Exploring Second Language Learning Patterns in Duolingo Data

Chuangyu Ding

December 2025

1 Introduction

This report investigates patterns of second language acquisition using learner interaction data from Duolingo, a large-scale language learning platform. Rather than focusing on individual items or short-term accuracy, the analysis takes a longitudinal perspective, asking how learners' performance evolves over time and whether this evolution differs systematically across linguistic structures and language pairs.

Large-scale learning platforms provide a unique opportunity to study second language learning in naturalistic settings. The Duolingo dataset contains millions of interactions from learners with diverse backgrounds, across multiple target languages and grammatical categories. Such data make it possible to move beyond isolated examples and examine broad regularities in learning behavior. At the same time, the data are noisy and highly heterogeneous: learners differ in proficiency and study habits, practice is irregularly spaced, and item difficulty varies widely. These characteristics make it difficult to draw clear conclusions from simple summaries of accuracy alone.

To address this, we take two complementary approaches.

First, we analyze the dataset directly and examine aggregate patterns observed in learner responses. By grouping interactions by grammatical category, or language pair, we obtain a descriptive view of how learners perform across different linguistic contexts. This approach reflects what can be seen directly from the data, without imposing strong modeling assumptions, but it is also sensitive to noise and confounding.

Second, we adopt a modeling-based perspective. We train a hazard-based survival model that treats forgetting as a time-to-event process, where the event corresponds to an incorrect response after a period of correct recall. The model estimates the probability of forgetting at each time interval, conditional on the item having been remembered up to that point. By combining these conditional probabilities, we obtain full forgetting curves that describe how recall probability decays over time under controlled conditions. By fixing most contextual features and varying only specific factors, such as grammatical category or language pair, the model allows us to isolate and visualize differences in forgetting dynamics

that are difficult to observe directly in the raw data. In this sense, the trained model provides an alternative lens on the dataset, complementing the descriptive analysis rather than replacing it.

By comparing these two perspectives—direct observation from raw data and patterns revealed through a trained model—this report aims to shed light on systematic differences in second language learning behavior. In particular, we explore whether the model recovers intuitively plausible distinctions between grammatical constructions and language pairs, and how these distinctions relate to patterns that can (or cannot) be observed directly in the data.

2 Results from Raw Data

2.1 Overall performance differences across languages

We begin with a simple descriptive analysis of overall learner performance in the Duolingo dataset. As a coarse measure of learning success, we use the perfection rate, defined as the proportion of sessions in which all items were answered correctly.

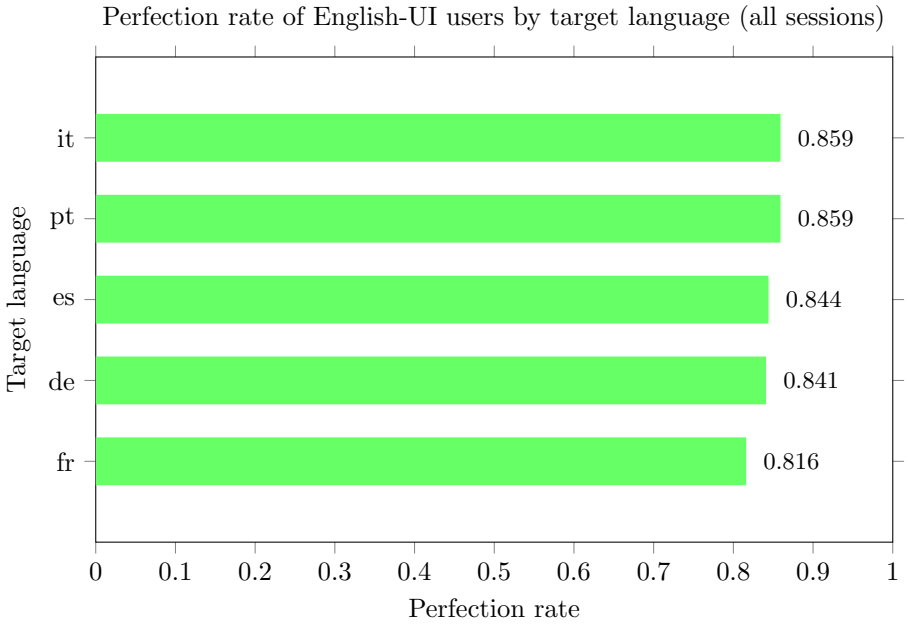

Figure 1: Perfection rate of English-UI users by target language (all sessions)

The Figure 1 shows the perfection rate of English-interface users across different target languages (Italian, Portuguese, Spanish, German, and French). Although all values are relatively high, clear differences can be observed. Italian and Portuguese exhibit the highest perfection rates, followed by Spanish and

German, while French shows the lowest overall performance among the five languages. This suggests that, even for learners with the same interface language, target languages differ in how easily learners achieve error-free sessions.

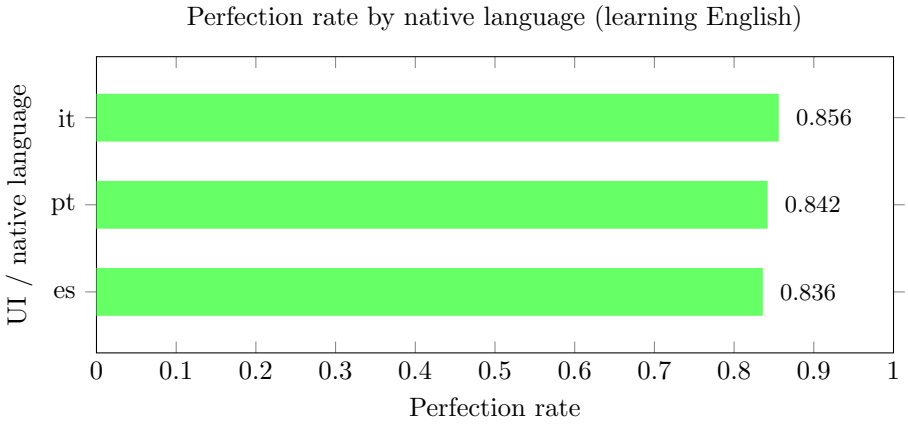

Figure 2: Perfection rate by native language for learners of English

To complement this view, the figure 2reverses the perspective and examines learners with different native or interface languages who are learning English. A similar pattern emerges: learners with Italian and Portuguese interfaces achieve higher perfection rates than Spanish-interface learners.

Taken together, these results suggest that second language learning performance varies systematically across language pairs. These differences may reflect a combination of factors, including typological similarity between languages, transfer effects from the native language, differences in grammatical complexity, or properties of the learning content itself. Importantly, these patterns are already visible through simple aggregation, without any modeling or temporal analysis.

2.2 Performance differences across grammatical points

We further zoom in from language-level comparisons to grammatical structure–level differences. The Table below shows the perfection rate of English-interface learners studying French, grouped by grammatical point. Each bar represents the proportion of sessions in which all items associated with a given grammatical category were answered correctly.

A clear overall pattern emerges from the results. Grammatical points for which French adds little beyond what English already expresses show the highest performance, while those that require learners to encode distinctions that English typically leaves implicit tend to be more difficult. The difference does not primarily lie in conceptual understanding, but in the number of decisions learners must make and the amount of grammatical information they must actively track when producing a sentence.

Table 1: Perfection rates by grammatical point for English-speaking learners of French.

Grammatical point	**Perfection rate (%)**
Future	65.8
Passé composé	68.9
Passive voice	70.5
Person agreement	71.4
Subjunctive (present / past)	72.2
Imparfait	74.0
Reflexive pronouns	74.1
Purpose clauses	76.2
Interrogatives	78.8
Temporal subordinate clauses	78.8
Relative pronouns	81.6
Gender agreement	81.6
Number agreement	82.9
Formulaic expressions	84.1
Causal subordinate clauses	89.6

2.2.1 Minimal added requirements: direct mapping and fixed expressions

This contrast is especially clear for causal subordinate clauses and formulaic expressions, which appear among the highest-performing categories. In causal clauses such as **Je reste à la maison parce qu'il pleut** , French closely mirrors English in both structure and usage. The causal relationship is explicitly marked in both languages, and no additional grammatical adjustments are required. Formulaic expressions such as **bien sûr** or **d'accord** similarly impose almost no internal requirements. They are learned and used as fixed expressions, without the need to analyze verb forms, agreement, or sentence structure. In both cases, French does not require learners to perform additional grammatical computations beyond recognizing the intended meaning, which contributes to the high stability of learner performance.

Number and gender agreement also show relatively strong performance, even though English does not require comparable marking. The reason appears to be that the additional information French demands is local, visible, and repetitive. Learners must mark plurality or gender on articles and adjectives, but these markings usually affect only one element at a time and are strongly reinforced by frequent exposure and correction. As a result, although French introduces extra grammatical requirements here, they do not substantially increase cognitive load during sentence production.

2.2.2 Moderate added requirements: obligatory form choice within familiar structures

Performance begins to decline for categories such as relative pronouns, interrogatives, and temporal or purpose clauses, where French introduces additional constraints on form choice that are largely absent in English. Relative constructions illustrate this particularly well. In both languages, relative clauses link information across parts of a sentence, as in **the book that I talked about** or **the person I spoke to**. English typically relies on a single flexible pattern, often using that or omitting the relative marker altogether. French, however, requires learners to select a specific form based on the underlying grammatical relationship, even when the intended meaning is the same. For example, learners must produce **le livre dont je parle** or **la personne à laquelle j'ai parlé**, choices that depend on the preposition associated with the verb rather than on meaning alone. This forces learners to identify the verb's argument structure and track prepositional relationships before selecting the correct relative form.

As a result, sentence production in French involves an additional analytical step that English does not require. Learners cannot rely on a default relative pattern but must instead pause to determine how elements are connected syntactically. This extra layer of decision-making increases processing demands and helps explain why performance in these categories is lower, despite the underlying concept of relativization being familiar to English speakers.

A similar pattern appears in interrogative structures. Although both languages offer multiple ways to form questions, French learners must choose among several conventionalized constructions, such as intonation-based questions or the use of est-ce que. These options are not interchangeable in all contexts, and selecting among them requires attention to form rather than meaning alone.

Temporal and purpose clauses further increase this burden, as French often requires specific verb forms triggered by particular conjunctions. English typically expresses the same meanings without visible changes to verb morphology, allowing learners to rely on semantic interpretation alone. In French, learners must recognize the type of clause and apply a corresponding grammatical transformation, adding another layer of processing.

2.2.3 High added requirements: mandatory interpretation and morphological encoding

The lowest performance is observed in verb tense categories, subjunctive forms, passive constructions, and person agreement, all of which place substantial additional demands on learners compared to English. Past tense usage is a particularly clear example. In sentences such as **Il pleuvait quand je suis arrivé**, French obliges learners to distinguish between ongoing background situations and completed events and to select verb forms accordingly. English can often express the same contrast more flexibly, without requiring learners to commit to a rigid grammatical classification.

In French, however, this initial choice between tense categories is only the

first step. Learners must also determine which auxiliary verb to use when forming the past tense and whether past participle agreement is required. For instance, producing **je suis arrivé** requires selecting *être* rather than *avoir*, a choice that depends on verb class rather than meaning alone. This choice then introduces an additional requirement: the past participle must agree in gender and number with the subject, as in **elle est arrivée**. In contrast, verbs formed with *avoir*, such as *j'ai mangé*, do not generally require agreement, unless further syntactic conditions are met.

As a result, learners are not only required to interpret the temporal structure of an event, but must also manage a cascade of form-related decisions that English largely avoids. English past tense forms do not involve auxiliary selection based on verb type, nor do they require agreement marking on the verb. French therefore adds multiple mandatory decision points—choosing the tense, selecting the auxiliary, and determining whether agreement applies—before a sentence can be correctly produced. This accumulation of interpretive and morphological requirements substantially increases processing load, helping to explain the lower performance observed in these categories.

Subjunctive constructions present a similar challenge. French requires specific verb forms in contexts expressing necessity, desire, or uncertainty, even when English conveys the same meanings without any visible verb change. Learners must not only recognize the triggering context but also recall and apply a distinct morphological form, despite the lack of a clear parallel in English. Passive voice and person agreement further illustrate how French increases grammatical obligations. French verbs consistently change according to the subject, and passive constructions involve explicit structural marking. English, by contrast, uses minimal verb inflection and relies more heavily on word order and auxiliary verbs, reducing the need for continuous grammatical monitoring.

Taken together, these results suggest that differences in learner performance are best explained by the cumulative grammatical requirements imposed by French rather than by abstract rule difficulty. When French demands that learners encode distinctions that English leaves unmarked, requires obligatory form changes, or forces multiple decisions to be made simultaneously, performance declines. Conversely, when meaning can be expressed without additional grammatical computation, learners perform reliably.

3 Results from Hazard model

While these descriptive analyses reveal clear differences across grammatical points, they are inherently limited by their static nature. The perfection rate collapses learner behavior across time and does not distinguish between grammatical structures that are initially difficult but quickly stabilized, and those that appear easy in early sessions but remain fragile over longer intervals. Moreover, static aggregates conflate multiple sources of variation, including learner proficiency, exposure frequency, and item repetition. As a result, it is difficult to tell whether observed differences reflect genuine differences in learning dynamics

or simply differences in practice patterns within the dataset. To address these limitations, we move beyond aggregate performance and introduce a time-aware modeling approach. Rather than asking which grammatical points have higher overall success rates, we ask how learner performance evolves as a function of time since last exposure, and whether different grammatical structures exhibit systematically different temporal patterns.

3.1 Model: a discrete-time hazard formulation

To model learning dynamics over time, we adopt a discrete-time hazard-based formulation. The key idea is to treat performance breakdown as a time-to-event process, where the event corresponds to the first session in which an error occurs after a period of successful recall.

3.1.1 Episode-level representation

Each learning episode begins at time $t = 0$, corresponding to a session in which all items are answered correctly. Let T denote the (random) time until the next incorrect session. If no error is observed before the end of the data sequence, the episode is treated as right-censored.

To handle irregularly spaced timestamps, time is discretized into K exponentially spaced bins,

$$0 = \tau_0 < \tau_1 < \cdots < \tau_K,$$

where τ_k represents the upper boundary of time bin k.

3.1.2 Discrete-time hazard

For an episode with feature vector x, we define the discrete-time hazard at bin k as

$$h_k(x) = \Pr\big(T \in (\tau_{k-1}, \tau_k] \mid T > \tau_{k-1}, x\big),$$

that is, the conditional probability that the episode ends in bin k, given that it has survived up to the beginning of that bin.

The model learns a function

$$h_k(x) = f_\theta(x, k),$$

where f_θ is implemented as a binary classifier that takes both episode-level features and the time-bin index as input.

3.1.3 Survival and forgetting curves

Given the predicted hazards $\{h_k(x)\}_{k=1}^{K}$, we can derive the survival function

$$S_k(x) = \Pr(T > \tau_k \mid x) = \prod_{j=1}^{k} \big(1 - h_j(x)\big),$$

which represents the probability that performance remains error-free beyond time τ_k.

We define the corresponding forgetting function as

$$F_k(x) = 1 - S_k(x),$$

which describes how the probability of failure accumulates over time. Plotting $F_k(x)$ as a function of τ_k yields a forgetting curve for the given episode features.

3.1.4 Model training

To train the model, each episode is expanded into a person-period representation. For each time bin k in which the episode is at risk, we create a training example with label

$$y_k = \begin{cases} 1, & \text{if the episode ends in bin } k, \\ 0, & \text{otherwise.} \end{cases}$$

The model parameters θ are learned by minimizing the binary cross-entropy loss over all person-period observations. In practice, we implement f_θ using CatBoost, which allows us to handle heterogeneous features such as learner history, grammatical categories, language pairs, and temporal context.

3.1.5 Use of the model for analysis

By fixing most features and varying only one factor at a time (e.g., grammatical category or language pair), we derive controlled forgetting curves that summarize how learning dynamics differ across linguistic conditions.

This formulation allows us to separate time effects from contextual variation and provides a principled way to compare learning behavior across different dimensions of second language acquisition.

3.2 Forgetting dynamics revealed by the hazard model

3.2.1 Interpretation of within-learner item-specific forgetting dynamics

Even after controlling for learner identity, substantial variation remains in how forgetting risk accumulates over time across different items. By fixing the learner and varying only item-level features, the hazard-based curves demonstrate that items induce systematically different temporal profiles of vulnerability to error. This indicates that forgetting behavior cannot be fully attributed to stable learner characteristics alone; instead, properties intrinsic to the learning material play an independent and measurable role in shaping when errors are most likely to occur.

A prominent feature of these curves is the steep increase in forgetting risk shortly after a successful review. This rapid early rise is consistent with classic

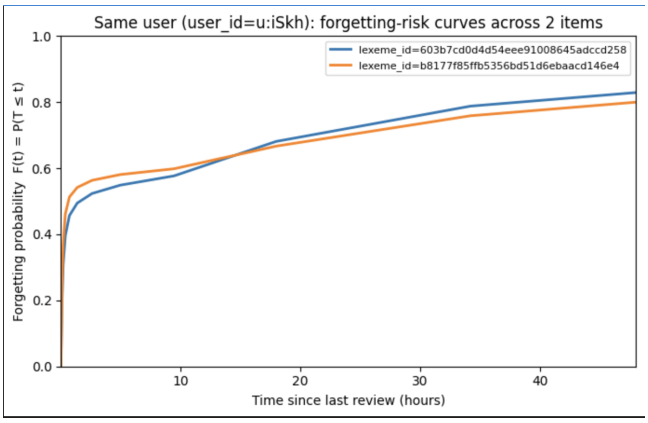

Figure 3: Forgetting-risk curves of one user

findings in memory psychology, which distinguish an initial phase of fast decay followed by a slower, more gradual decline. In the present setting, this effect is further amplified by our operational definition of forgetting: an event is recorded as soon as a learner produces any incorrect response within a session, even if other responses remain correct. As a result, the early portion of the curve reflects not only memory fragility but also a stringent success criterion, under which partial degradation of knowledge is sufficient to trigger a failure event. Importantly, differences between items are most pronounced in this early phase, suggesting that some items are inherently more prone to rapid destabilization immediately after review.

Taken together, these results highlight that forgetting is not characterized by a single time scale or difficulty parameter. Even within the same learner, different items exhibit distinct combinations of early fragility and long-term stability. The hazard-based formulation makes these temporal differences explicit, revealing item-specific forgetting dynamics that would be obscured by aggregate accuracy measures or static notions of item difficulty.

3.2.2 Grammar-category effects on forgetting dynamics

We now examine forgetting dynamics aggregated at the level of grammatical categories. Figure 4 shows the estimated cumulative probability of failure over time for different categories, obtained by evaluating the hazard model on category-specific feature configurations.

Across categories, forgetting risk rises rapidly shortly after review and then transitions into a slower phase of accumulation. However, the magnitude and persistence of this risk differ substantially by category. Verb form distinctions constitute the most stable category, exhibiting the lowest forgetting risk and the slowest accumulation over time. This robustness can be attributed to the fact that verb forms such as infinitives, participles, and gerunds are largely

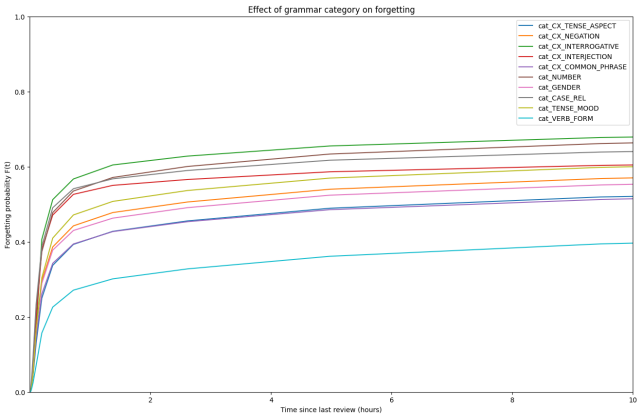

Figure 4: Forgetting-risk curves of different grammar categories

identifiable through local morphological cues and are often learned as part of fixed or semi-fixed constructions. Once acquired, their use typically involves recognizing a surface form and matching it to a limited set of options, rather than computing abstract syntactic relations or coordinating multiple grammatical elements across a sentence. As a result, successful performance on verb form items relies primarily on lexicalized or form-based representations that remain relatively stable after initial learning.

In contrast, interrogative constructions, number marking, and case-related categories exhibit consistently higher forgetting risk because successful performance depends on online structural decisions rather than surface form recognition. Interrogatives require learners to choose among multiple non-interchangeable constructions whose appropriateness is governed by syntactic and pragmatic constraints, making errors likely even when the communicative intent is clear. Number agreement, despite its conceptual simplicity, imposes distributed requirements across several sentence elements, so that a single missed agreement can trigger failure. Case-related and relational categories place the greatest burden on learners, as correct forms depend entirely on identifying abstract syntactic roles and mapping them onto language-specific realizations, a process that remains fragile and susceptible to cross-linguistic interference. Under a stringent definition of forgetting that treats any incorrect response as an event, these structurally demanding categories therefore show elevated and persistent forgetting risk over time.

3.2.3 Language pair effects on forgetting dynamics

We next examine forgetting dynamics aggregated at the level of language pairs. Figure 5 shows mean cumulative forgetting curves for different source–target language combinations, with shaded regions indicating 95% bootstrap confidence intervals. All curves are evaluated on the same time scale and derived

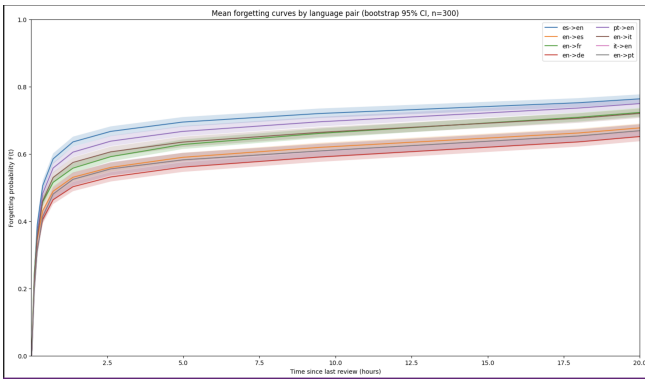

Figure 5: Forgetting-risk curves of different grammar categories

from the same hazard model, allowing direct comparison of how forgetting risk accumulates across language pairs.

Across all language pairs, the overall shape of the curves is remarkably consistent: forgetting risk increases rapidly shortly after review and then transitions into a slower, more gradual phase. This shared structure mirrors the general temporal pattern established in the earlier analyses and suggests that the basic dynamics of forgetting are largely invariant across language pairs. In other words, differences between language pairs do not manifest as qualitatively different decay regimes, but rather as systematic shifts within a common temporal framework.

Within this shared structure, however, clear and persistent differences emerge in the level of forgetting risk. Language pairs such as Spanish→English and Portuguese→English exhibit consistently higher forgetting probabilities across the entire observed interval, whereas pairs such as English→German and English→Portuguese remain systematically lower. Importantly, these separations are already visible in the early phase and are maintained over time, indicating that language-pair effects act as a relatively stable modifier of forgetting risk rather than being confined to a specific retention window.

The relative ordering of curves also remains largely unchanged as time progresses. Language pairs that show higher early vulnerability continue to accumulate risk at comparable or slightly higher rates in the long term, while those with lower early risk remain comparatively robust. This stability suggests that language-pair effects can be understood as global shifts in the baseline level of vulnerability, rather than as differences in the speed of transition between early and late forgetting phases.

From a linguistic perspective, these patterns are consistent with the idea that cross-linguistic similarity and structural alignment modulate learning stability. Language pairs that differ substantially in morphosyntactic structure or lexical realization appear to impose a higher overall burden on learners, resulting in elevated forgetting risk throughout the retention interval. Conversely,

pairs with greater overlap in grammatical categories or surface forms show lower cumulative risk, even though they follow the same general forgetting trajectory.

4 Discussion

A practical implication of the present modeling approach concerns the design of adaptive review schedules in language learning applications. Many widely used vocabulary learning systems are still based on variants of classical spaced repetition algorithms, such as the Leitner system, in which items are promoted or demoted between a small number of boxes associated with fixed review intervals. While these methods introduce adaptivity at the level of individual success or failure, they implicitly assume that all items share the same underlying forgetting dynamics and that difficulty can be adequately captured by coarse performance history alone. As a result, they do not explicitly account for systematic differences between items, grammatical categories, or language pairs, nor for the fact that some types of knowledge are intrinsically more fragile over time than others.

The hazard-based forgetting model studied in this report suggests an alternative, more fine-grained approach. By estimating a continuous forgetting probability as a function of time since last successful recall and conditioning this estimate on item- and language-related features, the model provides a principled measure of when an item is at elevated risk of failure. In such a framework, review scheduling need not rely on predefined box transitions or fixed interval doubling. Instead, a review can be triggered whenever the predicted cumulative forgetting probability exceeds a threshold,

$$\tau^*(x;\theta) = \min\{\tau_k : F(\tau_k \mid x) \geq \theta\},$$

where $F(\tau_k \mid x)$ denotes the model-predicted probability of failure by time τ_k since the last successful recall.

Importantly, the threshold θ itself need not be fixed globally. It can be interpreted as a control parameter governing the trade-off between memory retention and practice efficiency, and may therefore vary across learners and over time. For instance, a user-specific and time-varying threshold $\theta_{u,t}$ can be updated online based on recent learning outcomes,

$$\theta_{u,t+1} = \Pi_{(0,1)}\big(\theta_{u,t} + \eta\,(\widehat{R}_{u,t} - R_0)\big),$$

where $\widehat{R}_{u,t}$ summarizes observed or predicted recall performance over a recent window, R_0 is a target retention level, η is a step size, and $\Pi_{(0,1)}(\cdot)$ denotes projection onto the interval $(0,1)$. Under this feedback rule, thresholds decrease when observed performance falls below the target, triggering earlier reviews, and increase when performance is stable, reducing unnecessary practice.

More generally, threshold selection can be formulated as an optimization problem that trades off expected retention against review cost, for example by

minimizing the total number of review sessions subject to a constraint on expected recall performance. From this perspective, the trained forgetting model does not prescribe a single optimal schedule. Instead, it defines a space of adaptive review policies, ranging from fixed global thresholds to fully personalized and dynamically updated ones, whose performance can be empirically evaluated and compared using held-out data.

This view highlights how time-aware forgetting models can serve not only as analytical tools for understanding learning dynamics, but also as decision-level interfaces for the design of flexible and linguistically informed spaced repetition algorithms.