Surprisal and Crossword Clues difficulty: Evaluating Linguistic Processing between LLMs and Humans

Tommaso Iaquinta[1,*,†], Asya Zanollo[2,3,†], Achille Fusco[3,4,†], Kamyar Zeinalipour[1,†] and Cristiano Chesi[2,3,†]

[1] Università degli Studi di Siena (UNISI), Via Roma 56, 53100 Siena, Italy
[2] University School for Advanced Studies IUSS Pavia, Piazza della Vittoria 15, 27100 Pavia, Italy
[3] Laboratory for Neurocognition, Epistemology, and Theoretical Syntax - NeTS-IUSS Pavia
[4] Università degli Studi di Firenze, Piazza S. Marco 4, 50121 Firenze, Italy

Abstract

Crossword clue difficulty is traditionally judged by human setters, leaving automated puzzle generators without an objective yard-stick. We model difficulty as the *Surprisal* of the answer given the clue, estimating it with token probabilities from large language models. Comparing three models three causal LLMs-Llama-3-8B, Llama-2-7B, and Ita-GPT-2-121M. with 60 human solvers on 160 hand-balanced clues, Surprisal correlates negatively with accuracy (r = −0.62 for nominal clues). These results show that language-model Surprisal captures some of the cognitive load humans experience and that language-specific training and model scale both matter; the metric therefore enables adaptive crossword generation and provides a new test-bed for probing the alignment between human and model linguistic processing.

Keywords

surprisal, llm, gpt, crossword, education, linguistic games, puzzle, Crossword difficulty

1. Introduction

Crossword (CW) puzzles are among the most popular language games, captivating millions through newspapers, mobile apps, voice assistants, and even televised competitions [1, 2]. The enduring appeal of crosswords across formats stems from the careful calibration of clue difficulty, which can range from accessible, beginner-friendly prompts to highly intricate, expert-level challenges.

Despite advancements in automated puzzle generation, state-of-the-art systems like Dr. Fill [3] and the Berkeley Crossword Solver [1], while capable of outperforming many human solvers, still lack a reliable, objective measure to assess the challenge posed by the clues they generate. Traditional heuristics, such as clue length, grid density, historical solve statistics, and letter frequency, only weakly reflect human solving effort, failing to capture the subtle syntactic nuances, semantic leaps, and playful misdirection intrinsic to crossword difficulty [4].

Table 1
Linguistic properties for "piante che forniscono frutti per spremute, aranci" (plants that provide fruits for juice – orange trees).

Microcategory	bareNP:rel
Macrocategory	nominal
Accuracy	0.526
RTs ($\log_{10}$)	4.214
Surprisal	5.207

Table 2
Linguistic properties for "i mobili con le grucce, armadi" (the furniture with hangers – wardrobes).

Microcategory	defDP
Macrocategory	nominal
Accuracy	1.0
RTs ($\log_{10}$)	3.973
Surprisal	3.926

Meanwhile, psycholinguistics provides a promising, information-theoretic perspective [5] through the concept of *Surprisal*, defined as the negative logarithm of the probability of a word given its context. This metric reliably predicts human cognitive effort, correlating strongly with eye-tracking and self-paced reading measures [6, 7, 8, 9]. Leveraging modern large

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

[†] † These authors contributed equally.

✉ tommaso.iaquinta@unisi.it (T. Iaquinta); asya.zanollo@iusspavia.it (A. Zanollo); achille.fusco@iusspavia.it (A. Fusco); kamyar.zeinalipour2@unisi.it (K. Zeinalipour); cristiano.chesi@iusspavia.it (C. Chesi)

🌐 https://tommyiaq.me (T. Iaquinta); https://www.iusspavia.it/it/rubrica/asya-zanollo/ (A. Zanollo); https://github.com/achille-fusco (A. Fusco); https://kamyarzeinalipour.github.io/ (K. Zeinalipour); https://github.com/cristianochesi (C. Chesi)

iD 0009-0009-1262-6768 (T. Iaquinta); 0009-0001-3987-4843 (A. Zanollo); 0000-0002-5389-8884 (A. Fusco); 0009-0006-3014-2511 (K. Zeinalipour); 0000-0003-1935-1348 (C. Chesi) language models (LLMs), which naturally compute token probabilities, Surprisal becomes readily accessible. Recent studies further emphasize the influence of model scale and training domain on the alignment between model-derived Surprisal and human cognitive patterns [10, 11]. Notably, despite its potential, Surprisal has yet to be explored specifically as a metric for crossword difficulty.

Given the increasing prevalence and sophistication of automated CW generation systems, there is now a pressing need for a principled, data-driven metric capable of accurately gauging puzzle difficulty. Such a metric could facilitate adaptive tutoring tools, ensure fairness in online competitions, and provide richer psycholinguistic experimentation frameworks. In this paper, we propose and investigate token-level Surprisal, delivered by LLMs, as an innovative and robust candidate for objectively quantifying crossword puzzle difficulty. The current research represents the first attempt to apply the surprisal metric in the context of crossword puzzles, marking a novel approach to defining crossword difficulty through computational linguistics measures. To guide our investigation and evaluate the viability of token-level Surprisal as an effective measure, we formulate a central research question, summarized clearly below. From this overarching inquiry, we derive four specific, actionable research questions (RQs) designed to systematically unpack the predictive capabilities of Surprisal.

Main Question Can Surprisal computed by modern LLMs serve as a reliable, fine-grained predictor of how hard humans find a crossword clue?

To unpack this question, we address four research questions (RQs):

1. **RQ1**: To what extent does token-level Surprisal correlate with human-measured difficulty (accuracy and solving time) for clue–answer pairs?
2. **RQ2**: How do model family and size—Llama-3, Llama-2, Ita-GPT-2—affect predictive power?
3. **RQ3**: Which sentence-concatenation strategy (*clue cioè answer*, copular rewrites, topic–comment, etc.) yields the most reliable Surprisal estimate for each clue category?
4. **RQ4**: Can Surprisal-based grading drive an adaptive crossword-generation pipeline that targets specific solver skill levels?

Contributions

- **Fine-grained linguistic taxonomy & benchmark** —A curated set of 160 Italian clues spanning 20 syntactic categories, solved by 60 natives (2 880 judgments), provides accuracy and solving-time gold standards.
- **Surprisal estimation framework** —Five generic concatenation rules turn any clue–answer pair into a well-formed sentence with the answer in final position; open-source code computes multi-token Surprisal from any causal LM.
- **Empirical findings** —(i) Surprisal correlates strongly and negatively with accuracy (best $r = -0.57$) but only weakly with raw solving times—stronger after log transform. (ii) Ita-GPT-2 and Llama-3 outperform larger, non-specialised models. (iii) Predictive strength is category-dependent; metalinguistic and copular clues remain challenging. (iv) Picking the right concatenation rule per category boosts correlation by up to 0.15 r-points.
- **Recipe for adaptive generation** —A demonstrator workflow assigns category-specific Surprisal thresholds, selects clues at desired difficulty, and sketches integration with full-grid generation.
- **Open resources** —All data, annotation scripts, Surprisal code, and analysis notebooks are released to foster reproducibility and future research on cognitively informed puzzle generation.

Table 1 and 2 distils our guiding idea into one side-by-side snapshot. For two carefully matched clues the answer that GPT-2 finds more surprising (*aranci*) is also the one humans solve more slowly and less accurately, previewing our central claim: words that a transformer language model finds less predictable also slow humans down and trigger more errors. Numeric values are means; RTs are log-transformed.

Headline results Surprisal from Ita-GPT-2 and Llama-3 explains more than half the variance in human accuracy for nominal clues, evidencing a robust link between probabilistic prediction and perceived difficulty. The general guiding framework adopted in this study is exemplified in Table 5.

Paper layout Section 3.1 presents the dataset and taxonomy; Section 3.2 details concatenation rules and Surprisal computation; Section 5 reports human and model results; Section 6 applies the findings to adaptive generation; Section 7 concludes.

2. Related Work

2.1. Surprisal as a Psycholinguistic Metric

In recent years, Surprisal has been employed to evaluate LLMs performances in psycholinguistic studies, in correlation with online processing measures taken from corpora, like Reading Times (RTs) [12, 13, 14, 15, 16], and Event-Related Potentials (ERPs) [17, 18]. A key issue in comparing LLMs linguistic competence and Human competence consist in understanding at which human-like degree LLMs represent Natural Language (NL). Human linguistic competence does not rely on probability alone [19, 20] and it is structure-driven, in contrast to LLMs data-driven training [21, 22] and tend to underestimate syntax with respect to human processing, in virtue of their different mechanism of learning and understanding [13] In this scenario Surprisal represents a 'neutral' measure which can account also for differences deriving from various linguistic sources in a probabilistic framework. [23] The understanding of the difference between language in models and humans remains a central and extremely relevant point in all the comparative studies and in the analysis of the results. Following the line of research described above, we aim at investigating whether the same correlation - between processing difficulty and Surprisal values – holds also for CW clue-answer pairs. No prior work supplies a token-level, psycholinguistically grounded metric for per-clue difficulty. We import LLMs Surprisal, validate it against 60 human solvers, and show how it plugs into adaptive generation workflows.

2.2. LLMs and Cognitive Alignment

Large language models (LLMs) supply token probabilities out of the box, enabling fine-grained surprisal estimates. Layer-wise activations in GPT-, BERT- and Llama-style models predict fMRI and MEG responses to naturalistic text with striking accuracy [24, 25]. Model scale and training data modulate that alignment: bigger is not always better for eye-movement predictivity, whereas deeper layers in larger models often map best to slower neural signals [26]. Tokenisation also matters: sub-word splits can blur the link between model surprise and human lexical access; aggregating sub-tokens or using morphologically aware tokenisers improves fit [27]. By comparing three Italian-capable LLMs (Ita-GPT-2, Llama-2, Llama-3), we contribute new evidence on how family, size and training regime affect cognitive alignment in a puzzle-solving context.

2.3. Crossword Solving & Generation

AI interest in crosswords began with the probabilistic solver Proverb [28] and the web–based WebCrow system [29]. Dr. Fill later recast clue filling as a single-weighted CSP [3], while subsequent systems introduced neural rerankers and hybrid IR–NLP pipelines [30]. Large language models now push solver accuracy above 90 % on *New York Times* puzzles [31].

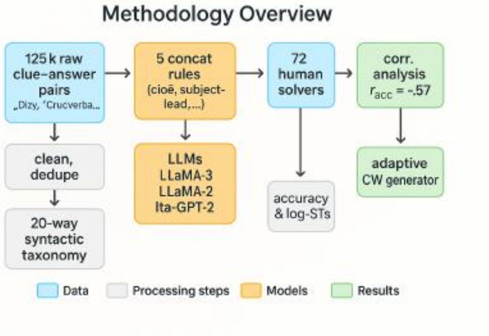
Methodology Overview
125 k raw clue–answer pairs
„Dizy, "Crucverba...
clean, dedupe
20-way syntactic taxonomy
5 concat rules (cioè, subject-lead,...)
LLMs LLaMA-3 LLaMA-2 Ita-GPT-2
72 human solvers
accuracy & log-STs
corr. analysis r_{acc} = -.57
adaptive CW generator
Data
Processing steps
Models
Results

Figure 1: Methodology overview. Colour-coded blocks show data (blue), processing (grey), models (orange) and results (green); arrows trace the workflow.

Grid construction and clue writing pose a different challenge. Early generators searched word-list constraints for Italian crosswords and beyond [32, 33], later adapting to Malay [34], Spanish [35] and Indian languages for education [36]. More recently, Zeinalipour and collaborators have spearheaded a multilingual, education-oriented research programme: Italian educational grids [37], the *WebCrow* French solver [38], Arabic generators—including both clue-focused ArabIcros [39] and a text-to-puzzle pipeline [40]—, a Turkish generator [41], and the Clue-Instruct dataset for pedagogy-centred clues [42]. Together, these works illustrate a fast-growing ecosystem of LLM-driven solvers and generators that operate across languages and educational settings.

Despite this progress, no prior work proposes an objective, cognitively grounded difficulty metric. Published systems label puzzles informally ("easy", "hard") or rely on surface heuristics (grid density, answer length). By linking LLM-derived surprisal to human accuracy and solving times, our study closes this evaluation gap and enables adaptive puzzle generation across languages.

3. Methodology

Our four–step pipeline (Fig.1) is: (1) scrape, clean, and tag approximately 125 000 Italian clue–answer pairs into 20 syntactic categories; (2) turn each pair into a sentence via five lightweight templates and compute answer–level surprisal with Llama − 3, Llama − 2, and Ita − GPT − 2; (3) obtain a human baseline from 60 native speakers solving 160 balanced clues, yielding accuracy and log-transformed solving times; and (4) correlate surprisal with those measures and use category-specific thresholds to power an adaptive crossword generator.

3.1. Data and Preprocessing

To evaluate the difficulty of crossword puzzles, we leveraged a comprehensive collection of Italian CW clues and answers. The sources of the clues-answer pairs are both internet sites that release solutions for CW clues, https://www.dizy.com/ and https://www.cruciverba.it/, that we scraped through apposite scripts. And also *pdf* versions of famous Italian CW papers like *Settimana Enigmistica* and *Repubblica*, that we suitably converted to clue-answer pairs. The various sources where than cleaned, merged and the duplicates were removed. This dataset consists of 125,600 entries that correspond to unique clue-answer pairs. It includes clues related to different domains, such as history, geography, literature, and pop culture. The dataset under investigation contains a diverse array of linguistic features, including grammatical structures, syntactic patterns, and lexical elements.

3.2. Linguistic Classification

The dataset of Italian clue-answer pairs has been syntactically analysed and different clue constructions have been categorized with the aim of investigating what kinds of structural operations can be applied to derive CW clues from well-formed sentences. Being based on the *syntax* of clue-answer pairs, the classification presented is language-dependent on Italian.

In general terms, clues have been initially distinguished into clausal and non-clausal structures depending on the presence or absence of an inflected verb in the matrix clause and, secondly, non-clausal clues can be articulated in different structures varying in the nature of their heads: Noun Phrases (NP), Determiner Phrases (DP), Prepositional Phrases (PP), Adjectival Phrases (AdjP) and Adverbial Phrases (AdvP).

Clausal clues, on the other side, represent syntactically relevant items in virtue of the presence of an inflected verb in the matrix clause and they can be categorized on that basis. Indeed these include clauses with verbal or nominal predicates (i.e. copular sentences), and relative clauses. These main categories differentiate internally, and some subcategories can be accordingly defined. Once some significant syntactic structures have been outlined we can proceed with the classification of our unstructured corpus. It is important to highlight that the proposed categorization is based on the generative grammar approach thus, in the computation of classification rules we considered the difference between the parser (dependencies) and our hierarchical categorization. Categories have been identified on the basis of the type of head, and then further specified by additional features (if any) like in the case of DP which can be of type definite or indefinite.

First of all a qualitative data analysis has been carried out using Regular Expressions (RegEx) and Part-of-Speech (PoS) tagging that have been employed to extract examples of different syntactic constructions and see whether their distribution was significant or not. The extraction has then been improved using the python library spaCy [43] and the dataset has been parsed using the \nlp function which allows us to identify the head node of each clue. We identified 20 pertinent clue typologies for our experiment summarized in Table 3. For further details see the original work on CW linguistic analysis [44].

4. Experimental Setup

The research question that guides our experiment is whether the probability of LLMs token can be used to predict the difficulty of a clue-answer pair. The underlying assumption is that Surprisal, as a complexity metric, correlates to online measures of processing difficulty. For this reason, we can consider Surprisal in relation to measures that we took as index of the difficulty of a CW clue, which is expected to be visible in:

- Response Times (RTs): how long does it take to solve the clue, i.e. reading,guessing and typing the answer;
- Accuracy: How accurate is the answer.

Consequently, a trivial answer would have low Surprisal, which means a high probability, and vice versa we can consider high Surprisal, or low probability of the target word, as indicating a non-obvious, original answer. Several psycholinguistic studies investigate language processing in predicting next word, but no use of CW data have been found on this task. Finding the word-answer, given a definition, could be considered a type of next word prediction task. In this case not only the probability of the word must be considered, but more than that the Accuracy. Indeed, the right choice of the exact word needed to fill the grid characterizes a CW task. The current experimental proposal configures as an explorative approach for a psycholinguistic treatment of CW language, and as an attempt to investigate LLMs abilities to grasp different levels of surprise, linguistic originality in CW clues. The experimental setup consists of two different paths, the results of which will be compared.

- **Human Experiment:** the first step consists of a Solving Task to test participants and collect human responses. The absence of already annotated corpora for CW language leads to the limitation of having a constrained number of tested items, for reasons of time and because they are hand-designed.
- **LLMs Surprisal Calculation:** this limitation is not encountered on the LLMs side, with which

Macrocategory	Typologies	Examples
copular	cop:missSubj, copular sentence with subject omission	Fu Cancelliere della Germania dal 1949 al 1963 = *Adenauer*
copular	cop:clitic, copular sentence with a clitic in object position	Venere **ne** era la dea = *bellezza*
copular	cop:pron, copular sentence with a pronoun in object position	È celebre quella di Trinità dei Monti = *scalinata*
verbal predicate	act:missSubj, active verbal sentences with subject omission	Risiede in uno spazio geografico determinato = *abitante*
verbal predicate	act:clitic, active verbal sentences with a clitic in object position	La segue il medico = *ammalata*
verbal predicate	act:pron, active verbal sentences with a pronoun in object position	Quelli d'America hanno per capitale Washington = *Stati uniti*
verbal predicate	pass:missSubj, passive sentence with subject omission	È detta Il Continente Bianco = *Antartide*
verbal predicate	pass:other, other kinds of passive sentences	Vi furono ritrovati noti bronzi = *Riace*
verbal predicate	imp_refl:missSubj, active sentence with impersonal pronoun or reflexive verb with subject omission	Si reca spesso al catasto = *geometra*
verbal predicate	imp_refl:other, other kinds of active sentence with impersonal pronoun or reflexive verb	Che si riferisce all'Università = *accademico*
infinitive	inf_VP, infinitival verb phrases (VP)	Investire di un grado = *nominare*
nominal	bare_NP, bare noun phrases (NP)	Infuso paglierino = *tè*
nominal	bare_NP:rel, bare NP followed by a relative clause	Cilindri commestibili che vengono affettati = *polpettoni*
nominal	def_DP, definite determiner phrases (DP)	Il conto delle spese da farsi = *preventivo*
nominal	def_DP:rel, DP followed by a relative clause	Lo Stato di cui fanno parte le Isole Azzorre = *Portogallo*
nominal	ind_DP, indefinite DP	Una brutta abitudine perdonabile = *vizietto*
prepositional	PP, prepositional phrases	Davanti a Rodrigo = *Don*
adjectival	adjP, adjectival phrases	Probo, retto = *onesto*
adjectival	adjP:pron, adjectival phrases with pronoun	Pittoresco quello siciliano = *carretto*
metalinguistic	two-letters answer	Il centro di Matera = *TE*

Table 3
Typologies of linguistic clues with corresponding examples and macro-categories

the entire dataset can be used without particular time-issues. LLMs will assign word probabilities to the clue-answer pairs and Surprisal will be automatically measured starting from this output.

- **Experimental Results:** finally, the comparison between Surprisal values and human measures will tell us whether LLMs are able to correctly predict the difficulty of a clue-answer pair.

4.1. Solving Task

Starting from our reference dataset, a set of clue-answer pairs has been selected consisting of a limited number of 8 items for 20 categories presented in 3.2. A total of 160 items have been organized into four lists, all equally representative of the categories. Hence, a subject was presented with one of these four lists and asked to solve 40 CW clues. 60 Italian native speakers were recruited for the experiment. Participants were presented with a clue, and they had to guess the solution, having at their disposal only the length of the answer, represented as a grid, and its initial letter. No time constraint was given during the experiment. For each subject and each item (2880 data points) in the experimental list we collected:

- The string representing the given answer.
- RT (response time) was measured as the interval in milliseconds between the appearance of the crossword clue and the submission of the answer. This includes reading, comprehension, and typing time.

Results will be presented in the following sections.

4.2. LLMs Surprisal Calculation

To assess how predictable crossword answers are for a language model, we use the notion of *surprisal*, defined as the negative logarithm of a token's predicted probability. In the case of full-word answers, we compute:

$$\text{AnswerSurprisal} = -\log\big(P(\text{answer})\big) \tag{1}$$

where $P(\text{answer})$ denotes the probability assigned by the model to the answer. Because we work with *causal language models*—which predict the next token based only on the left-hand context—this surprisal is computed as *last word surprisal* by placing the answer at the *end* of a concatenated input, typically of the form `clue + answer`. This ensures that the model encounters the clue as context before attempting to generate or evaluate the answer, in line with the left-to-right autoregressive mechanism of causal models.

Crossword answers may consist of multiple tokens, as in: *I bambini possono riceverla dopo i sette anni = prima comunione* ('kids can receive it after the seventh year = first communion'). In these cases, the surprisal must refer to the entire answer sequence. Letting the answer consist of tokens $t_1, t_2, \ldots, t_n$, the surprisal becomes:

$$\text{AnswerSurprisal} \ = \ -\sum_{i=1}^{n} \log\bigl(P(t_i)\bigr) \tag{2}$$

This captures the cumulative surprisal of all the answer tokens, assuming the clue and previous answer tokens have already been processed.

In some cases, however, the format of the input may place the answer at the *beginning* of the sequence, rather than at the end, recalling a topicalized structure [45, 46, 47, 48, 49]. The interesting thing is that, given how the clues are phrased (as definitions or comments), the most general structure would actually be that of topic + comment in which the comment or clue provides relevant information about the answer that represents accordingly the topic of the clue. This structure then constitutes the most suitable strategy of concatenation in line with the CW puzzle logic. For such *reverse concatenations* (e.g., `answer + clue`), however, standard Answer Surprisal is no longer applicable because causal models, in virtue of their incremental progressive nature, cannot condition on future tokens. To address this, we introduce a complementary measure: **Surprisal Difference**. This measure is used in all the concatenation rules that do not permit to use the standard Answer Surprisal like the Topic-based rule. So concatenation rules that have the answer at the end use *AnswerSurprisal* while concatenation rules that have the answer in the beginning use *SuprisalDifference* as their surprisal score.

Surprisal Difference compares the surprisal of the clue in isolation with the surprisal of the same clue following the answer. It captures how much the presence of the answer facilitates (or reduces the unexpectedness of) the clue:

$$\text{SurprisalDiff} = S(a + c) - S(c) \tag{3}$$

where $S(\cdot)$ denotes surprisal, c is the clue, and a is the answer.

This difference provides an interpretable surprisal-based signal even when the answer appears before the clue, a configuration that, as said, arises in certain experimental concatenation schemes. The assumption is that if the answer helps predict the clue, the clue's surprisal should be lower when preceded by the answer.

Both Answer Surprisal and Surprisal Difference rely on the autoregressive, left-to-right prediction behavior of causal models. For each concatenation strategy, the suitable Surprisal measure is calculated. To ensure linguistically accurate tokenization and probability estimates, we use models that are pre-trained or fine-tuned on Italian data.

4.2.1. Experimental items preparation for models Surprisal

Complete sentences composed of clue and answer are given in input to the models, thus it must be faced the issue of concatenating clue and answer in grammatical and coherent structures without substantially modifying the clue style, syntactic characterization and meaning and having the answer as final word so as to calculate its Surprisal value after the context represented by the clue.

In most cases, the answer maintains a synonymy relationship with the clue, which can often be expressed using the Italian adverb *cioè*. This allows for an automatic concatenation of clue-answer pairs, forming sentences where the answer appears as the final word, such as **<clue> cioè <answer>**.

To analyze how different concatenation strategies impact Surprisal values, various concatenation rules have been applied to the dataset, ensuring that each clue-answer pair is formatted appropriately for model evaluation. The employed concatenation methods are:

Different concatenations has been then employed:

Cioè rule `<clue> cioè ART <answer>`

Subject-based rule `ART <answer> <clue>`

Topic-based rule `ART <answer> , <clue>`

Copular rule `ART <answer>` *VERB(TO BE)* `<clue>`

Inverse-copular rule `<clue>` *VERB(TO BE)* `ART <answer>`

Prompt rule `Sei un cruciverbista esperto. Ti verrà fornita una definizione a cui dovrai rispondere correttamente. La definizione è: <clue>. La risposta ha <answer length> lettere, inizia con <answer's first letter>, <answer>`

These different formulations allow for a comparative analysis of Surprisal variations across clue structures, ensuring that the most effective concatenation strategy can be identified for each category.

For each item in the dataset, the model will calculate the probability of each token, then the token composing the answer are used to estimate the Surprisal of the answer given the other tokens. High Surprisal values at the answer final word will tell us that the answer is unexpected in that context, and consequently harder to guess. Different types of Surprisal are so defined by means of how data are labelled, by means of the different concatenation rules. This opens the door to fine-grained investigation in different directions. One rule could work better with some categories than the others in enabling the model to do more reliable predictions. The possibility exists of elaborating specific rules for each structure of clue-answer pair, in order to make input items as realistic as possible and hence improve the model performance in predicting human responses. To evaluate models' performances in predicting Accuracy and RTs, Surprisal values will be compared with results collected in the human experiment. The comparison should highlight:

- A positive correlation between Surprisal and RTs;
- A negative correlation between Surprisal and Accuracy.

Different Surprisal have been calculated with different models and with different concatenations rules. Pearson coefficient will tell us more on the correlation between these variables, human data and Surprisal (for the three models employed). For both Accuracy and RTs we will have:

- A global comparison, which tells us whether each model's Surprisal output is in a significant correlation with human measures;
- The correlations between Surprisal and Accuracy or RTs for each category, to observe whether more relevant correlations are there for some of the categories.

5. Experimental Results

The experimental results focus on the correlation between Surprisal values and human performance in solving CW clue-answer pairs. We tested this approach on three models: Llama-3-8B [1] , Llama-2-7B [2], and Ita-GPT-2 Medium-121M [3]. The mean Accuracy of participants in the human experiment was found to be 0.63.

5.1. Correlation Analysis

To examine the relationship between Surprisal values and human Accuracy, we first conducted a Pearson correlation analysis using mean per-item accuracy scores. The results revealed a negative correlation, consistent with our hypothesis that higher Surprisal values correspond to more difficult clues. Among the tested models, Llama3 and Ita-GPT2 yielded higher Pearson coefficients, which may reflect Llama3's extensive multilingual capacity and Ita-GPT2's fine-tuning on Italian. Figure 2 illustrates the correlation between Surprisal and Accuracy for the three models on a representative concatenation rule. In addition, Tables 11, 12, and 13 in the Appendix report a Generalized Linear Mixed Model (GLMM) analysis, which incorporates individual variability without aggregating accuracy values. This analysis further confirms Surprisal as a significant predictor of Accuracy, and therefore of clue difficulty.

We also investigated the relationship between surprisal and response times (RTs) using a series of Linear Mixed Models (LMMs) fitted separately for each concatenation type. RTs were log-transformed to correct for positive skew and stabilize variance, in line with standard psycholinguistic practice. This transformation helped reduce the impact of outliers and enabled the use of parametric modeling techniques. In each model, surprisal was included as a fixed effect, and subject-specific intercepts were modeled as random effects to account for baseline variation across participants.

The results consistently showed a statistically significant positive relationship between surprisal and log-transformed RTs across all concatenation types as summarized in table 4 for Llama3 and the other two models in the appendix (table 14, 15). This indicates that clues with higher surprisal values led to longer response times, supporting the hypothesis that surprisal reflects processing difficulty. Although the magnitude of the effect varied by concatenation rule, all coefficients were positive, and confidence intervals did not include zero.

These findings demonstrate that surprisal is a robust predictor of reading latency in the crossword task, even under minimal context and with sparse surface cues. Importantly, this effect emerges despite the lack of explicit time pressure, suggesting that surprisal exerts an automatic influence on processing effort.

While the overall pattern is clear, future research could further refine the temporal precision of RTs by decomposing the overall response into distinct phases. Specifically, logging (i) the time to initiate typing, (ii) the typing duration, and (iii) the post-completion delay would help distinguish comprehension time from motor and decision-related delays. This would allow a more direct mapping between linguistic difficulty and behavioral latency, providing an even clearer picture of the cognitive

[1] meta-llama/Meta-Llama-3-8B
[2] meta-llama/Llama-2-7b-hf
[3] GroNLP/gpt2-small-italian

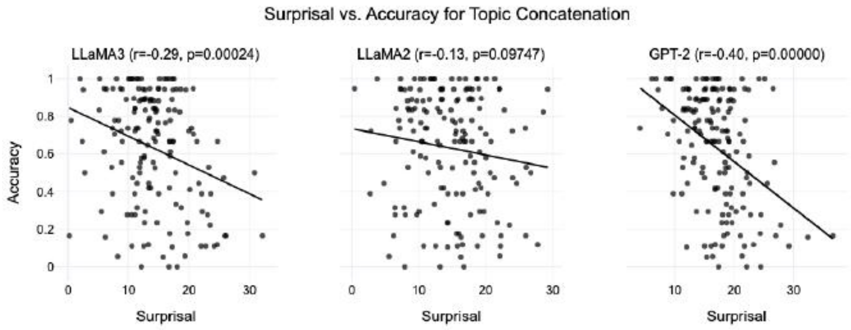

Figure 2: Correlation between Surprisal and Accuracy for the three models with Topic Concatenation.

Concatenation Type	Coef	Std.Err	z	p-value	CI Lower	CI Upper
concatenation_topic_art	0.023	0.003	6.983	0.0000	0.017	0.030
concatenation_subj_art	0.029	0.003	9.726	0.0000	0.023	0.035
concatenation_cioè_art	0.034	0.005	6.600	0.0000	0.024	0.044
concatenation_cop	0.018	0.003	6.671	0.0000	0.013	0.024
concatenation_inv_cop	0.044	0.005	7.996	0.0000	0.033	0.055
concatenation_prompt	0.065	0.005	12.665	0.0000	0.055	0.075
solution	0.026	0.002	14.370	0.0000	0.023	0.030

Table 4
Linear Mixed Model results: effect of surprisal on log-transformed RT by concatenation type for Llama3

mechanisms involved.

Macro Category	Concat. type	r	p
infinitive	topic_art	-0.59	0.123
verb_pred	subj_art	**-0.32**	**0.0177**
metalinguistic	cop	-0.45	0.259
nominal	topic_art	**-0.62**	**2.41e-05**
copular	prompt	-0.12	0.578
prepositional	topic_art	-0.59	0.126
adjectival	cioè_art	-0.44	0.0884

Table 5
Best correlation coefficients (r) and p-values for each macro category and concatenation type (Ita-GPT-2 Medium-121M).

5.1.1. Correlation in Different Categories

To further investigate how Surprisal correlates with human performance across different types of clues, we analyzed the correlation separately for different macrocategories and individual categories. The results are visualized in Figures 3 for the Ita-GPT-2 model. Our findings indicate that the strength of the correlation between Surprisal and Accuracy varies significantly depending on the type of clue. In particular, two categories showed notably weak correlations:

- **Metalinguistic Clues:** This category exhibited no correlation between Surprisal and Accuracy. A likely explanation is the difficulty transformers face when processing metalinguistic cues, such as wordplays and abbreviations. Since these models rely on token probabilities, and not on single characters they struggle to accurately predict non-standard or unconventional relationships between clues and answers, which are common in metalinguistic clues.
- **Copular Clues:** The correlation was also absent for copular structures. One probable reason is that the *cioè* concatenation rule does not naturally fit the syntactic structure of these clues. Copular constructions often require a more flexible paraphrasing strategy, rather than a simple equivalence statement, leading to suboptimal Surprisal estimations.

Other categories, particularly nominal and verbal predicate structures, displayed stronger correlations, suggesting that Surprisal works better for categories where the clue-answer relationship is more straightforwardly semantic rather than dependent on linguistic nuances like wordplay or syntactic constraints.

A more robust analysis with GLMMs, to account for individual variability, will require more data for each cat-

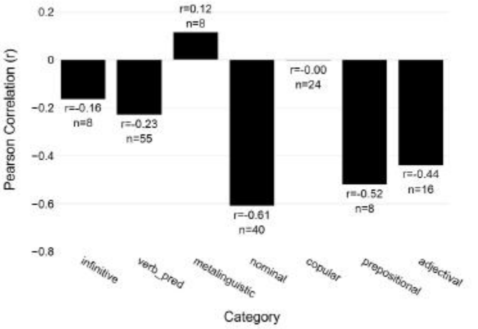

Figure 3: Correlation between Surprisal and Accuracy across different macrocategories for concatenation rule *cioè_art* and model GPT-2.

egory. We leave this further effort to future experimental work.

5.2. Effect of Concatenation Strategies

We also explored the impact of different concatenation strategies on model performance. The concatenation method influenced Surprisal values differently across clue categories. Some structures benefited from the *cioè* rule, while others yielded more reliable Surprisal estimates under different approach.

Table 5 shows, for each macro category, the concatenation that yields the best correlation results and it's value. These results highlight the importance of category-specific approaches when applying Surprisal-based difficulty estimation.

5.3. Summary of Findings

Overall, our findings confirm that Surprisal serves as a useful predictor of CW puzzle difficulty, particularly when considering Accuracy as a measure of challenge. However, its predictive power for solving times remains limited, likely due to the nature of short CW clues. The choice of concatenation strategy also plays a crucial role in model performance, suggesting that tailored approaches could further refine Surprisal-based difficulty estimations.

6. Conclusion

This paper provides the first cognitively grounded, automatic gauge of crossword–clue difficulty. We compiled a 160-item Italian benchmark (2 880 human judgements), converted each clue–answer pair into well-formed sentences with five templates, and estimated token-level Surprisal with three causal LLMs (Ita-GPT-2-121M, Llama-2-7B, Llama-3-8B).

Answers to the research questions

1. **RQ1:** Higher Surprisal predicts lower solver accuracy (best $r = -0.57$) and longer log-RTs, showing that information-theoretic "surprise" mirrors cognitive load.
2. **RQ2:** Language match beats raw size: the Italian-specific Ita-GPT-2 and multilingual Llama-3 surpass the larger, English-leaning Llama-2.
3. **RQ3:** No single template suffices. Topic–comment placement works best for nominal and verbal clues, the *cioè* rule for many adjectival/infinitival ones, while copular and metalinguistic items need ad-hoc rewrites; selecting the best rule per macro-category adds up to 0.15 r-points.
4. **RQ4:** Category-specific Surprisal thresholds separate "easy", "medium" and "hard" clues, enabling an adaptive generator that targets any solver level.

Main finding. LLM-derived Surprisal is a reliable, fine-grained predictor of human crossword difficulty, explaining more than half of the variance in accuracy for the most common clue types.

Limitations (i) Italian-only data; other languages may need new tokenisers. (ii) The 160-item set limits power for rare structures. (iii) RTs blend reading, reasoning and typing; keystroke logs would isolate comprehension latency. (iv) Only decoder-style LLMs were tested; encoder–decoder or retrieval-augmented models might align differently. (v) Clues were scored in isolation, ignoring cross-checks within full grids.

Future work

1. Scale the benchmark to thousands of clues, multiple languages and complete grids.
2. Log richer behaviour (eye-tracking, keystrokes, EEG) to separate processing stages.
3. Probe new architectures and character-level tokenisers for closer cognitive fidelity.
4. Fuse Surprisal with real-time solver profiles for personalised tutoring.
5. Couple Surprisal-based clue ranking with constraint-based fills to deliver fully adaptive crosswords.

Anchoring puzzle evaluation in probabilistic language theory links NLP, psycholinguistics and game AI, promising crosswords that scale from novice amusement to expert challenge while offering a fresh lens on human–machine language alignment.

References

[1] E. Wallace, N. Tomlin, A. Xu, et al., Automated crossword solving, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2022, pp. 2968–2981.

[2] S. Kulshreshtha, O. Kovaleva, N. Shivagunde, A. Rumshisky, Down and across: Introducing crossword-solving as a new nlp benchmark, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2022, pp. 2648–2659.

[3] M. L. Ginsberg, Dr. fill: Crosswords and an implemented solver for singly weighted csps, Journal of Artificial Intelligence Research 42 (2011) 851–886.

[4] R. Leban, How do crosshare difficulty ratings work?, https://crosshare.org/articles/crossword-difficulty-ratings, 2021. Accessed 4 June 2025.

[5] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423.

[6] J. Hale, A probabilistic earley parser as a psycholinguistic model, in: Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2001, pp. 159–166.

[7] R. Levy, Expectation-based syntactic comprehension, Cognition 106 (2008) 1126–1177.

[8] V. Demberg, F. Keller, Data from eye-tracking corpora as evidence for theories of incremental parsing, Cognition 109 (2008) 193–210.

[9] N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic, Cognition 128 (2013) 302–319.

[10] H. Touvron, T. Lavril, G. Izacard, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[11] M. A. Research, The llama 3 herd of models, Meta Research Blog (2024). Accessed 4 June 2025.

[12] P. Arehalli, R. Futrell, Syntactic surprisal from neural models predicts, but underestimates, human garden-path difficulty, in: Proceedings of the 26th Conference on Computational Natural Language Learning, Association for Computational Linguistics, 2022, pp. 269–283.

[13] B.-D. Oh, W. Schuler, Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?, Transactions of the Association for Computational Linguistics 11 (2023) 336–350.

[14] T. Liu, I. Škrjanec, V. Demberg, Temperature-scaling surprisal estimates improve fit to human reading times–but does it do so for the" right reasons"?, arXiv preprint arXiv:2311.09325 (2023).

[15] S. Nair, P. Resnik, Words, subwords, and morphemes: what really matters in the surprisal-reading time relationship?, arXiv preprint arXiv:2310.17774 (2023).

[16] E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, R. Levy, On the predictive power of neural language models for human real-time comprehension behavior, arXiv preprint arXiv:2006.01912 (2020).

[17] B. Krieger, H. Brouwer, C. Aurnhammer, M. W. Crocker, On the limits of llm surprisal as functional explanation of erps, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 46, 2024.

[18] E. Huber, S. Sauppe, A. Isasi-Isasmendi, I. Bornkessel-Schlesewsky, P. Merlo, B. Bickel, Surprisal from language models can predict erps in processing predicate-argument structures only if enriched by an agent preference principle, Neurobiology of language 5 (2024) 167–200.

[19] D. Jurafsky, Probabilistic modeling in psycholinguistics: Linguistic comprehension and production, Probabilistic linguistics 21 (2003) 1–30.

[20] M. Greco, A. Cometa, F. Artoni, R. Frank, A. Moro, False perspectives on human language: Why statistics needs linguistics, Frontiers in Language Sciences 2 (2023) 1178932.

[21] M. Wilson, J. Petty, R. Frank, How abstract is linguistic generalization in large language models? experiments with argument structure, Transactions of the Association for Computational Linguistics 11 (2023) 1377–1395.

[22] J. Hale, M. Stanojević, Do llms learn a true syntactic universal?, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 17106–17119.

[23] S. Slaats, A. E. Martin, What's surprising about surprisal, Computational Brain & Behavior (2025) 1–16.

[24] M. Schrimpf, I. Blank, N. Kanwisher, E. Fedorenko, The neural architecture of language is grounded in predictive deep networks, Science 374 (2021) 105–111.

[25] C. Caucheteux, J.-R. King, Brains and algorithms partially converge in natural language processing, Communications Biology 5 (2022) 1–10.

[26] C. Shain, E. Wilcox, R. Levy, Large language models still diverge from humans in predictive processing: a mega-study, Psychological Science (2024).

[27] A. Goodkind, K. Bicknell, Predictive power of word frequency and surprisal for reading times, in: Proceedings of CogSci, 2018.

[28] M. Littman, K. Ho, S. Shell, J. O'Neill, PROVERB: A probabilistic crossword solver, in: AAAI, 1999.

[29] M. Ernandes, G. Angelini, M. Gori, WebCrow: A

web-based system for crossword solving, in: AAAI, 2005.

[30] D. R. Radev, R. Zhang, S. Wilson, Cruciform: Solving crosswords with nlp, in: Workshop on Structured Prediction for NLP, 2016.

[31] S. Saha, S. Chakraborty, S. Saha, U. Garain, Language models are crossword solvers, arXiv preprint arXiv:2406.09043 (2024).

[32] L. Rigutini, M. Maggini, M. Gori, Automatic generation of crossword puzzles, in: IEA/AIE, 2008.

[33] L. Rigutini, M. Maggini, M. Gori, Automatic crossword puzzle generation and its educational applications, in: AI*IA, 2012.

[34] H. Ranaivo-Malançon, M. R. Sazali, Automatic fill-in crosswords in malay and english, Journal of Computer Science (2013).

[35] A. Esteche, R. Rosito, Automatic generation of spanish crossword puzzles from news, in: Proceedings of Clei, 2017.

[36] A. Arora, A. Kumar, SEEKH: Generating educational crosswords for indian languages, in: International Conference on Educational Data, 2019.

[37] K. Zeinalipour, T. Iaquinta, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, Italian crossword generator: Enhancing education through interactive word puzzles, in: Proceedings of CLiC-it, 2023.

[38] G. Angelini, M. Ernandes, T. Iaquinta, C. Stehlé, F. Simões, K. Zeinalipour, A. Zugarini, M. Gori, The webcrow french crossword solver, arXiv preprint arXiv:2311.15626 (2023).

[39] K. Zeinalipour, M. Z. Saad, M. Maggini, M. Gori, ArabIcros: Ai-powered arabic crossword puzzle generation for educational applications, arXiv preprint arXiv:2312.01339 (2023).

[40] K. Zeinalipour, M. Z. Saad, M. Maggini, M. Gori, From arabic text to puzzles: Llm-driven development of arabic educational crosswords, in: Proceedings of the Workshop on Language Models for Low-Resource Languages, 2025.

[41] K. Zeinalipour, Y. G. Keptiğ, M. Maggini, L. Rigutini, M. Gori, A turkish educational crossword puzzle generator, arXiv preprint arXiv:2405.07035 (2024).

[42] A. Zugarini, K. Zeinalipour, S. S. Kadali, M. Maggini, M. Gori, L. Rigutini, Clue-instruct: Text-based clue generation for educational crossword puzzles, in: Proceedings of LREC-COLING, 2024.

[43] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, et al., spacy: Industrial-strength natural language processing in python (2020).

[44] K. Zeinalipour, T. Iaquinta, A. Zanollo, G. Angelini, L. Rigutini, M. Maggini, M. Gori, et al., Italian crossword generator: an in-depth linguistic analysis in educational word puzzles, IJCOL 11 (2025) 47–72.

[45] L. Rizzi, On the form of chains: Criterial positions and ecp effects (2006).

[46] T. Reinhart, Pragmatics and linguistics: an analysis of sentence topics (1981).

[47] S. Cruschina, The syntactic role of discourse-related features (2009).

[48] S. Cruschina, Topicalization in Romance Languages, 2021.

[49] S. Cruschina, Topicalization, dislocation and clitic resumption, 2022.

7. Appendices

In the following section we report the complete results for all llms and concatenation rules divided by macro category and languege model, The appendix already contains one correlation table for each model; see their individual captions.

Table 6

Concatenation type with highest correlation coefficients (r) and p-values for each macro-category (Llama2).

Macro Category	Concatenation Type	r	p
infinitive	concatenation_prompt	**-0.72**	**0.0428**
verb_pred	concatenation_cioè_art	**-0.34**	**0.0119**
metalinguistic	concatenation_topic_art	-0.28	0.506
nominal	concatenation_topic_art	**-0.37**	**0.0175**
copular	concatenation_cioè_art	-0.25	0.243
prepositional	concatenation_inv_cop	**-0.80**	**0.0168**
adjectival	concatenation_prompt	-0.41	0.111

Table 7

Best correlation coefficients (r) and p-values for each macro-category and concatenation type (Llama3).

Macro Category	Concatenation Type	r	p
infinitive	concatenation_subj_art	-0.51	0.192
verb_pred	concatenation_prompt	**-0.37**	**0.00614**
metalinguistic	concatenation_cioè_art	-0.42	0.305
nominal	concatenation_topic_art	**-0.45**	**0.00385**
copular	concatenation_cioè_art	-0.26	0.215
prepositional	concatenation_topic_art	-0.56	0.150
adjectival	concatenation_prompt	**-0.60**	**0.0142**

Table 8

Best correlation coefficients (r) and p-values for each category using Llama3.

Category	Concatenation Type	r	p
inf_VP	concatenation_subj_art	-0.51	0.192
pass:other	concatenation_cop	-0.29	0.482
metalinguistic	concatenation_cioè_art	-0.42	0.305
imp_refl:missSubj	concatenation_topic_art	**-0.95**	**0.00023**
def_DP	concatenation_topic_art	-0.63	0.093
cop:missSubj	concatenation_prompt	**-0.74**	**0.0373**
PP	concatenation_topic_art	-0.56	0.150
cop:pron	concatenation_cop	-0.46	0.257
ind_DP	concatenation_inv_cop	-0.54	0.168
cop:clitic	concatenation_subj_art	-0.43	0.290
bare_NP:rel	concatenation_cioè_art	-0.65	0.083
adjP:pron	concatenation_prompt	-0.53	0.180
bare_NP	concatenation_cop	-0.39	0.342
adjP	concatenation_cioè_art	**-0.72**	**0.0432**
act:pron	concatenation_inv_cop	-0.58	0.128
act:missSubj	concatenation_cop	-0.43	0.284
def_DP:rel	concatenation_inv_cop	-0.54	0.165
imp_refl:other	concatenation_cioè_art	**-0.79**	**0.0334**
act:clitic	concatenation_subj_art	-0.60	0.114
pass:missSubj	concatenation_prompt	-0.70	0.0543

Table 9
Best correlation coefficients (r) and p-values for each category using Llama2.

Category	Concatenation Type	r	p
inf_VP	concatenation_prompt	**-0.72**	**0.0428**
pass:other	concatenation_prompt	-0.38	0.348
metalinguistic	concatenation_topic_art	-0.28	0.506
imp_refl:missSubj	concatenation_cioè_art	**-0.90**	**0.00256**
def_DP	concatenation_topic_art	-0.48	0.234
cop:missSubj	concatenation_inv_cop	-0.31	0.450
PP	concatenation_inv_cop	**-0.80**	**0.0168**
cop:pron	concatenation_prompt	-0.49	0.217
ind_DP	concatenation_cioè_art	-0.45	0.262
cop:clitic	concatenation_prompt	-0.35	0.396
bare_NP:rel	concatenation_cop	**-0.78**	**0.0217**
adjP:pron	concatenation_cioè_art	-0.25	0.552
bare_NP	concatenation_topic_art	-0.56	0.145
adjP	concatenation_prompt	**-0.86**	**0.00606**
act:pron	concatenation_topic_art	-0.51	0.194
act:missSubj	concatenation_cop	-0.33	0.424
def_DP:rel	concatenation_inv_cop	-0.61	0.111
imp_refl:other	concatenation_cioè_art	**-0.78**	**0.0367**
act:clitic	concatenation_topic_art	**-0.84**	**0.00979**
pass:missSubj	concatenation_cioè_art	-0.58	0.134

Table 10
Best correlation coefficients (r) and p-values for for each category using GPT-2.

Category	Concatenation Type	r	p
inf_VP	concatenation_topic_art	-0.59	0.123
pass:other	concatenation_prompt	-0.22	0.608
metalinguistic	concatenation_cop	-0.45	0.259
imp_refl:missSubj	concatenation_topic_art	**-0.91**	**0.00163**
def_DP	concatenation_cioè_art	**-0.92**	**0.00111**
cop:missSubj	concatenation_prompt	-0.4	0.326
PP	concatenation_topic_art	-0.59	0.126
cop:pron	concatenation_cop	-0.22	0.605
ind_DP	concatenation_cioè_art	-0.48	0.226
cop:clitic	concatenation_prompt	-0.08	0.853
bare_NP:rel	concatenation_cioè_art	-0.66	0.0725
adjP:pron	concatenation_cioè_art	-0.35	0.391
bare_NP	concatenation_topic_art	**-0.71**	**0.0493**
adjP	concatenation_topic_art	-0.69	0.0591
act:pron	concatenation_subj_art	-0.55	0.158
act:missSubj	concatenation_cioè_art	-0.13	0.765
def_DP:rel	concatenation_cioè_art	-0.44	0.271
imp_refl:other	concatenation_prompt	**-0.78**	**0.0385**
act:clitic	concatenation_cop	**-0.93**	**0.00075**
pass:missSubj	concatenation_topic_art	**-0.76**	**0.0285**

Table 11
Logistic Mixed Model results: effect of surprisal on accuracy by concatenation type for Llama3

Concatenation Type	Coef	Std.Err	z	p-value	CI Lower	CI Upper
concatenation_topic_art	-0.064	0.008	-7.700	0.0000	-0.080	-0.048
concatenation_subj_art	-0.063	0.007	-8.414	0.0000	-0.078	-0.048
concatenation_cioè_art	-0.108	0.013	-8.431	0.0000	-0.133	-0.083
concatenation_cop	-0.033	0.007	-4.865	0.0000	-0.046	-0.020
concatenation_inv_cop	-0.111	0.014	-8.177	0.0000	-0.137	-0.084
concatenation_prompt	**-0.157**	0.014	-11.315	0.0000	-0.184	-0.130

Table 12
Logistic Mixed Model results: effect of surprisal on accuracy by concatenation type for Llama2

Concatenation Type	Coef	Std.Err	z	p-value	CI Lower	CI Upper
concatenation_topic_art	-0.032	0.008	-4.032	0.0001	-0.048	-0.017
concatenation_subj_art	-0.034	0.008	-4.428	0.0000	-0.049	-0.019
concatenation_cioè_art	**-0.114**	0.012	-9.178	0.0000	-0.138	-0.089
concatenation_cop	-0.007	0.007	-0.924	0.3560	-0.021	0.007
concatenation_inv_cop	-0.059	0.010	-5.870	0.0000	-0.079	-0.039
concatenation_prompt	-0.016	0.004	-3.675	0.0002	-0.024	-0.007

Table 13
Logistic Mixed Model results: effect of surprisal on accuracy by concatenation type for GPT-2

Concatenation Type	Coef	Std.Err	z	p-value	CI Lower	CI Upper
concatenation_topic_art	-0.113	0.010	-11.114	0.0000	-0.133	-0.093
concatenation_subj_art	-0.029	0.005	-5.726	0.0000	-0.039	-0.019
concatenation_cioè_art	**-0.116**	0.011	-10.886	0.0000	-0.137	-0.095
concatenation_cop	-0.012	0.005	-2.413	0.0158	-0.022	-0.002
concatenation_prompt	-0.107	0.011	-9.406	0.0000	-0.130	-0.085
concatenation_inv_cop	-0.008	0.011	-0.701	0.4830	-0.029	0.014

Table 14
Linear Mixed Model results: effect of surprisal on log-transformed RT by concatenation type for Llama2

Concatenation Type	Coef	Std.Err	z	p-value	CI Lower	CI Upper
concatenation_topic_art	0.012	0.003	3.489	0.0005	0.005	0.018
concatenation_subj_art	0.019	0.003	5.935	0.0000	0.013	0.025
concatenation_cioè_art	0.046	0.005	9.280	0.0000	0.036	0.056
concatenation_cop	0.011	0.003	3.643	0.0003	0.005	0.017
concatenation_inv_cop	0.022	0.004	5.240	0.0000	0.014	0.030
concatenation_prompt	0.013	0.002	7.225	0.0000	0.009	0.016

Table 15
Linear Mixed Model results: effect of surprisal on log-transformed RT by concatenation type for GPT-2

Concatenation Type	Coef	Std.Err	z	p-value	CI Lower	CI Upper
concatenation_topic_art	0.034	0.004	8.890	0.0000	0.027	0.041
concatenation_subj_art	0.015	0.002	7.126	0.0000	0.011	0.019
concatenation_cioè_art	0.043	0.004	10.779	0.0000	0.035	0.051
concatenation_cop	0.010	0.002	4.676	0.0000	0.006	0.014
concatenation_prompt	0.058	0.004	13.215	0.0000	0.049	0.066
concatenation_inv_cop	-0.009	0.005	-1.944	0.0519	-0.018	0.000