Context aware recommender system via LLM in context learning

Fernanda B. de Pinho[1] **, André C. A. Nascimento**[1]

[1]Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

nandasuna.dev@gmail.com, andre.camara@ufrpe.br

Abstract. *Augmentative and Alternative Communication (AAC) provides strategies and tools that support individuals with speech and motor impairments. Among high-tech AAC systems, pictogram-based interfaces are especially widespread. Yet as personal vocabularies grow, navigating large pictogram grids becomes both cognitively and physically demanding, a burden that context-aware pictogram recommendation can alleviate. Prior approaches rely on probabilistic models or supervised classifiers, leaving the use of Large Language Models (LLMs) under the non-tuning paradigm unexplored for AAC pictogram recommendation. We present a two-stage recommender grounded in the GLLM4Rec framework. Stage 1 selects candidate pictograms using temporally decayed frequency combined with contextual modeling. Stage 2 reranks these candidates through a structured prompt that injects the user's history, spatial cluster, and temporal context as few-shot examples, without updating any model parameters. Crucially, the LLM reasons over raw numeric pictogram identifiers, isolating pattern-based inference from semantic knowledge. In a preliminary evaluation on logs from a real-world AAC application, we adopt a rolling temporal validation protocol across two users (55 folds, 1,770 test interactions). LLM reranking improves over the Stage 1 probabilistic baseline on the same data and protocol, reaching Recall@5* $= 0.556 \pm 0.216$ *and MRR@10* $= 0.348 \pm 0.172$ *(vs.* 0.496 *and* 0.297*). The gains concentrate at the top of the ranking (*$K \leq 5$*), whereas the probabilistic baseline recovers more items at* $K = 10$*—exposing a precision–coverage trade-off that is particularly relevant to AAC interface design.*

Resumo. *Comunicação Aumentativa e Alternativa (CAA) oferece estratégias e ferramentas que apoiam indivíduos com deficiências de fala e motoras. Entre os sistemas de CAA de alta tecnologia, as interfaces baseadas em pictogramas são especialmente difundidas. No entanto, à medida que os vocabulários pessoais crescem, navegar por grandes grades de pictogramas torna-se exigente tanto do ponto de vista cognitivo quanto físico, um ônus que a recomendação de pictogramas sensível ao contexto pode atenuar. Abordagens anteriores baseiam-se em modelos probabilísticos ou classificadores supervisionados, deixando inexplorado o uso de Modelos de Linguagem de Grande Escala (LLMs) sob o paradigma sem ajuste fino (non-tuning) para a recomendação de pictogramas em CAA. Apresentamos um recomendador em dois estágios fundamentado no arcabouço GLLM4Rec. O Estágio 1 seleciona pictogramas candidatos utilizando frequência com decaimento temporal combinada à modelagem contextual. O*

Estágio 2 reordena esses candidatos por meio de um prompt estruturado que injeta o histórico do usuário, o agrupamento espacial e o contexto temporal como exemplos few-shot, sem atualizar nenhum parâmetro do modelo. Crucialmente, o LLM raciocina sobre identificadores numéricos brutos dos pictogramas, isolando a inferência baseada em padrões do conhecimento semântico. Em uma avaliação preliminar com registros de uma aplicação de CAA do mundo real, adotamos um protocolo de validação temporal deslizante (rolling) com dois usuários (55 folds, 1,770 interações de teste). A reordenação por LLM supera a linha de base probabilística do Estágio 1 nos mesmos dados e protocolo, alcançando Recall@5 $= 0.556 \pm 0.216$ *e MRR@10* $= 0.348 \pm 0.172$ *(vs.* 0.496 *e* 0.297*). Os ganhos concentram-se no topo do ranking (*$K \leq 5$*), ao passo que a linha de base probabilística recupera mais itens em* $K = 10$*—revelando um trade-off entre precisão e cobertura particularmente relevante para o projeto de interfaces de CAA.*

1. Introduction

AAC systems support people with speech, language, or motor impairments through a spectrum of strategies, ranging from low-tech symbol boards to high-tech software applications (Alfuraih et al. 2024). Among these, pictogram-based AAC is one of the most widely adopted modalities: users communicate by selecting sequences of visual cards, each pairing a symbolic image with a short text that denotes a word, activity, or concept. Such systems typically rely on grid-based interfaces, as in the Picture Exchange Communication System (PECS), where users browse many symbols to assemble messages or express basic needs (Srinivasan et al. 2022).

As personal vocabularies grow to hundreds of symbols, this browsing imposes substantial cognitive and physical demands. The burden falls hardest on early communicators, neurodivergent users, and those with significant motor challenges, for whom searching becomes a major obstacle to spontaneous communication (Zastudil et al. 2024). Context-aware recommendation can reduce this overhead by predicting the card a user is likely to select next based on temporal patterns, location, and interaction history. Indeed, prior work has shown that time of day, day of week, and GPS cluster are strong predictors of communicative intent (Garcia et al. 2016; Neamtu et al. 2019; Evangeline and Moorthy 2024).

A path that, to the best of our knowledge, remains unexplored in the AAC context is the use of Large Language Models (LLMs) within a *non-tuning* paradigm (Wu et al. 2023), in which the model is queried through carefully crafted prompts rather than retrained on domain data. Because it requires no fine-tuning and hence no GPU infrastructure, this approach adapts through in-context learning and could, in principle, transfer across users. This raises question for AAC deployment: can a non-tuning LLM meaningfully reduce the search burden faced by real AAC users — without requiring GPU infrastructure or user-specific model training?

This paper presents exploratory results from applying the non-tuning GLLM4Rec framework to a sample of a real AAC interaction dataset, and makes three contributions:

(i) The first AAC recommendation system built on the GLLM4Rec non-tuning paradigm, in which candidate selection and LLM-based reranking are kept as separate, reproducible stages.

(ii) A two-stage pipeline that reproduces the probabilistic preprocessing of (Anonymous 2026) in Stage 1 and applies a five-component structured prompt with few-shot examples drawn from the user's own history in Stage 2.
(iii) An empirical evaluation on a sample of real AAC application logs, conducted under the same rolling-window protocol as the reference baseline, which reveals a precision–coverage trade-off between LLM and probabilistic ranking with practical implications for AAC interface design.

The remainder of the paper is organized as follows. Section 2 provides the necessary background and surveys related literature. Section 3 details the proposed methodology. Sections 4–5 describe the experimental setup and report the empirical results. Finally, Section 6 offers concluding remarks.

2. Related Work

AAC encompasses strategies and tools designed to support or replace natural speech for individuals affected by conditions such as autism spectrum disorder, cerebral palsy, or acquired brain injury (Alfuraih et al. 2024). High-tech AAC systems bring the PECS paradigm to digital mobile devices, e.g., Livox (Neamtu et al. 2019), Proloquo2Go (AssistiveWare 2025), and Tobii Dynavox (Tobii Dynavox 2019), presenting pictogram grids through which users construct messages one card at a time. While the PECS paradigm (Srinivasan et al. 2022) has been shown to develop requesting skills across contexts, growing vocabularies place mounting demands on users, particularly those with motor limitations (Zastudil et al. 2024). Adaptive recommendation addresses this by surfacing the most contextually relevant cards before the user begins searching.

Research on next-pictogram prediction can be grouped by learning paradigm: tuning-based approaches, which require training on domain data, and rule-based or supervised approaches, which incorporate contextual signals without deep architectures.

Among tuning-based approaches, PictoBERT (Pereira et al. 2022) adapts BERT (Bidirectional Encoder Representations from Transformers) to next-pictogram prediction, and PrAACT (Pereira et al. 2024) extends this toward user-adaptive vocabularies. Both require training on domain data and target next-item prediction rather than ranking, emphasizing phrase-composition usage patterns.

Complementing these, rule-based and supervised approaches have addressed context-awareness without relying on LLMs. Garcia et al. [2016] showed that GPS-based profile adjustment improves AAC interaction efficiency, and Evangeline and Moorthy [2024] proposed a supervised context-aware system based on linguistic normalization and semantic retrieval. Even so, research combining context-aware modeling with LLM-based reasoning in AAC remains scarce.

On the LLM side, Wu et al. [2023] classify LLM-based recommendation into *tuning* (parameter updates on domain data) and *non-tuning* (inference-only, prompt-driven) paradigms. The non-tuning approach exploits pre-trained knowledge via zero-shot, few-shot, and chain-of-thought prompting, requiring no training infrastructure or labeled data beyond what fits in the context window. A structured non-tuning recommendation prompt comprises five components: Role Instruction, Few-shot Examples, Task Description, Behavior Injection, and Format Indicator. This paradigm faces two notable challenges. The

first is out-of-corpus generation: the LLM may hallucinate items absent from the candidate set, necessitating a grounding step to filter responses to valid items only. The second concerns transferability: although LLMs achieve competitive performance in domains where items carry natural-language descriptions, movies, books, and products, it remains unclear how well they extend to bare numeric identifiers, where the model cannot draw on pre-trained item semantics. To the best of our knowledge, no prior work has applied non-tuning LLM recommendation, and, specifically, the GLLM4Rec framework, to AAC.

Table 1 consolidates the discussed works along for dimensions: the underlying method, the target task, the learning paradigm (tuning, non-tuning, supervised, or rule-based), and whether the approach leverages LLMs and targets AAC. Prior pictogram-recommendation work concentrates on the setting under the tuning paradigm (Pereira et al. 2022; Pereira et al. 2024), while few rely on rule-based or supervised pipelines without LLMs (Garcia et al. 2016; Evangeline and Moorthy 2024). Conversely, non-tuning LLM recommendation frameworks, such as GLLM4Rec (Wu et al. 2023), have been validated in general-purpose domains but not in AAC. Our work occupies the intersection left open by this landscape: a non-tuning, LLM-based approach to pictogram recommendation in AAC.

Table 1. Comparison of related work on next-pictogram and LLM-based recommendation, contrasted with our contribution.

Work	Method	Paradigm	LLM	AAC
(Pereira et al. 2022)	BERT for next-item prediction	Tuning	No	Yes
(Pereira et al. 2024)	BERT with user-adaptive vocabulary	Tuning	No	Yes
(Garcia et al. 2016)	GPS-based profile adjustment	Rule-based	No	Yes
(Evangeline and Moorthy 2024)	Linguistic normalization + semantic retrieval	Supervised	No	Yes
(Wu et al. 2023)	Prompt-driven LLM recommendation	Non-tuning	Yes	No
Ours	**GLLM4Rec applied to AAC**	**Non-tuning**	**Yes**	**Yes**

3. Method

(Anonymous 2026) propose a context-aware pictogram recommendation system for AAC that models spatial context through DBSCAN clustering of GPS coordinates and temporal context through fuzzy Gaussian membership functions over the hour of day, combining both with temporally decayed usage frequency into a single ranking score, under a rolling temporal validation protocol. They evaluate this probabilistic model against a Random Forest classifier trained on the same contextual features, using real interaction logs from a deployed AAC application comprising 671,022 interactions from 27 users with a considerably larger and sparser vocabulary than ours (median of 216 unique pictograms per user, reaching 463 at the 90th percentile). Their results show that the probabilistic approach consistently outperforms the Random Forest across all metrics (Recall@5 = 0.280 vs. 0.161; MRR@10 = 0.173 vs. 0.097), establishing it as a strong, interpretable baseline for

AAC pictogram recommendation. Our Stage 1 (Eqs. 1–3) reproduces this probabilistic formulation — temporally decayed frequency combined with fuzzy temporal and spatial modeling — which is why we treat its output as the natural non-LLM baseline against which we measure the effect of Stage 2's LLM reranking. The system follows a two-stage pipeline (Figure 1). Stage 1 selects candidate pictograms from the training window through probabilistic scoring alone, with no LLM involved. Stage 2 then reranks those candidates using a structured prompt.

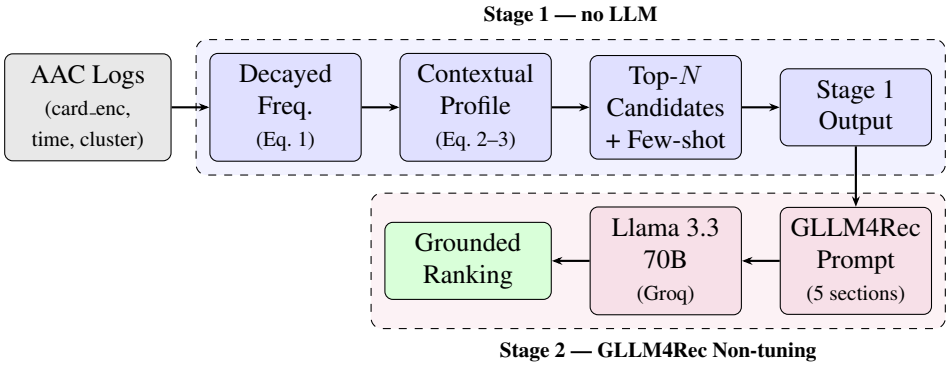

Figure 1. Two-stage pipeline of the GLLM4Rec Non-tuning system. Stage 1 runs entirely in Python with no API calls; Stage 2 submits a structured five-component prompt to Llama 3.3 70B via the Groq API and grounds the output to the Stage 1 candidate set.

3.1. Stage 1: Candidate Selection

In Stage 1, every past interaction contributes to the frequency estimate with a weight that decays exponentially over time, following a half-life function so that recent interactions carry more influence than older ones. Formally, let $\mathcal{H}$ denote the set of interactions in the current training window, p_i the pictogram selected at interaction i, t_i its timestamp, and t_{now} the current inference time. The weight of interaction i is

$$w_i; =; e^{-\lambda,(t_{\text{now}}-t_i)}, \qquad \lambda = \frac{\ln 2}{t_{1/2}}, \tag{1}$$

where the decay rate λ is derived from a predefined half-life $t_{1/2}$.

Each interaction also carries spatial and temporal context. Its location is obtained by applying DBSCAN clustering to the geographical coordinates and is then encoded as a one-hot vector $\mathbf{g} \in 0, 1^C$ over the C recurring clusters. Its temporal context is encoded by a fuzzy Gaussian membership function over the hour of day: the day is partitioned into K overlapping periods, each a Gaussian centered at a reference hour c_k with spread σ_k, yielding a membership vector $\mathbf{t} \in [0, 1]^K$ with $t_k = \exp!\big(-(h - c_k)^2/2\sigma_k^2\big)$ ($K = 6$, $c_k \in 0, 4, 8, 12, 16, 20$, $\sigma_k = 3$,h for all k).

Let c be the current GPS cluster and $k = \arg\max_k \mu_k(h_{\text{now}})$ the dominant temporal period of the current hour. Treating location and time as conditionally independent given the pictogram (naive Bayes), the probability of selecting candidate $i \in H$ is the posterior

$$P\big(i \mid c, k\big) = \frac{P(i), P\big(c \mid i\big), P\big(k \mid i\big)}{\sum_{j \in H} P(j), P\big(c \mid j\big), P\big(k \mid j\big)}, \tag{2}$$

where the prior is the recency-weighted usage of each pictogram, $P(i) = \hat{f}(i)$ (Eq. 1), and the location and temporal likelihoods are its decayed-frequency-weighted profiles over past occurrences $H_i \subseteq H$:

$$P\big(c \mid i\big) = \frac{\sum_{n \in H_i} w_n, \mathbf{1}[c_n = c]}{\sum_{n \in H_i} w_n}, \qquad P\big(k \mid i\big) = \frac{\sum_{n \in H_i} w_n, \mu_k(h_n)}{\sum_{n \in H_i} w_n}, \tag{3}$$

with decay weights w_n as in Eq. 1.

The top-N items are passed to Stage 2. The probabilistic baseline in our comparison is precisely this Stage 1 ordering-Eq. 2 without any LLM reranking. Stage 1 additionally extracts F few-shot examples from the most recent training interactions; each example records the previous

3.1. Stage 2: GLLM4Rec Prompt and LLM Ranking

Stage 2 assembles a five-component prompt following (Wu et al. 2023):

(1) **Role Instruction:** casts the LLM as an AAC specialist with expertise in pictogram usage patterns.
(2) **Few-shot Examples:** provide F in-context examples from the user's own history. Each pair consists of a contextual state, period, cluster, day, and a recent sequence of pictogram IDs, along with the chosen pictogram, allowing the LLM to infer personal patterns without parameter updates.
(3) **Task Description:** asks the LLM to rank the Stage 1 candidates from most to least likely, weighing temporal frequency, context, and the few-shot evidence.
(4) **Behavior Injection:** supplies the current state, dominant period with its full membership vector, cluster, day of week, recent pictogram ID history, and each candidate's decayed frequency.
(5) **Format Indicator:** requires a JSON list of integers to enable reliable parsing.

Figure 2 shows the complete reranking prompt used in our experiments, instantiated for a representative test interaction. The prompt injects the decayed frequency $\hat{\mathbf{f}}(\mathbf{i})$ as a compact proxy for the full posterior, while the spatial and temporal context is conveyed separately through the behavior-injection fields.

3.2. Output Grounding and Language Model

The LLM's response is parsed as a JSON integer list; if parsing fails, any integers mentioned in the text are extracted instead. Values outside the candidate set are removed, duplicates are dropped, and the original order is preserved. If grounding yields an empty result, the system falls back to the Stage 1 ordering, and the fallback rate is tracked as a reliability diagnostic.

Stage 2 uses **Llama 3.3 70B** (Meta AI 2024) via the Groq API, chosen for its low-latency free-tier access, which makes the experiment reproducible without GPU resources. The temperature is set to 0.1 for near-deterministic outputs. No fine-tuning, prompt-tuning, or adapter training is applied at any stage.

```
## Section 1: Role Instruction
You are an expert in AAC (Augmentative and Alternative Communication)
with deep knowledge of pictogram usage patterns and user behavior.
Your task is to predict which pictogram a user will select next based
on their usage history, temporal context, and location.

## Section 2: Few-shot Examples (In-Context Learning)
Example 1:
  Period: morning | Cluster: 2 | Day: Monday
  History: [42, 17, 8, 33, 42]
  Selected: 42
Example 2:
  Period: afternoon | Cluster: 1 | Day: Tuesday
  History: [17, 42, 55, 42, 17]
  Selected: 17
Example 3:
  Period: night | Cluster: 1 | Day: Monday
  History: [8, 33, 42, 17, 8]
  Selected: 8
[... F=5 examples total ...]

## Section 3: Task Description
Considering ONLY the candidates listed below, rank them from most to
least likely for this user, taking into account:
  1. Historical frequency (with recent temporal decay)
  2. Contextual patterns (time of day, location cluster, day of week)
  3. History examples (Few-shot learning)
Return the order from most to least likely.

## Section 4: Behavior Injection (Current Context)
Current Context:
  Period: morning (dawn=0.05, morning=0.87, night=0.08, evening=0.00)
  Cluster: 2  |  Day: Friday
  Recent history: [42, 17, 42, 8, 42]
Candidates and Decayed Frequencies:
  42: 0.85  |  17: 0.47  |  8: 0.31  |  33: 0.22  |  55: 0.18
  [... N=20 candidates total ...]

## Section 5: Format Indicator
Return ONLY a list with the candidate IDs inside brackets [],
from most to least likely. Do not include explanations.
  Expected format: [42, 17, 8, 33, 55]
```

Figure 2. Complete reranking prompt (GLLM4Rec, $F{=}5$, $N{=}20$) instantiated for a representative test interaction. Card identifiers are raw integers; no pictogram labels are provided to the LLM.

4. Experimental Setup

4.1. Dataset and Protocol

We use real interaction logs from a commercial AAC application, from a sample of 2 users, namely $user_a$ (38 folds, 824 interactions, complete evaluation) and $user_b$ (17 folds, 946 interactions). The combined results span 55 folds and 1,770 test interactions. Each record is defined by the pictogram ID, DBSCAN cluster ($\varepsilon = 400\,\text{m}$), as well as the $\mathbf{g}$ and $\mathbf{t}$ contextual vectors. The pictogram vocabulary is 55 items per user; each user has 2 recurring spatial clusters.

We adopt the rolling temporal validation protocol to prevent data leakage (Ji et al. 2023), in which the first fold initializes training. Then, for each test fold $fold_s$, all model components are rebuilt from folds 1 through $s{-}1$. Predictions are generated sequentially; after each prediction, the ground truth (pictogram ID) is appended to the rolling context. Stage 1 and GLLM4Rec rankings are evaluated simultaneously under identical conditions.

4.2. Metrics and Configuration

Ground truth p^* is the pictogram the user actually selected at test s interaction, as recorded in the AAC application log immediately after the context used for prediction. Recall@K and MRR@K are reported for $K \in \{3, 5, 10\}$, aggregated as mean $\pm$ std across folds:

$$\text{Recall@}K = \mathbf{1}[p^* \in \text{ranking}_{:K}] \tag{4}$$

$$\text{MRR@}K = \frac{1}{\text{rank}(p^*)} \text{ if } \text{rank}(p^*) \leq K, \text{ else } 0 \tag{5}$$

The main experiment uses $N = 20$ candidates and $K = 5$ few-shot examples (FS=5 C=20). Fixed parameters: half-life $= 7$ days, $\sigma = 3$ h, *history_window* $= 5$, temperature $= 0.1$.

5. Results and Discussion

To motivate the quantitative results that follow, Figure 3 first illustrates a concrete reranking example.

Context: morning | cluster 2 | Friday | recent history: `[42, 17, 42, 8, 42]`

Rank	Stage 1 (Prob)	Stage 2 (LLM)
1	9 *ball*	9 *ball*
2	83 *stand up*	71 *pizza* ↑
3	80 *soft taco*	**26 *choc. donut*** ↑↑ GT
4	60 *monsters*	76 *recliner*
5	71 *pizza*	22 *cheeseburger*
—	26 *choc. donut* (× absent)	—

Figure 3. Schematic AAC board recommendations for a representative interaction (morning, cluster 2, Friday, recent history `[42,17,42,8,42]`). The Stage 1 probabilistic ranking places the ground truth (card 26, *chocolate donut*) outside the top-5 (bottom row, shaded red). Stage 2 LLM reranking promotes it to position 3 (shaded green). This corresponds to Example A in Table 3.

Table 2 reports Recall and MRR for GLLM4Rec Non-tuning and the Stage 1 probabilistic baseline. The two systems share data, folds, and candidate sets, so the difference between them isolates the effect of LLM reranking. For context, we also list the reference values from (Anonymous 2026); because that study draws on a much larger user base and vocabulary, those rows are not a like-for-like comparison, and we do not treat exceeding them as evidence of method superiority (see Limitations).

On identical folds, GLLM4Rec outperforms the probabilistic baseline at R@3 (+0.110), R@5 (+0.060), MRR@3 (+0.077), MRR@5 (+0.067), and MRR@10 (+0.051). This gain holds even though the LLM receives no textual descriptions, only raw pictogram IDs, which suggests that few-shot examples, combined with temporal and spatial context, give the model enough signal to capture personal usage patterns through in-context learning, without any access to pictogram semantics.

Table 2. Recall@K (↑) and MRR@K (↑), mean ± std for $user_a$ and $user_b$ data. Best of the two comparable systems per metric in bold. Reference rows from a larger experiment († (Anonymous 2026), considering 27 users) are shown for context only and are *not* directly comparable.

Model	R@3	R@5	R@10	MRR@3	MRR@5	MRR@10
GLLM4Rec (ours)	**0.467±0.216**	**0.556±0.216**	0.616±0.229	**0.319±0.174**	**0.340±0.170**	**0.348±0.172**
Probabilistic	0.357±0.244	0.496±0.233	**0.676±0.224**	0.242±0.178	0.273±0.166	0.297±0.159
Reference values (different scale, not comparable):						
Prob. Baseline†	0.208	0.280	0.400	0.140	0.157	0.173
Random Forest†	0.111	0.161	0.254	0.073	0.084	0.097

The ordering reverses at R@10, where the baseline leads (0.676 vs. 0.616, $\Delta = -0.060$), revealing a precision–coverage trade-off: the LLM concentrates probability mass on fewer items, improving top-3 and top-5 precision at the expense of broader coverage. The practical implication depends on the interface. AAC systems typically surface only 3–5 suggestions, a regime that favors the LLM; scenarios that call for a wider safety net, for instance, a caregiver reviewing all plausible options, may still be better served by the probabilistic ordering.

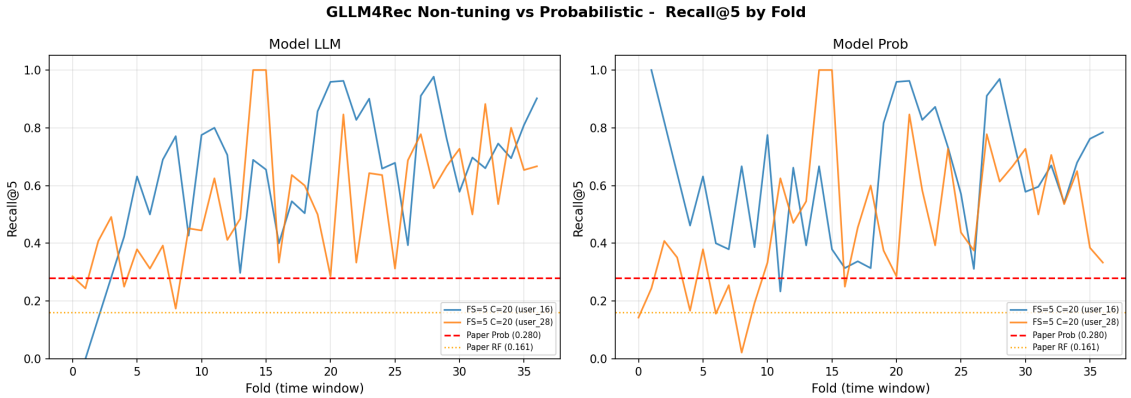

Figure 4. Recall@5 per fold for GLLM4Rec (left) and Probabilistic (right) on $user_a$ (orange) and $user_b$ (blue). Dashed lines: Probabilistic baseline (red, 0.280) and Random Forest (orange, 0.161) from (Anonymous 2026).

Figure 4 tracks Recall@5 across folds. The high variance ($\sigma \approx 0.22$–0.24) reflects the natural irregularity of personal AAC usage: neither model is stable from fold to fold. The two curves largely overlap, with the LLM gaining ground in higher-activity periods. Coefficients of variation exceed 40% for both models, which underscores why single-fold snapshots are insufficient and why the rolling aggregate is necessary.

To complement these aggregate metrics, Table 3 presents three representative reranking examples from the evaluation of user_a, each illustrating a distinct behavior of the LLM relative to the Stage 1 probabilistic ordering.

These three cases capture the qualitatively distinct outcomes observed across folds. In Example A, the LLM recovers a ground-truth item from outside the probabilistic top-5 and promotes it to position 3, the case where reranking adds the most value. In Example B, both models place the ground truth at the same rank, but the LLM rearranges the surrounding candidates, reflecting a different confidence distribution. In Example C, the LLM reproduces the Stage 1 ranking exactly, indicating that when the probabilistic signal is already strong, the LLM leaves it intact rather than introducing noise. Together,

Table 3. Reranking examples (user_a, F=5, N=20). GT = ground truth; ↑ = promoted; ↓ = demoted relative to probabilistic ranking. Labels shown for readability only — the LLM receives raw card_enc integers.

Rank	Probabilistic	LLM	Chg.
Example A — GT: 26 *chocolate donut* (Fold 0, Rec. 0). GT absent from probabilistic top-5; LLM promotes it to position 3.			
1	9 *ball*	9 *ball*	—
2	83 *stand up*	71 *pizza*	↑
3	80 *soft taco*	**26 *choc. donut*** (GT)	↑
4	60 *monsters*	76 *recliner*	—
5	71 *pizza*	22 *cheeseburger*	—
Example B — GT: 33 *crumb donut* (Fold 2, Rec. 0). Both rank GT at position 2; LLM reorders remaining candidates.			
1	53 *ipad*	71 *pizza*	↑
2	**33 *crumb donut*** (GT)	**33 *crumb donut*** (GT)	—
3	71 *pizza*	53 *ipad*	↓
4	15 *breakfast bowl*	15 *breakfast bowl*	—
5	101 *yes or no?*	26 *choc. donut*	—
Example C — GT: 53 *ipad* (Fold 1, Rec. 2). LLM confirms probabilistic order.			
1	**53 *ipad*** (GT)	**53 *ipad*** (GT)	—
2	2 *3*	2 *3*	—
3	71 *pizza*	71 *pizza*	—
4	15 *breakfast bowl*	15 *breakfast bowl*	—
5	81 *stand*	81 *stand*	—

these cases support the aggregate finding that LLM reranking sharpens precision at small cutoffs without degrading the cases that Stage 1 already handles well.

These findings, however, rest on a deliberately narrow evaluation, and several constraints bound what can be concluded from them. (i) $user_b$ is evaluated on only 17 of its 83 folds, owing to computational and time constraints. (ii) The ablation covers a single configuration, limited by API rate constraints. (iii) Most importantly, the comparison with (Anonymous 2026) is not like-for-like, as detailed in Section 3, our setting uses 2 users and a 55-item vocabulary against their 27 users and considerably larger vocabulary (median 216 items/user), so the reference rows in Table 2 situate our results in scale, not in direct superiority. The controlled comparison remains GLLM4Rec reranking versus our own Stage 1 baseline. (iv) Finally, the LLM fallback rate, the fraction of calls for which grounding yielded no valid items, will be reported once the full evaluation completes.

6. Conclusion

We presented an application of the GLLM4Rec non-tuning paradigm to AAC pictogram recommendation. The two-stage system first selects candidates through temporally decayed

frequency and fuzzy Gaussian modeling (Stage 1), then reranks them via a structured five-component prompt submitted to Llama 3.3 70B without any parameter update (Stage 2).

In a preliminary evaluation across 55 folds and 1,770 test interactions from sample users, LLM reranking improves over the Stage 1 probabilistic baseline on identical folds, reaching Recall@5 $= 0.556 \pm 0.216$ and MRR@10 $= 0.348 \pm 0.172$ (vs. 0.496 and 0.297). The LLM attains this from raw numeric card identifiers alone, offering preliminary evidence that pattern-based LLM recommendation is viable even in low-semantic-content domains.

The results also expose a precision–coverage trade-off: the LLM ranks the correct pictogram higher at $K \leq 5$, whereas the probabilistic model recovers more items at $K = 10$. Since most AAC interfaces display only 3–5 suggestions, this regime favors the LLM, and combining both signals, with broad probabilistic retrieval followed by LLM reranking, is a natural next step.

Several directions follow from this work: completing the evaluation on both users under the full 83-fold protocol; running the six-configuration ablation to characterize hyperparameter sensitivity; reporting the LLM fallback rate; introducing pictogram text labels (Phase 2) to measure how much semantic knowledge contributes beyond the pattern-based baseline established here; most important, conducting a live deployment study with real AAC users to evaluate the system's practical usability and interface impact beyond retrospective log analysis.

References

[Alfuraih et al. 2024] Alfuraih, R. K., Almalki, N. S., and AlNemary, F. M. (2024). Effectiveness of picture exchange communication system in developing requesting skills for children with multiple disabilities. *Frontiers in Psychology*, 15.

[Anonymous 2026] Anonymous (2026). Context-aware recommendation system for augmentative and alternative communication. Manuscript submitted for publication; details omitted for double-blind review.

[AssistiveWare 2025] AssistiveWare (2025). Proloquo2go: Symbol-based AAC application.

[Evangeline and Moorthy 2024] Evangeline, S. B. and Moorthy, A. D. (2024). Improving communication for people with speech disabilities: Exploring context-aware recommendation systems in assistive technology. In *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*, pages 1–6.

[Garcia et al. 2016] Garcia, L. F., de Oliveira, L. C., and de Matos, D. M. (2016). Evaluating pictogram prediction in a location-aware augmentative and alternative communication system. *Assistive Technology*, 28(2):83–92.

[Ji et al. 2023] Ji, Y., Sun, A., Zhang, J., and Li, C. (2023). A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems*, 41(3).

[Meta AI 2024] Meta AI (2024). Llama 3 model card.

[Neamtu et al. 2019] Neamtu, R., Camara, A., Pereira, C., and Ferreira, R. (2019). Using artificial intelligence for augmentative alternative communication for children with disabilities. In *Lecture Notes in Computer Science*, volume 11746, pages 234–243.

[Pereira et al. 2022] Pereira, J. A., Macêdo, D., Zanchettin, C., de Oliveira, A. L. I., and do Nascimento Fidalgo, R. (2022). PictoBERT: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202:117231.

[Pereira et al. 2024] Pereira, J. A., Zanchettin, C., and do Nascimento Fidalgo, R. (2024). PrAACT: Predictive augmentative and alternative communication with transformers. *Expert Systems with Applications*, 240:122417.

[Srinivasan et al. 2022] Srinivasan, S., Patel, S., Khade, A., Bedi, G., Mohite, J., Sen, A., and Poovaiah, R. (2022). Efficacy of a novel augmentative and alternative communication system in promoting requesting skills in young children with autism spectrum disorder in india: A pilot study. *Autism & Developmental Language Impairments*, 7.

[Tobii Dynavox 2019] Tobii Dynavox (2019). My tobii dynavox.

[Wu et al. 2023] Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., Xiong, H., and Chen, E. (2023). A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*.

[Zastudil et al. 2024] Zastudil, C., Holyfield, C., Smith, J. A., Nguyen, H., and MacNeil, S. (2024). Predictive anchoring: A novel interaction to support contextualized suggestions for grid displays.