $p = 0.41$). Direction was consistent: no rater × dimension cell favored the seed prompt. Restricting to passages where both raters agreed on a preference, the optimized prompt led **8 to 5 on content preservation** and **9 to 5 on Spanish fluency** (see Table 4). Inter-rater reliability was again only slight-to-fair ($\kappa = 0.27$ content, $\kappa = 0.12$ fluency), consistent with the Set 1 finding that fine-grained pairwise translation judgments produce noisy paired labels at the passage level.

The qualitative comments supplied the most convergent evidence: rater 1 repeatedly flagged the seed prompt's preamble-artifact failure mode (e.g., *"Aquí está la traducción del pasaje..."*, *"all the conversations with the AI tool"*) — precisely the behavior the optimized system prompt's Zero-Addition Policy was written to suppress. We read these results as directional support for GEPA-optimized prompting on the smallest model, strongest on content preservation, albeit with $n = 30$ underpowered for a confirmatory statistical claim.

5. Discussion

Human evaluations suggest that model scale matters for this type of translation. The lower parameter model Aya 23 8B was clearly judged as the overall consensus worst in most scenarios. However, Rater 1 picked Aya 23 8B as Best on 21 of 81 individual passages with standards tagged; Rater 2 on 32; and both raters agreed on 10 (see Table 5). Aya 23 8B actually received more Best ratings than GPT-4o, despite being roughly 1/250th GPT-4o's rumored parameter count. This suggests prompt optimization can help draw out the model's better tendencies.

Table 5
Per-model forced-rank quality counts (Set 1 headline subset, $n = 81$ pages, 243 labels per rater). "Both Best" / "Both Worst" are pages where both raters independently selected the same model.

	Rater 1			Rater 2			Both raters agree	
Model	Worst	Middle	Best	Worst	Middle	Best	Best	Worst
Aya 23 8B	30	30	21	35	14	**32**	10	**22**
Gemini 2.5 Pro	27	19	**35**	22	28	31	**14**	15
GPT-4o	23	**33**	25	25	**39**	17	6	18
Totals (agreed)							30	55

Cohere's open-weight Aya Expanse 32B is a natural next target. Future experiments will explore how this larger open model compares with the best closed, trillion-plus parameter closed models with the added benefit of locality. In resource, cost, and privacy-constrained settings, locality is a catch-all solution. Regardless, 8B remains the sweet spot at which most consumer laptops with a GPU or M-series processor can capably serve inference, and it will continue to be a target for optimization and experimentation.

We are further encouraged by the results of GEPA optimization. As the seed prompt evolved, it became apparent that the reflection process allowed the preferences and directives of human experts to gradually fold into the optimized prompt. We used the larger Qwen 3.6-27B to optimize the smaller Aya 23 8B. Tests by the Mosaic Research Team at Databricks show that prompts optimized on a large model can lead to significant performance gains when used with smaller models, resulting in inference rates that are 20–90 times cheaper in implementation than using large models. They hypothesized that the gains on smaller models are likely due to the large model computing reasoning steps "ahead of time" and writing them to the prompt for smaller models to execute [35]. This process requires zero gradient updating, and thus does not require expensive GPU time or technical expertise. For targeted tasks for which human experts can give ratings and feedback, GEPA prompt optimization can be an impactful practice.

The translations provided by the three different methods (seed prompt, agent retry, and GEPA optimized) reveal some interesting dynamics. Consider the example of a passage on electromagnets, using the Aya 23 8B model (see Table 6).

In the translation using the agent-based retry system at a 2nd–3rd-grade target, "Electromagnets" is translated to "Los imanes mágicos" ("magical magnets"), a simplification our human evaluator felt changed the meaning of the source text. Aya 23 8B's translation under the seed prompt alone (no agent-based retries) at the same target instead used "imanes eléctricos." This is less precise than the canonical Spanish technical term "electroimanes," but is recognizably scientific and far more accurate than "magical magnets." This could mean that the agent-based retry system balances two objectives that occasionally compete with one another: (1) to tune the Fernández–Huerta index of the passage toward the target grade, and (2) to include scientific terminology, which can sometimes have more syllables or length than the average grade-level word. When grade-level pressure dominates the loop, the retry agent can simplify its way out of the source's technical content entirely.

The GEPA-optimized prompt on Aya 23 8B for the same English text also produces "imanes eléctricos" and preserves additional technical vocabulary including "corriente eléctrica," "bobinas de alambre," and

Table 6
Physical Science lesson on Magnetism, targeted at students in grades 2nd–3rd (ages 7–8), Agent-based Retry system, with FH index of 93.72

Category	Content
Original	How are electromagnets used? Electromagnets are often more useful than permanent magnets because: you can turn them on and off by switching the current on and off; you can make them stronger by increasing the current and/or the coils of wire. Electromagnets are used in many things people use every day...
Translation	Los imanes mágicos se encienden y se apagan con un cambio simple. ¡Pueden ser muy fuertes! Se usan en cosas como la alarma cuando abres la puerta y en los juguetes que se mueven, como los trenes.
Human Comment	"Terrible translation, the vocabulary is too simple for students. Key terms are too simplified and it does not have the same complexity and explanation of the source."

Table 7
Fernández–Huerta scores on the electromagnets passage (target grade 2nd–3rd, age 7–8) across three Aya 23 8B configurations. All FH values computed with a single consistent formula; distance from target is to the band floor of ≥ 100.

Configuration	FH	FH band → grade	Dist. from target
Target	≥ 100	very easy / $\leq$ grade 5	0
Retry agent (5 attempts, "imanes mágicos")	**93.72**	very easy / $\leq$ grade 5	−6
Seed prompt (single-shot)	57.31	somewhat difficult / grade 11–12	−43
GEPA-optimized (single-shot)	46.14	difficult / university	−54

"imanes permanentes." This terminology retention is consistent with the content-fidelity emphasis of prompt optimization, but it also pushes the FH index further from the target: the GEPA-optimized translation scores FH = 46.14, 54 points below the 2nd–3rd-grade band floor of 100 (Table 7). Removing these scientific compounds ("imanes eléctricos," "imanes permanentes," "corriente eléctrica," "bobinas de alambre") and recomputing FH lifts the score to 66.44; removing additionally "motores" and "guitarras eléctricas" lifts it to 71.67. Even at the more aggressive stripping level the vocabulary-blind FH score remains roughly 22 points below the retry agent's 93.72. The retry agent's FH advantage on this passage therefore comes from more than compound-avoidance alone: the agent-based method also compresses its output to three short sentences and substitutes everyday referents ("la alarma cuando abres la puerta," "los juguetes que se mueven, como los trenes") for the source's pedagogical examples (doorbells, hobby trains and cars, electric guitars), trading content fidelity for readability gain.

One promising extension of this research could be to test an agent-based system that removes scientific or domain-specific vocabulary *before* calculating FH index, allowing the system to push the complexity of the surrounding terms towards the target grade level while retaining the accuracy and pedagogical usefulness of the translation. There is also considerable room to test different GEPA optimization methods; such as using translations from larger models as training data, longer rollouts, and larger reflection models.

For future research we hope to further test a number of alternative approaches to agent-based translation and prompt optimization. It is worth noting that, in our context, testing means subjecting the translations to scrutiny by educators and students. While automated metrics like FH can be useful for optimization or an approximation of grade-level complexity, we are ultimately interested in creating robust translation systems that educators feel comfortable deploying via online textbooks or classroom settings. We see this pipeline less as a standalone translator than as a component of an intelligent textbook or learning management system: rather than serving a fixed translation, a system could re-level a passage on demand the moment a student opens it.

5.1. Limitations

The configuration of the agent-based retry system led to some noteworthy limitations. Of the 125 passages we had initially used for the retry agents, the 98 passages that human evaluators reviewed were those in which all three models produced successful translations. Of the remaining 27 passages, at least one of the three models either timed out, errored out, or could not reach the target FH index within five retries. This could bias the other 98 passages towards those that can be easily translated.

6. Conclusion

Our work highlights the complexity of using readability metrics, agent-based frameworks, and prompt optimization to improve textbook translation quality for educators. While closed models (GPT-4o and Gemini 2.5 Pro) did provide better simplified translations according to human raters, the victory was not nearly absolute, as the local model (Aya 23 8B) did manage to beat the closed models in many cases. GEPA optimization allowed for the evolution of content preservation and formatting preferences, resulting in improved translations. While we did not ultimately find the optimal balance of simplification and preservation, we provide a clear road map for future exploration.

Although we focused on translating English to Spanish, the overall process is replicable for any language pair. GEPA-optimization using human evaluations of the agent-based retry system output provides a way for researchers and engineers to leverage multiple forms of subject matter expertise to improve model performance. The optimized prompt we provide in our supplementary materials can help to create more useful English to Spanish textbook translations with the closed or local LLM of the user's choice. Ultimately, this workflow offers a blueprint for scaling high-fidelity, grade-appropriate curriculum adaptation without requiring a proportional increase in human expert time.

7. Acknowledgments

Thanks to Jose Chavez for providing some essential background on the shortcomings of LLM translation in English Language Development classrooms in California.

8. Declaration on Generative AI

During the preparation of this work, the authors used Anthropic's Claude in order to: generate and debug research code; produce plots from experimental results via that code; and check grammar, spelling, and typos. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

9. Code and Supplementary materials

Data and materials are available on the OSF Project Overview.

References

[1] U.S. Department of Education, Teacher shortage areas nationwide listing, 2024–2025, https://tsa.ed.gov, 2024. Accessed: 2026-05-19.

[2] National Center for Education Statistics, Most U.S. public elementary and secondary schools faced hiring challenges for the start of the 2024–25 academic year, Press Release, October 17, 2024, https://nces.ed.gov/whatsnew/press_releases/10_17_2024.asp, 2024.

[3] McGraw Hill, McGraw Hill Introduces New AI Capabilities in Its Connect Digital Course Solution for Higher Education, Press release, 2026.

URL: https://www.mheducation.com/about-us/news-insights/press-releases/mcgraw-hill-introduces-new-ai-capabilities-in-its-connect-digital-course-solution-for-higher-education.html, accessed: 2026-06-15.

[4] J. Sugarman, Beyond Teaching English: Supporting High School Completion by Immigrant and Refugee Students, Technical Report, Migration Policy Institute, Washington, DC, 2017.

[5] L. A. Agrawal, S. Tan, D. Soylu, N. Ziems, R. Khare, K. Opsahl-Ong, A. Singhvi, H. Shandilya, M. J. Ryan, M. Jiang, C. Potts, K. Sen, A. G. Dimakis, I. Stoica, D. Klein, M. Zaharia, O. Khattab, Gepa: Reflective prompt evolution can outperform reinforcement learning, 2025. URL: https://arxiv.org/abs/2507.19457. arXiv:2507.19457.

[6] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer, The mathematics of statistical machine translation: Parameter estimation, Computational Linguistics 19 (1993) 263–311.

[7] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, F. M. Tyers, Apertium: A free/open-source platform for rule-based machine translation, Machine Translation 25 (2011) 127–144.

[8] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, 2015.

[9] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).

[10] S. Läubli, R. Sennrich, M. Volk, Has machine translation achieved human parity? A case for document-level evaluation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4791–4796. URL: https://aclanthology.org/D18-1512/.

[11] I. Alpizar-Chacon, S. Sosnovsky, Interlingua: Linking textbooks across different languages, in: Proceedings of the First Workshop on Intelligent Textbooks (iTextbooks'19) co-located with the 20th International Conference on Artificial Intelligence in Education (AIED 2019), volume 2384 of *CEUR Workshop Proceedings*, CEUR-WS.org, Chicago, IL, USA, 2019, pp. 103–116. URL: https://ceur-ws.org/Vol-2384/paper11.pdf.

[12] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 1877–1901.

[13] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, H. H. Awadalla, How good are gpt models at machine translation? a comprehensive evaluation, 2023. URL: https://arxiv.org/abs/2302.09210. arXiv:2302.09210.

[14] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, Philadelphia, PA, 2002, pp. 311–318.

[15] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 2685–2702.

[16] I. M. Barrio-Cantalejo, P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, P. Hernando, Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient, Anales del Sistema Sanitario de Navarra 31 (2008) 135–152. doi:10.4321/S1137-66272008000300004.

[17] F. Szigriszt Pazos, Sistemas predictivos de legibilidad del mensaje escrito: Fórmula de perspicuidad, Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain, 1993.

[18] J. Fernandez Huerta, Medidas sencillas de lecturabilidad, Consigna 214 (1959) 29–32.

[19] B. G. Johnson, B. Jerome, J. S. Dittel, R. Van Campenhout, Improving textbook accessibility through AI simplification: Readability improvements and meaning preservation, in: Proceedings of the Sixth Workshop on Intelligent Textbooks (iTextbooks'25) co-located with the 26th International Conference on Artificial Intelligence in Education (AIED 2025), volume 4010 of *CEUR Workshop Proceedings*, CEUR-WS.org, Palermo, Italy, 2025. URL: https://ceur-ws.org/Vol-4010/itb25_s2p2.pdf.

[20] R. Stodden, L. Kallmeyer, A multi-lingual and cross-domain analysis of features for text simplification, in: N. Gala, R. Wilkens (Eds.), Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI), European Language Resources Association, Marseille, France, 2020, pp. 77–84. URL: https://aclanthology.org/2020.readi-1.12/.

[21] P. Martínez, L. Moreno, A. Ramos, Exploring large language models to generate Easy to Read content, Frontiers in Computer Science 6 (2024). doi:10.3389/fcomp.2024.1394705.

[22] M. Wu, J. Xu, Y. Yuan, G. Haffari, L. Wang, W. Luo, K. Zhang, (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts, 2025. URL: https://arxiv.org/abs/2405.11804. arXiv:2405.11804.

[23] G. Wang, J. Hu, S. Ali, MAATS: Multi-agent automated translation system based on MQM, arXiv preprint arXiv:2505.14848 (2025).

[24] Z. Feng, J. Su, J. Zheng, J. Ren, Y. Zhang, J. Wu, H. Wang, Z. Liu, M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation, 2025. URL: https://arxiv.org/abs/2412.20127. arXiv:2412.20127.

[25] A. Peter, M. Dang, M. Liu, J. Dominguez, N. Lohia, MATT: Enhancing low-resource language translation via multi-agent workflow, SMU Data Science Review 8 (2024).

[26] R. Flesch, A new readability yardstick, Journal of Applied Psychology 32 (1948) 221–233.

[27] OpenStax, OpenStax free textbooks, https://openstax.org, 2024. Accessed: 2025-01-15.

[28] CK-12 Foundation, CK-12 free educational resources, https://www.ck12.org, 2024. Accessed: 2025-01-15.

[29] NGSS Lead States, Next Generation Science Standards: For States, By States, The National Academies Press, Washington, DC, 2013. URL: https://www.nextgenscience.org.

[30] V. Aryabumi, J. Dang, D. Talupuru, S. Dash, D. Cairuz, H. Lin, B. Venkitesh, M. Smith, J. A. Campos, Y. C. Tan, K. Marchisio, M. Bartolo, S. Ruder, A. Locatelli, J. Kreutzer, N. Frosst, A. Gomez, P. Blunsom, M. Fadaee, A. Üstün, S. Hooker, Aya 23: Open weight releases to further multilingual progress, 2024. URL: https://arxiv.org/abs/2405.15032. arXiv:2405.15032.

[31] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, et al., AutoGen: Enabling next-gen LLM applications via multi-agent conversation, arXiv preprint arXiv:2308.08155 (2023).

[32] Chan Zuckerberg Initiative, Learning standards knowledge graph, https://github.com/learningequality/standards, 2025. Accessed: 2025-01-15.

[33] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[34] Qwen Team, Qwen3.6, GitHub, 2026. URL: https://github.com/QwenLM/Qwen3.6, accessed: 2026-05-18.

[35] Mosaic AI Research Team, Building state-of-the-art enterprise agents 90x cheaper with automated prompt optimization, Databricks Blog, https://www.databricks.com/blog/building-state-art-enterprise-agents-90x-cheaper-automated-prompt-optimization, 2025.