ELiRF-UPV at SimpleText 2026: Decoupling Action Prediction and Controlled Generation for Text Simplification

ELiRF-UPV at CLEF 2026

Marc Hurtado[1,*,†], Vicent Ahuir[1,†], Antonio Molina[1,†] and María-José Castro-Bleda[1,2,†]

[1] Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camí de Vera, sn, València
[2] Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Universitat Politècnica de València, Camí de Vera, sn, València

Abstract

In this work, we describe the ELiRF-UPV participation in the CLEF 2026 SimpleText Track, Simplify Scientific Text. Our approach for text simplification decouples the process into two stages: planning and generation. First, a model predicts sentence-level simplification actions, which are then used to guide the final simplification. For the generation stage, we explore and compare two primary strategies. The first one is LLM-based, where we apply prompting techniques. For the second strategy, we fine-tune an encoder-decoder-based automatic simplification model. This design allows the models to learn new actions without suffering from catastrophic forgetting of their pre-trained knowledge. In addition, we apply Direct Preference Optimization (DPO) training to strictly align the model's outputs with the predicted control instructions, mitigating unwarranted changes and ensuring accurate simplifications. A core principle of our work is its alignment with Green AI guidelines; by using small open-source models, we aim to democratize controlled text generation and reduce the high computational and environmental costs associated with massive proprietary models.

Keywords

Text Simplification, Biomedical Literature, Controlled Text Generation, Action Prediction, Green AI, Direct Preference Optimization, Parameter-Efficient Fine-Tuning,

1. Introduction

The extensive specialized vocabulary and complex structures of biomedical literature make it difficult for non-expert audiences to understand the meaning of the text, leading to a barrier that can promote misinformation. To mitigate this, automated scientific text simplification has gained substantial traction. The recent state-of-the-art approaches, including the top-performing systems in last year's SimpleText competition, have been heavily dominated by prompting massive, private Large Language Models (LLMs). Even though these approaches are effective, their reliance on closed APIs limits reproducibility and incurs high computational, energy, and water costs. Beyond simple prompting, different techniques allow small open models to achieve competitive, reproducible results locally. Specifically, input-level conditioning guides generation by injecting explicit control tokens into the source sequence, whereas model-level conditioning adapts the internal architecture to inherently process these constraints.

ELiRF-UPV is participating in the CLEF 2026 SimpleText Track [1] [2]. Our methodology relies on a two-module architecture dividing the task into planning and generation. An initial module predicts specific simplification actions, which are then processed by a generation module using either advanced

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
mhurben@upv.es (M. Hurtado); vahuir@upv.es (V. Ahuir); amolina@upv.es (A. Molina); mcastro@dsic.upv.es (M. Castro-Bleda)
0009-0005-5651-5243 (M. Hurtado); 0000-0001-5636-651X (V. Ahuir); 0000-0001-6537-8803 (A. Molina); 0000-0003-1001-8258 (M. Castro-Bleda)

prompting techniques or a combination of fine-tuning and Direct Preference Optimization (DPO) applied to encoder-decoder models, [3] .

A core principle of our approach is its alignment with Green AI guidelines. Instead of relying on massive, proprietary models with high computational, energy, and water costs, we advocate for the use of efficient and optimized open models. By doing so, we aim to democratize access to controlled text generation technologies, demonstrating that it is possible to achieve highly competitive simplification systems that can run in local environments while minimizing environmental impact.

2. Tasks and Datasets

Specifically, we present our proposal for Task 1 (Simplify Scientific Text), addressing the challenge by providing results for both sentence-level (Task 1.1) and document-level simplification (Task 1.2). [4]

Task 1 focuses on the simplification of biomedical literature, utilizing abstracts and their corresponding Plain Language Summaries (PLS) from Cochrane systematic reviews. The objective is to make highly technical medical texts accessible to non-expert audiences. To achieve this, Task 1.1 asks systems to simplify individual "shredded" sentences extracted from these abstracts. Conversely, Task 1.2 requires the simplification of entire abstracts as a single document, with the constraint that the generated simplification must be readable on its own and its overall message must remain consistent with the conclusions of the full review.

The evaluation of the current edition is carried out on the official 2026 test set. This new set presents a dataset divided into 11 different languages, with a total of 48,809 sentences for task 1.1 and 8,069 documents. However, the main evaluation is based on the newly published Cochrane abstracts from the last year. Specifically, this core evaluation subset focuses exclusively on the English versions and is structurally restricted to the "Main results" and "Authors' conclusions" sections of both the abstracts and their paired plain language summaries.

Despite the characteristics of this new 2026 test set, due to the absence of a new official training corpus for this edition, our system has been trained using the Cochrane-auto dataset [5]. This corpus consists of 1,085 documents, structured into 4,171 paragraphs and 14,719 sentences, all in English. An example of the data is shown in the Table 1

The main strength of this corpus is that it includes the label of the specific transformation (the action) applied to each source sentence during the simplification process. The possible actions are: Delete, Copy, Rephrase, Split, Merge.

It is important to highlight a methodological modification that we have introduced to the reference labels of the corpus. While the original *Cochrane-auto* dataset defines the merge blocks, the sequence of sentences to be joined with the labels Merge and None, we have decided to change the labeling scheme. Specifically, we assign the `<Merge_B>` label to the first sentence of a block and the `<Merge_I>` label to the remaining sentences of the same merge block. We implemented this change because it provides a much clearer and more understandable representation of the sequence boundaries.

Sentence-level example (Task 1.1)	
Complex	The one included trial suggests no evidence of an effect of fat supplementation of human milk on short-term growth and feeding intolerance in preterm infants.
Action	*<Split>*
Simple	Addition of extra fat to human milk for preterm infants showed no clear benefits with regards to short-term rates of weight gain, length gain, and head growth. There was no evidence that the extra fat increased the risk of feeding intolerance.
Document-level example (Task 1.2)	
Complex	'This review includes a total of 1897 children and 729 adults in 39 trials.', 'Thirty-three trials were conducted in the emergency room and equivalent community settings, and six trials were on inpatients with acute asthma (207 children and 28 adults).', 'The method of delivery of beta2-agonist did not show a significant difference in hospital admission rates.', 'In adults, the risk ratio (RR) of admission for spacer versus nebuliser was 0.94 (95% CI 0.61 to 1.43).', 'The risk ratio for children was 0.71 (95% CI 0.47 to 1.08, moderate quality evidence).', 'In children, length of stay in the emergency department was significantly shorter when the spacer was used.', 'The mean duration in the emergency department for children given nebulised treatment was 103 minutes, and for children given treatment via spacers 33 minutes less (95% CI -43 to -24 minutes, moderate quality evidence).', 'Length of stay in the emergency department for adults was similar for the two delivery methods.', 'Peak flow and forced expiratory volume were also similar for the two delivery methods.', 'Pulse rate was lower for spacer in children, mean difference -5% baseline (95% CI -8% to -2%, moderate quality evidence), as was the risk of developing tremor (RR 0.64; 95% CI 0.44 to 0.95, moderate quality evidence).'
Action	*<Rephrase>* → *<Rephrase>* → *<Rephrase>* → *<Delete>* → *<Delete>* → *<Rephrase>* → *<Split>* → *<Split>* → *<Delete>* → *<Rephrase>*
Simple	'We found 39 clinical trials involving 1897 children and 729 adults.', "Thirty-three of the trials were conducted in an emergency room (or emergency department) and community settings (such as a GP's surgery), and six trials were on inpatients (people in hospital) with acute asthma (207 children and 28 adults).", 'Taking beta-agonists through either a spacer or a nebuliser in the emergency department did not make a difference to the number of adults being admitted to hospital, whilst in children we can be fairly confident that nebulisers are not better than spacers at preventing admissions.', 'In children, the length of stay in the emergency department was significantly shorter when the spacer was used instead of a nebuliser.', 'The average stay in the emergency department for children given nebulised treatment was 103 minutes.', 'Children given treatment via spacers spent an average of 33 minutes less.', 'In adults, the length of stay in the emergency department was similar for the two delivery methods.', 'However the adult studies were conducted slightly differently which may have made it more difficult to show a difference in the length of stay in the emergency department.', 'Pulse rate was lower in children taking beta-agonists through a spacer (mean difference -5% baseline), and there was a lower risk of developing tremor.'

Table 1
Examples of sentence-level and document-level simplification extracted from the Cochrane-auto corpus, illustrating the transformation flow from complex to simple text.

3. System description

3.1. Architecture

To carry out this task, we propose an architecture based on two distinct modules: an initial planning module responsible for predicting the specific simplification action for each sentence, followed by a generation module that uses these predicted actions to guide the final text production. For this generation module, we explore and compare two different strategies. On the one hand, we apply advanced prompting techniques on the Gemma 4 model by directly injecting the predicted actions into its context. On the other hand, we introduce an architectural modification to encoder-decoder models, specifically Flan-T5 [6] and Bart [7], combined with supervised fine-tuning and DPO. This approach allows the models to inherently integrate and process these explicit instructions during the simplification process. The architecture schema is detailed in Figure 1

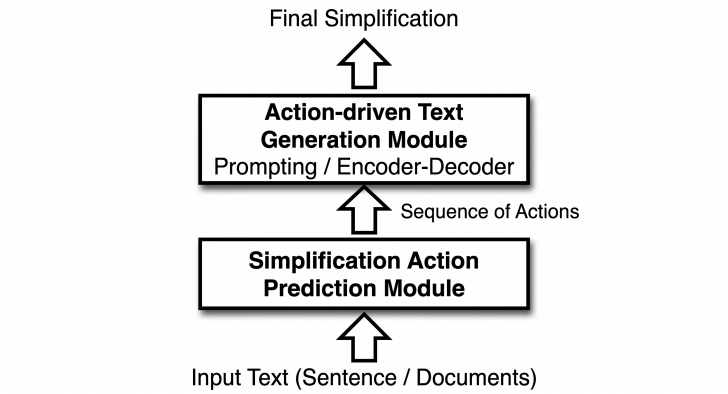

Figure 1: Architecture Diagram of the Text Simplification System

3.2. Simplification Action Prediction

In the context of the SimpleText26 shared task, we frame the action prediction as a sentence-level sequence-labeling over the 5 editing actions defined by the shared task: `Copy`, `Delete`, `Rephrase`, `Split`, and `Merge` (split into two – `Merge_B` and `Merge_I`). Given a document, the system predicts exactly one action per input sentence, so that the full output is a structurally valid sequence of edits that can be executed downstream.

3.2.1. Architecture

The system that we propose aims to tackle the following task: *given a document, the system should decide which action for simplification should be done for each sentence.* The possible actions are the following ones: $\mathcal{A} = \{\texttt{Copy}, \texttt{Delete}, \texttt{Merge\_B}, \texttt{Merge\_I}, \texttt{Rephrase}, \texttt{Split}\}$, with $C = |\mathcal{A}| = 6$. A document is a sequence of N sentences $\mathbf{s} = (s_1, \ldots, s_N)$ and the model assigns exactly one action $y_i \in \mathcal{A}$ to each sentence. The two-tag pair `Merge_B`/`Merge_I` encodes the multi-sentence merge groups annotated in the dataset in IOB style, so an arbitrary-length merge span can be described without enlarging the label set. Sentences are concatenated with a dedicated separator token `<sent>` before being passed to the encoder:

$$\mathbf{x} = [\texttt{<s>}, \texttt{<sent>}, s_1, \texttt{<sent>}, s_2, \ldots, \texttt{<sent>}, s_N, \texttt{</s>}].$$

The underlying model is a hybrid architecture built around a shared RoBERTa [8], large version, encoder f_θ that processes the full document in a single pass, so that every sentence representation benefits from cross-sentence attention:

$$H = f_\theta(\mathbf{x}) \in \mathbb{R}^{L \times d}.$$

A learnable sentence pooler π then collapses, for each sentence i, the contextual states between its '<sent>' marker and the next separator into a single vector

$$\mathbf{h}_i = \pi(H, i) \in \mathbb{R}^{d_p}, \qquad i = 1, \ldots, N.$$

On top of this representation, we attach three specialist binary branches, indexed by $b \in \mathcal{B} = \{\texttt{delete}, \texttt{copy}, \texttt{rephrase}\}$, that probe three orthogonal semantic axes – *"is this information relevant?"*, *"is this readable as-is?"* and *"does this need rewriting?"*. Each branch consists of a low-dimensional projection $W_b^{\text{proj}} \in \mathbb{R}^{d_b \times d_p}$ and a binary classifier $W_b^{\text{cls}} \in \mathbb{R}^{2 \times d_b}$:

$$\mathbf{z}_i^{(b)} = W_b^{\text{proj}}\, \text{sg}(\mathbf{h}_i), \qquad \boldsymbol{\ell}_i^{(b)} = W_b^{\text{cls}}\, \mathbf{z}_i^{(b)} \in \mathbb{R}^2,$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator that decouples the binary objectives from the shared pooler. Each branch is trained against the one-vs-rest indicator $t_i^{(b)} = \mathbb{1}[y_i = y_b^\star]$ derived from the dataset's gold actions, with binary cross-entropy,

$$\mathcal{L}_b = -\frac{1}{N} \sum_{i=1}^{N} \log p\Big(t_i^{(b)} \mid \boldsymbol{\ell}_i^{(b)}\Big)\,.$$

The pooled vector and the (detached) branch projections are concatenated and passed through a linear fusion head to produce per-sentence emission scores over the C actions:

$$\mathbf{u}_i = \big[\mathbf{h}_i\,;\, \text{sg}(\mathbf{z}_i^{(\texttt{delete})})\,;\, \text{sg}(\mathbf{z}_i^{(\texttt{copy})})\,;\, \text{sg}(\mathbf{z}_i^{(\texttt{rephrase})})\big], \qquad \mathbf{e}_i = W^{\text{emit}}\, \text{Dropout}(\mathbf{u}_i) \in \mathbb{R}^C.$$

The stop-gradients between the branches and the fusion layer fully decouple the specialist objectives from the main classifier: the binary heads provide stable auxiliary features without being reshaped by the main loss, while the fusion head and pooler are shaped only by the document-level objective.

The emissions are then passed through a linear-chain CRF that scores a full sequence with a learnable transition matrix $T \in \mathbb{R}^{C \times C}$ and boundary potentials $\boldsymbol{\pi}, \boldsymbol{\tau} \in \mathbb{R}^C$,

$$\Psi(\mathbf{y} \mid \mathbf{e}) = \pi_{y_1} + \sum_{i=1}^{N} e_{i,y_i} + \sum_{i=2}^{N} T_{y_{i-1},y_i} + \tau_{y_N}.$$

Structural validity is enforced by masking forbidden transitions and boundaries to $-\infty$,

$$T_{a,b} = -\infty \;\; \forall (a,b) \in \mathcal{F}, \qquad \pi_a = -\infty \;\; \forall a \in \mathcal{F}_{\text{start}}, \qquad \tau_a = -\infty \;\; \forall a \in \mathcal{F}_{\text{end}},$$

where $\mathcal{F}$, $\mathcal{F}_{\text{start}}$ and $\mathcal{F}_{\text{end}}$ encode the IOB rules implied by the SimpleText26 annotation scheme — `Merge_I` may only follow `Merge_B` or `Merge_I`, `Merge_B` must be followed by `Merge_I`, `Merge_I` cannot begin a document and `Merge_B` cannot end one — so that any decoded sequence corresponds to a well-formed group of merges by construction. The total training loss combines the CRF negative log-likelihood, an auxiliary class-weighted cross-entropy on the emissions that compensates the heavy imbalance observed in the dataset between frequent (`Copy`, `Delete`) and rare (`Split`, `Merge_*`) actions, and the three branch BCE terms:

$$\mathcal{L}_{\text{CRF}} = -\Psi(\mathbf{y}^\star \mid \mathbf{e}) + \log \sum_{\mathbf{y} \in \mathcal{A}^N} \exp \Psi(\mathbf{y} \mid \mathbf{e}), \qquad \mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} w_{y_i^\star} \log \text{softmax}(\mathbf{e}_i)_{y_i^\star},$$

$$\mathcal{L} = \mathcal{L}_{\text{CRF}} + \alpha\, \mathcal{L}_{\text{CE}} + \sum_{b \in \mathcal{B}} \lambda_b\, \mathcal{L}_b.$$

For our final prediction system, we train $S = 5$ independent replicas with different random seeds and combine them at inference time with an emission-level ensemble. Each replica produces pre-CRF emissions $\mathbf{e}_i^{(s)}$ and CRF parameters $(T^{(s)}, \boldsymbol{\pi}^{(s)}, \boldsymbol{\tau}^{(s)})$, which are averaged across seeds:

$$\bar{\mathbf{e}}_i = \frac{1}{S} \sum_{s=1}^{S} \mathbf{e}_i^{(s)}, \qquad \bar{T} = \frac{1}{S} \sum_{s=1}^{S} T^{(s)}, \qquad \bar{\boldsymbol{\pi}} = \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\pi}^{(s)}, \qquad \bar{\boldsymbol{\tau}} = \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\tau}^{(s)}.$$

The averaged parameters are loaded into a fresh linear-chain CRF $\bar{\Psi}$ that re-applies the structural masks to $\bar{T}$, $\bar{\boldsymbol{\pi}}$ and $\bar{\boldsymbol{\tau}}$. Averaging at the emission level rather than voting on hard labels preserves the relative confidence between actions, while averaging the transitions keeps the structural prior consistent across replicas. Because the RoBERTa encoder is limited to $L_{\max} = 512$ sub-word tokens, long test documents are processed with a sliding-window scheme: sentences are pre-tokenised and greedily packed into windows $W_k = [a_k, b_k)$ with $\sum_{i=a_k}^{b_k-1}(1 + |s_i|) + 2 \leq L_{\max}$, and consecutive windows share an overlap of $\Delta = 4$ sentences so that border sentences still receive sufficient left and right context. The ensemble emissions of each sentence are accumulated across all windows in which it appears and averaged,

$$\tilde{\mathbf{e}}_i = \frac{1}{|\mathcal{W}_i|} \sum_{k \in \mathcal{W}_i} \bar{\mathbf{e}}_i^{(k)}, \qquad \mathcal{W}_i = \{k : a_k \leq i < b_k\},$$

and a single constrained Viterbi pass under $\bar{\Psi}$ over $(\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_N)$,

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{A}^N} \bar{\Psi}(\mathbf{y} \mid \tilde{\mathbf{e}}),$$

yields one globally consistent action sequence per document, independent of where window boundaries fall with respect to the encoder budget.

3.2.2. Simplification Action Prediction Performance

We evaluate the action prediction sub-task with the per-class F1-score, reported per action and as the unweighted macro average. The macro variant gives the same weight to every action and is therefore sensitive to the rarer classes (`Split`, `Merge_B`, `Merge_I`) that dominate the label imbalance of the provided dataset, while the weighted average reflects the overall sentence-level accuracy of the system on the empirical class distribution. Table 2 reports the per-action and aggregate F1-scores of our system.

Table 2
Per-action and aggregate F1-scores of action simplification prediction system on the validation and test splits of the dataset.

Action	Validation	Test
`Copy`	0.293	0.253
`Delete`	0.489	0.440
`Merge_B`	0.090	0.115
`Merge_I`	0.165	0.123
`Rephrase`	0.548	0.534
`Split`	0.132	0.226
Macro	0.286	0.282
Weighted	0.458	0.434

Two observations stand out. First, the ranking of actions is consistent across the two splits and in line with the distribution of the dataset: `Rephrase` and `Delete` – the two most frequent classes, and the ones most clearly anchored in the input by the specialist branches – are predicted markedly better than the structurally constrained `Merge_B`/`Merge_I` pair, whose support is small and whose prediction additionally depends on the IOB transitions being correctly aligned along the sequence. Second, the macro F1 is essentially stable between validation (0.286) and test (0.282), and the weighted F1 only drops mildly (0.458 $\rightarrow$ 0.434), which indicates that the system generalises without a noticeable validation/test gap; the most visible movement is on `Split`, whose F1 actually improves on the test split (0.132 $\rightarrow$ 0.226), and on `Delete`, which drops by about five points. The gap between the macro F1 (which gives equal weight to every action) and the weighted F1 (which reflects the empirical class distribution) further quantifies how much of the system's overall accuracy comes from correctly handling the frequent actions, and how much headroom remains on the rarer ones.

3.3. Action-driven Text Generation

Once the actions are predicted, they are used to explicitly guide the final text generation process. To accomplish this, we detail two primary generation strategies.

3.3.1. LLM Prompting Approach

For our first generation strategy, we apply advanced prompting techniques to the open-weight model Gemma 4 (31B). This model was chosen because its relatively compact size enables local inference, perfectly aligning with our Green AI principles by avoiding reliance on private APIs and significantly reducing the energy footprint of the simplification process.

The prompt design is firmly grounded in the *National Institutes of Health* (NIH) guidelines for written health materials, ensuring that the generated text targets approximately an eighth-grade US reading level, appropriate for the general public. Crucially, the text generation is explicitly guided by the output of our initial planning module. To achieve this, the System Prompt is structured to define the six available simplification actions (`<Split>`, `<Copy>`, `<Delete>`, `<Rephrase>`, `<Merge_B>`, and `<Merge_I>`).

During the inference phase, the User Prompt supplies the model with the complete complex text to provide global context. This is immediately followed by a sequential breakdown of the text, where each source sentence is strictly paired with the exact simplification action predicted by the planning module. The complete design of both the System Prompt and the User Prompt utilized for this strategy is detailed in Appendix A.

3.3.2. Fine-tuning Approach

The fundamental basis of the fine-tuning approach is the use of new special tokens, which will indicate to the model how it should behave. In our case, a series of tokens is added for each aspect of the conditioning; specifically, the new action tokens actions are: `<Copy>`, `<Rephrase>`, `<Delete>`, `<Split>`, `<Merge_B>`, and `<Merge_I>`. In addition to action tokens, we defined an special token indicating the beginning of a sentence, `<Sentence>`.

This token creation not only involves modifying the tokenizer so that it understands these custom strings and transforms them into numbers to obtain the embeddings, but it also requires modifying the transformer architecture itself.

Regardless of the encoder-decoder model used, we cannot simply add new tokens to the original embedding matrix, since adding more tokens creates new rows in the matrix. This generates a hidden problem during training: since we are using a pre-trained model with knowledge acquired through extensive training on diverse corpora, this matrix would contain many meaningful tokens alongside 7 new tokens that the model has never seen and does not know what they mean.

Faced with this problem, intuition would tell us that the most direct solution is to apply a freeze to the original tokens to protect them, leaving only the new action tokens to be trained. However, in practice we encounter a major technical limitation of frameworks like PyTorch, upon which Hugging Face Transformers are based: embedding matrices do not natively allow partial freezing. The "requires_grad" attribute applies to the entire matrix (either all of it is frozen, or no row is frozen). While it is possible that one could try to bypass this limitation by using backward hooks (functions that intercept and modify the gradient before weight updates to zero out the gradients of old tokens). Despite being functional, this technique is highly inefficient. Because the engine would still compute the gradients for all tokens during the backward pass, unnecessarily consuming

computation time and GPU memory, only to end up discarding 99% of the gradient calculations at the end.

To solve this inefficiency, our approach proposes a structural separation combined with Parameter-Efficient Fine-Tuning (PEFT) methods, [9]. Specifically, we introduce a direct architectural modification by implementing an auxiliary, independent embedding matrix dedicated exclusively to the new control tokens.

3.3.3. Model Input

We use a dual conditioning approach, which implies that we indicate the actions in both the encoder and the decoder parts. This presents several advantages; for example, including the actions in the encoder can help the model focus more on certain parts of the sentence over others depending on the predicted action. By passing these actions again as decoder_input_ids, we reinforce the model with the actions it must perform, so that at the time of generation it is already aware of the actions to apply.

For the first subtask, simplification at the sentence level, there are two actions that do not require reasoning to perform the simplification action. The first is the copy action; when we detect that it has been predicted, we simply copy the complex text as the prediction. The second, the delete action, also receives special treatment: it is simply deleted and an empty string is passed as the prediction.

Therefore, we structure the model input as follows: first, the actions we want to apply are set in the encoder, for example, `<Split>`, then we assign the `<Sentence>` token to indicate the beginning of the complex sentence, yielding as result "`<Split> <Sentence>` This review includes a total of 1897 children and 729 adults in 39 trials." Before it enters the decoder, we intervene the input again and reintroduce the actions. Specifically, we pass these control tags as an initial prefix to the decoder (decoder_input_ids), preceded by the start-of-sequence token, which in the case of the Flan-T5 model is `<pad>` and for BART is `<cls>`. This way, the decoder starts its generation process having directly processed the intended instructions for that instance.

Regarding the second subtask, at the document level, the input preparation requires a more complex procedure given the limitations of the model's context window (512 tokens). To process complete documents without truncating relevant information, we apply a sequential clustering and block processing strategy, divided into two phases:

First, we establish "atomic or indivisible units"; this means that if the prediction module outputs a sequence of sentences to be combined, the `<Merge_B>` and `<Merge_I>` segments are immediately clustered. This guarantees that they are never split across different inference blocks, as they must be evaluated jointly. Second, these clusters are iteratively assembled to form larger text blocks. We use the tokenizer to continuously monitor the total block length, ensuring it stays within a safety margin to prevent truncating sentences, which would otherwise degrade performance.

Once the block is formed, the dual conditioning strategy is applied to multiple sequences. For the encoder, we concatenate all the predicted actions for the block followed by the original sentences, where each sentence is preceded by its respective marker. In this way, an input for a three-sentence block could have this structure:

```
<Split><Rephrase><Delete> <Sentence> Complex sentence 1. <Sentence>Complex sentence 2. <Sentence> Complex sentence 3.
```

In parallel, the decoder again receives the continuous string of actions (`<Split><Rephrase><Delete>`) as the initial prompt. Finally, the simplified texts generated

for each of the blocks are sequentially concatenated to reconstruct and obtain the final prediction of the entire document.

3.3.4. Direct Preference Optimization

After the initial supervised fine-tuning phase, a manual inspection of the generated simplifications revealed that the model frequently introduced lexical modifications even when the explicitly provided action was <Copy>. Conversely, it tended to leave the original sentence unchanged when the tag required a rewrite (<Rephrase>). To mitigate this behavior and enforce stricter adherence to the control instructions, we introduced a secondary training phase leveraging DPO. The algorithm requires a dataset formed by triplets (x, y_w, y_l), where x is the input (the complex text along with the actions), y_w is the preferred response (chosen), and y_l is a rejected response. To build this corpus in an automated manner adapted to our task, we designed a generation and evaluation pipeline composed of two phases:

1. Candidate Generation and Evaluation (Dataset Construction). For each training instance, we use the previously trained SFT model to generate four different candidates. To ensure the syntactic diversity of these hypotheses, we apply a generation strategy based on *Diverse Beam Search*, configured with 4 beam groups and a diversity penalty of 1.5. Once the candidates are generated, we proceed to evaluate them to determine which represents the preferred output and which the rejected one. This scoring is based on two pillars:

- **Alignment Model (CRF-BERT):** Each candidate is evaluated using a sequential alignment model (based on BERT and CRF) that scores the quality of the alignment between the original text and the generated simplification. The choice of this model is not arbitrary; it is the exact same alignment system developed and used by the creators of the original Cochrane-auto corpus. Therefore, it serves as the most reliable and appropriate metric for assessing the quality of simplifications within this specialized dataset domain.
- **Control Correction Heuristics:** To resolve the specific problems observed in SFT, we apply strict rules to the base score. We severely penalize those candidates that maintain information when the tag is <Rephrase>, or that alter text when the tag is <Copy>, to force the model to perform the required simplification.

The candidate achieving the highest final score is designated as y_w (chosen response) while the remaining candidates are designated as y_l (rejected responses), when their scores fall below that of the best candidate by a predefined minimum margin.

2. DPO Training. Using the newly constructed preference dataset, we initialize the DPO training phase. We configure two models: the active model, whose weights will be updated, and a reference model, whose weights are kept frozen. Both models are initialized with the weights obtained from the SFT phase, which includes the custom embedding matrix designed to accommodate the control tokens. The training process directly optimizes the policy of the active model relative to the reference model via the standard DPO loss function. Through this alignment phase, the model not only learns the characteristics of a high-quality simplification but also actively identifies the patterns that lead to rejected outputs, thereby effectively calibrating its sensitivity to the action tags.

4. Results and Discussion

We evaluate our models across three settings: the Cochrane-auto validation set, its internal test set, and the CLEF 2026 SimpleText benchmark. Several simplification examples are provided in the Appendix B. To thoroughly assess performance, the evaluation utilizes three key metrics: SARI [10], which measures simplification quality by evaluating the accuracy of words kept, added, or deleted compared to human references; BLEU [11], which evaluates the general similarity and n-gram overlap between

the generated text and the reference texts; and FKGL [12] (Flesch-Kincaid Grade Level), which assesses readability by indicating the US school grade level required to understand the generated text easily.

Systems	SARI↑		BLEU↑		FKLG↓	
	Val	Test	Val	Test	Val	Test
Gemma4:31b-No_Actions	39.32	39.1	12.69	11.72	**9.25**	**9.43**
Gemma4:31b-Action	40.1	38.99	18.77	16.16	10.96	11.15
Flan-T5-No_Actions	34.82	32.04	27.15	28.80	12.83	11.22
Flan-T5-Actions	43.90	39.92	25.59	24.60	11.93	12.54
Flan-T5-DPO_Actions	**45.66**	**42.60**	11.58	11.51	18.48	66.65
BART-No_Actions	34.98	33.21	**35.19**	**30.50**	13.20	14.13
BART-Actions	39.47	39.09	28.15	27.57	13.66	13.77

Table 3
Metrics for Document Simplification Dataset

Systems	SARI↑		BLEU↑		FKLG↓	
	Val	Test	Val	Test	Val	Test
Gemma4:31b-No_Actions	38.17	37.43	5.95	5.67	9.1	9.3
Gemma4:31b-Action	49.38	47.08	10.79	9.23	**7.7**	**7.75**
Flan-T5-No_Actions	38.52	37.10	24.15	23.80	8.71	9.04
Flan-T5-Actions	48.88	46.68	**26.71**	**24.95**	9.66	9.74
Flan-T5-DPO_Actions	**55.88**	**52.78**	10.50	10.10	13.77	12.87
BART-No_Actions	37.45	34.63	25.10	23.49	8.39	8.74
BART-Actions	51.38	48.85	14.06	13.25	11.99	11.98

Table 4
Metrics for Sentence Simplification Dataset

A detailed analysis of Tables 3 and 4, which present the document and sentence-level results for both the validation and test sets, reveals the performance improvement provided by action-based strategies (Actions) compared to the baseline configurations (No_Actions). This qualitative leap is generalized and strongly evident in both prompting methods and classic encoder-decoder architectures, as the action system manages to guide the models much more precisely.

Comparing the two subtasks clearly shows that sentence-level evaluation produces significantly higher metrics compared to the document level. This phenomenon is primarily due to the success of the Copy and Delete actions predictions, which achieve a perfect score of 100 in SARI, whereas errors are not as heavily penalized due to the action model's tendency to predict Rephrase.

Finally, the overall evaluation solidifies Flan-T5 as the superior model; therefore, it was selected for fine-tuning using DPO. This additional training positively enhances the results to the point where the Flan-T5-DPO_Actions variant manages to compete closely with Gemma4. In this way, a more compact model is capable of challenging and even outperforming models with a vastly larger number of parameters. However, an unexpected anomaly was observed in the FKGL metric for the Flan-T5-DPO_Actions model on the test set, which recorded an unusually high score of 66.65, whereas the other configurations typically score between 11 and 13. The underlying cause of this outlier is currently being investigated.

Systems	SARI↑	BLEU↑	FKLG↓
Gemma4:31b-No_Actions	**46.22**	**9.79**	**7.77**
Gemma4:31b-Action	45.77	8.65	9.86
Flan-T5-No_Actions	35.52	3.04	10.77
Flan-T5-Actions	36.27	5.47	12.77
Flan-T5-DPO_Actions	37.56	6.23	12.68
BART-Actions	36.80	10.43	12.17

Table 5
Metrics for 2026 Test Documents

Systems	SARI↑	BLEU↑	FKLG↓
Gemma4:31b-No_Actions	**46.36**	9.91	**7.87**
Gemma4:31b-Action	46.35	11.54	9.17
Flan-T5-No_Actions	29.95	**16.20**	11.93
Flan-T5-Actions	35.72	13.73	11.80
Flan-T5-DPO_Actions	36.90	12.64	12.55
BART-Actions	35.26	11.74	14.29

Table 6
Metrics for 2026 Test Sentences

Evaluating the results of the official 2026 task reveals a significant performance drop in both document and sentence simplification compared to prior evaluations. Currently, the exact cause of this decline remains unknown; it will be necessary to wait for the references to be made public in order to analyze and study the reasons for this drop in metrics in depth.

Unlike to previous phases where Flan-T5 achieved the highest results, this new test scenario highlights that its architecture does not generalize as well. While Gemma4, on the other hand, performs significantly better, taking a clear lead in SARI scores with values nearing 46 in both subtasks.

Nevertheless, a consistent, and positive improvement in performance can still be observed when using actions, which outperform the baseline version. Similarly, the application of DPO training also enhances the results.

5. Future Work

For future work, we propose several lines of research to continue improving the controlled generation of biomedical text efficiently. On the one hand, we plan to evaluate our action injection strategy on newer open models and explore more advanced prompting techniques to refine control in the generation stage. On the other hand, regarding the Flan-T5 based approach, we propose an evolution of our architectural modification. Currently, structural control is achieved by introducing the simplification action discretely. Alternatively, our goal is to investigate a fuzzy integration of the actions. This soft integration (soft prompting/embeddings) would allow the model to learn the nuances of each action more flexibly.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and ERDF "ERDF A way of making Europe" (project PID2024-155948OB-C55), and by the grant PDC2025-164848-I00 funded by MICIU/AEI/10.13039/501100011033. And also, is partially supported by the Spanish Secretaría de Estado

de Digitalización e Inteligencia Artificial (TSI-100930-2023-9).

Declaration on Generative AI

During the preparation of this work, the authors used Gemini to translate portions of the initial drafts from Catalan to English, and to rewrite specific fragments to improve the academic tone and overall readability. Additionally, Gemini was employed to generate the image for the architectural visualization. After using these tools, the authors thoroughly reviewed and edited the generated text and take full responsibility for the final content and visual representations in the manuscript.

References

[1] L. Ermakova, et al., Overview of CLEF 2026 SimpleText track: Simplify scientific texts (and nothing more), in: M. Hagen, et al. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2026), LNCS, Springer-Verlag, 2026.

[2] L. Ermakova, H. Azarbonyad, J. Bakker, G. K. Shahi, B. Vendeville, J. Kamps, CLEF 2026 SimpleText track, in: R. Campos, et al. (Eds.), Advances in Information Retrieval. ECIR 2026, volume 16486 of *Lecture Notes in Computer Science*, Springer, Cham, 2026. URL: https://doi.org/10.1007/978-3-032-21321-1. doi:10.1007/978-3-032-21321-1.

[3] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, Advances in neural information processing systems 36 (2023) 53728–53741.

[4] J. Bakker, et al., Overview of the CLEF 2026 SimpleText task 1: Simplify scientific text, in: G. Faggioli, et al. (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[5] J. Bakker, J. Kamps, Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts, in: M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, R. Stodden (Eds.), Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 41–51. URL: https://aclanthology.org/2024.tsar-1.5/. doi:10.18653/v1/2024.tsar-1.5.

[6] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, Journal of Machine Learning Research 25 (2024) 1–53.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, arXiv e-prints (2019) arXiv:1910.13461. doi:10.48550/arXiv.1910.13461. arXiv:1910.13461.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, arXiv e-prints (2021) arXiv:2106.09685. doi:10.48550/arXiv.2106.09685. arXiv:2106.09685.

[10] W. Xu, C. Napoles, E. Pavlick, Q. Chen, C. Callison-Burch, Optimizing statistical machine translation for text simplification, Transactions of the Association for Computational Linguistics 4 (2016) 401–415. URL: https://aclanthology.org/Q16-1029/. doi:10.1162/tacl_a_00107.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics,

Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[12] J. P. Kincaid, et al., Development and test of a computer readability editing system (CRES), Final Report, 1980. June 1978 through December 1979.

A. Prompts Used for Text Generation

The following section details the prompts used to guide the Gemma 4 model. It includes both the System Prompts and User Prompts for the action-driven controlled generation and the action-free generation.

```
You are SimpleText-LLM, an expert biomedical text simplifier. Based on NIH
↪  guidelines for written health materials.

### AVAILABLE ACTIONS ###
    • <Copy>: Output the exact same sentence without any modifications.
    • <Rephrase>: Rewrite the sentence to simplify its structure and vocabulary,
    ↪  making it easy to understand for lay readers.
    • <Delete>: Omit the sentence entirely.
    • <Split>: Break down a complex sentence into two or more shorter, simple
    ↪  sentences.
    • <Merge_B>: Treat this sentence as the Beginning of a merged thought. Combine
    ↪  its core idea seamlessly with the surrounding context provided.
    • <Merge_I>: Treat this sentence as Inside a merged thought. Synthesize it
    ↪  together with the previous context.

### INPUT FORMAT ###
    • The user will input a source text and a list of its sentences tagged with
    ↪  specific action commands. Process each sentence exactly as tagged and
    ↪  return the final simplified text.

### ESSENTIAL RULES ###
    • Audience Write for readers at about a US 8th-grade level (K8 or smart 13-14
    ↪  year old student).
    • Splitting If a sentence contains more than one idea, split it into shorter
    ↪  sentences.
    • Omission If a sentence is irrelevant to lay readers (for example, detailed
    ↪  measurement methods), delete it.
    • Jargon Replace professional terms with common words. If no plain synonym
    ↪  exists, keep the term once and add a brief parenthetical gloss.
    • Statistics Remove p-values, confidence intervals, and similar numbers unless
    ↪  they are essential for understanding.
    • Voice Use active voice when possible.
    • Pronouns Resolve ambiguous pronouns or other references.
    • Subheadings Remove IMRAD labels, such as 'Background:', 'Introduction:',
    ↪  'METHODS:', 'Results:', 'Discussion:' or integrate them into a full
    ↪  sentence.
    • Output Return only the final adapted string. Do not include conversational
    ↪  filler.
```

Figure 2: System prompt used for the action-driven controlled generation (*action-based*). It defines the available action space and the strict input format.

```
### TASK ###
    - Plain-language sentence adaptation (based on NIH guidelines for written health
    ↪  materials)

### ESSENTIAL RULES ###
    • Audience Write for readers at about a US 8th-grade level (K8 or smart 13-14
    ↪  year old student).
    • Splitting If a sentence contains more than one idea, split it into shorter
    ↪  sentences inside the same pair of single quotes; never merge content from
    ↪  different source items.
    • Omission If a sentence is irrelevant to lay readers (for example, detailed
    ↪  measurement methods), output the empty string ''.
    • Jargon Replace professional terms with common words. If no plain synonym
    ↪  exists, keep the term once and add a brief parenthetical gloss.
    • Statistics Remove p-values, confidence intervals, and similar numbers unless
    ↪  they are essential for understanding.
    • Voice Use active voice when possible.
    • Pronouns Resolve ambiguous pronouns or other references.
    • Subheadings Remove IMRAD labels, such as 'Background:', 'Introduction:',
    ↪  'METHODS:', 'Results:', 'Discussion:' or integrate them into a full
    ↪  sentence.
    • Output Return only the final simplified sentence as string.

### QUICK EXAMPLES ###
    • Simplify 'Myocardial infarction is a leading cause of mortality worldwide.' →
    ↪  'A heart attack is a major cause of death worldwide.'
    • Carry over 'Metabolism is essential for life.' → 'Metabolism is essential for
    ↪  life.'
    • Omit 'Blood pressure was measured with a sphygmomanometer.' → ''
    • Split 'Cardiovascular disease is the leading cause of mortality, and it is
    ↪  influenced by genetics as well as lifestyle.' → 'Heart disease is the
    ↪  leading cause of death. Genetics and lifestyle also influence it.'

### SOURCE TEXT ###
{{Source_text}}

### SOURCE TEXT SPLITED INTO SENTENCES WITH ACTIONS ###
{{Source_text_actions}}
```

Figure 3: User prompt used for the action-driven controlled generation (*action-based*). It provides the global context and details each sentence paired with its specific simplification action.

```
You are SimpleText-LLM, an expert biomedical text simplifier. Based on NIH
↪  guidelines for written health materials.

### ESSENTIAL RULES ###
    • Audience Write for readers at about a US 8th-grade level (K8 or smart 13-14
    ↪  year old student).
    • Language If the input text is in a language other than English, generate the
    ↪  simplified output in that exact same language.
    • Splitting If a sentence contains more than one idea
    • Omission If a sentence is irrelevant to lay readers (for example, detailed
    ↪  measurement methods), output the empty string '' for that element.
    • Jargon Replace professional terms with common words. If no plain synonym
    ↪  exists, keep the term once and add a brief parenthetical gloss.
    • Statistics Remove p-values, confidence intervals, and similar numbers unless
    ↪  they are essential for understanding.
    • Voice Use active voice when possible.
    • Pronouns Resolve ambiguous pronouns or other references.
    • Subheadings Remove IMRAD labels, such as 'Background:', 'Introduction:',
    ↪  'METHODS:', 'Results:', 'Discussion:' or integrate them into a full
    ↪  sentence.
    • Output Return only the final simplified document as string.
```

Figure 4: System prompt used for the autonomous, action-free generation (*action-free baseline*). It provides global simplification guidelines, relying entirely on the model's autonomous capabilities.

```
### TASK ###
    - Plain-language sentence adaptation (based on NIH guidelines for written health
    ↪  materials)

### ESSENTIAL RULES ###
    • Audience Write for readers at about a US 8th-grade level (K8 or smart 13-14
    ↪  year old student).
    • Language If the input text is in a language other than English, generate the
    ↪  simplified output in that exact same language.
    • Splitting If a sentence contains more than one idea, split it into shorter
    ↪  sentences inside the same pair of single quotes; never merge content from
    ↪  different source items.
    • Omission If a sentence is irrelevant to lay readers (for example, detailed
    ↪  measurement methods), output the empty string ''.
    • Jargon Replace professional terms with common words. If no plain synonym
    ↪  exists, keep the term once and add a brief parenthetical gloss.
    • Statistics Remove p-values, confidence intervals, and similar numbers unless
    ↪  they are essential for understanding.
    • Voice Use active voice when possible.
    • Pronouns Resolve ambiguous pronouns or other references.
    • Subheadings Remove IMRAD labels, such as 'Background:', 'Introduction:',
    ↪  'METHODS:', 'Results:', 'Discussion:' or integrate them into a full
    ↪  sentence.
    • Output Return only the final simplified sentence as string.

### QUICK EXAMPLES ###
    • Simplify 'Myocardial infarction is a leading cause of mortality worldwide.' →
    ↪  'A heart attack is a major cause of death worldwide.'
    • Carry over 'Metabolism is essential for life.' → 'Metabolism is essential for
    ↪  life.'
    • Omit 'Blood pressure was measured with a sphygmomanometer.' → ''
    • Split 'Cardiovascular disease is the leading cause of mortality, and it is
    ↪  influenced by genetics as well as lifestyle.' → 'Heart disease is the
    ↪  leading cause of death. Genetics and lifestyle also influence it.'

### SOURCE TEXT ###
{{Source_text}}
```

Figure 5: User prompt used for the autonomous, action-free generation (*action-free baseline*). It introduces the instructions and solely provides the complete source text.

B. Qualitative Analysis of Simplification Outputs

This section provides a qualitative analysis of the models' performance on the 2026 test set. To illustrate this, we provide a direct comparison of our two primary generation strategies. Figure 6 presents a sentence-level simplification (Task 1.1), while Figure 7 showcases a document-level simplification (Task 1.2). Both figures include the original complex text, the sequence of predicted actions, the outputs generated by Gemma 4 (guided by actions) and Flan-T5-DPO (guided by actions).

Complex Text:

3-option trigger PPH diagnosis plus MOTIVE bundle versus visual estimation-based diagnosis plus usual care (direct evidence; 1 study, 170,956 participants) reduces PPH $\geq$ 500 mL (RR 0.48, 95% CI 0.39 to 0.58; high-certainty evidence), and PPH $\geq$ 1000 mL (RR 0.34, 95% CI 0.26 to 0.44; high-certainty).

Predicted Action: `<Rephrase>`

Gemma4:31b-Action:

Using a 3-option trigger to diagnose postpartum hemorrhage (heavy bleeding after childbirth), combined with the MOTIVE care bundle, reduces the amount of blood loss. This approach is more effective than using visual estimation to diagnose bleeding and providing usual care. It reduces the risk of losing 500 mL or more of blood. It also reduces the risk of losing 1,000 mL or more of blood.

Flan-T5-DPO_Actions:

Three Option trigger PPH diagnosis plus MOTIVE bundle versus direct evidence (one study, 170,956 participants) reduces PPH 500 mL and PPH 1000 mL (high-certainty evidence).

Figure 6: Example of Sentence-Level Simplification (Task 1.1) comparing the performance of Gemma 4 and Flan-T5-DPO.

Complex Text:

We identified two non-Cochrane systematic reviews relevant to rehabilitation and the EPOC category of financial arrangements. We excluded one review from the synthesis due to low methodological quality. The other review included 18 primary studies, but only one study reported data relevant to rehabilitation. This study, conducted in a single country (China), compared insured and uninsured children aged birth to 17 years with intellectual disabilities, focusing on the likelihood of receiving at least one rehabilitation service. The review reported that, based on this single study, enrollment in social health insurance, national health insurance, or community-based health insurance schemes may be associated with increased utilization of mental health services, including outpatient rehabilitation. However, due to the complexity and variability of mental health service utilization across contexts, the review authors noted that it was difficult to draw definitive conclusions about the overall impact of insurance enrollment on mental health outcomes and services. We identified two non-Cochrane systematic reviews aligned with the EPOC classification of financial arrangements that addressed rehabilitation. One review including one single-center study on rehabilitation offered low-certainty evidence on insurance schemes and access to rehabilitation services. As a result, this overview cannot provide a clear understanding of the effects of financial arrangements for rehabilitation or offer meaningful signposts to end-users, including policymakers. Future Cochrane overviews in HPSR should broaden their scope to include qualitative and mixed-methods study designs. Robust primary studies and inclusive systematic reviews are urgently needed to guide financial policy decisions that integrate rehabilitation into resilient and equitable health systems. PC, CK, and SN were supported and funded by the Italian Ministry of Health (Ricerca Corrente). The funder played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. Protocol (2025): DOI 10.23736/S1973-9087.24.08833-6.

Predicted Actions: <Delete> → <Delete> → <Delete> → <Rephrase> → <Rephrase> → <Delete> → <Merge_B> → <Merge_I> → <Delete> → <Delete> → <Delete> → <Delete> → <Copy> → <Delete>

Gemma4:31b-Action:

We found two systematic reviews (summaries of many studies) about rehabilitation and how it is paid for. We removed one review because it was not done well. The other review looked at 18 studies, but only one study had useful information about rehabilitation. This study took place in China. It compared children from birth to 17 years old with intellectual disabilities. Some children had insurance and some did not. The study looked at whether children received at least one rehabilitation service. The review found that having social, national, or community health insurance may help people use mental health services more often. This includes outpatient rehabilitation (rehab that does not require an overnight stay in a hospital). However, using mental health services is complex and differs by location. Because of this, the authors said it is hard to be certain about how insurance affects mental health services. We found that the evidence on how insurance affects access to rehab is weak. Therefore, this report cannot give a clear answer on how payment plans affect rehabilitation. It also cannot provide a clear guide for policymakers. Future reports should include different types of study designs. We urgently need better studies to help leaders create fair health systems that include rehabilitation.

Flan-T5-DPO_Actions:

We identified 2 non-Cochrane systematic reviews relevant to rehabilitation and the EPOC category of financial arrangements. We excluded one review from the synthesis due to low methodological quality. The other review included 18 primary studies, but only one study reported data relevant to rehabilitation. This study compared insured and uninsured children aged birth to 17 years with intellectual disabilities, focusing on the likelihood of receiving at least one rehabilitation service. The review reported that enrollment in social health insurance, national health insurance, or community-based health insurance schemes may be associated with increased utilization of mental health services, including outpatient rehabilitation. However, due to the complexity and variability of mental health service utilization across contexts, the review authors noted that it was difficult to draw definitive conclusions about the overall impact of insurance enrollment on mental health outcomes and services. We identified two non-Cochrane systematic reviews aligned with the EPOC category of financial arrangements that addressed rehabilitation. One review including one single-center study on rehabilitation offered low-certainty evidence on insurance schemes and access to rehabilitation services. As a result, this overview cannot provide a clear understanding of the effects of financial arrangements for rehabilitation or offer meaningful signposts to end-users, including policymakers. Future Cochrane overviews in HPSR should broaden their scope to include qualitative and mixed-methods studies. Robust primary studies and inclusive systematic reviews are urgently needed to guide financial policy decisions that integrate rehabilitation into resilient and equitable health systems. PC, CK, and SN were supported and funded by the Italian Ministry of Health (Ricerca Corrente). The funder played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Figure 7: Example of Document-Level Simplification (Task 1.2) illustrating sequential action adherence and model generalization.