# Italian-based Large Language Models at the Syntax-Semantics Interface: the Case of Instrumental Role

Alice Suozzi*
Università di Venezia, Ca' Foscari

Simone Mazzoli**
Università di Venezia, Ca' Foscari

Gianluca E. Lebani†
Università di Venezia, Ca' Foscari

*The linguistic competence of Large Language Models (LLMs) has been the focus of extensive investigation in recent years. Yet, the syntax-semantics interface remains a relatively understudied aspect of LLMs' linguistic abilities. This study aims to address this gap by focusing on the Instrumental role in Italian. In this language, Instruments can always be syntactically omitted, yet they remain semantically present, as they are recoverable either from the verb meaning alone (when the verb is presented in isolation) or from the interaction between the verb meaning and that of its internal argument (when the verb appears within a syntactic context).*

*To assess the ability of LLMs to semantically determine the most appropriate Instrument(s) from the verb meaning, we conducted two experiments based on psycholinguistically inspired tasks, comparing the performance of GePpeTto and Minerva models (350M, 1B, 3B and 7B) to that of Italian speakers. In the first experiment, verbs were presented in isolation, while in the second, they were presented within a syntactic context. Our findings indicate that the performance of LLMs is influenced by the semantic selectivity of verbs, the presence or absence of a clausal context and model characteristics.*

## 1. Introduction

In the last decade, Large Language Models (LLMs) have become the benchmark technology in Natural Language Processing. The impressive results they achieve in understanding and generating text have raised increasingly pertinent questions about the nature of the linguistic knowledge encoded within LLMs and what this can reveal about human language. Specifically, what type of knowledge do these models encode? Can the linguistic representations they generate be leveraged for psycholinguistic studies? (Linzen and Baroni 2022; Blank 2023; Wilcox et al. 2023).

* QuaCLing Lab, Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Dorsoduro 1075, 30123 Venice, Italy
E-mail: alice.suozzi@unive.it
** QuaCLing Lab, Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Dorsoduro 1075, 30123 Venice, Italy
E-mail: simone.mazzoli@unive.it
† QuaCLing Lab, Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Dorsoduro 1075, 30123 Venice, Italy
European Centre for Living Technology (ECLT), Ca' Bottacin, Dorsoduro 3911, 30123 Venice, Italy
E-mail: gianluca.lebani@unive.it

© 2025 Associazione Italiana di Linguistica Computazionale

Moreover, their human-like performance on a wide range of linguistic tasks has reignited the long-standing debate regarding the innate mechanisms required for language acquisition (Linzen and Baroni 2022). Some argue that LLMs' linguistic capabilities undermine the need for innate linguistic equipment (Piantadosi 2023), while others contend that their limitations on certain tasks provide evidence for the existence of such innate mechanisms (Kodner, Payne, and Heinz 2023).

Linguistic phenomena situated at the syntax-semantics interface form a crucial aspect of human linguistic competence and provide an ideal testing ground for assessing the linguistic capabilities of LLMs. This is particularly true because, unlike purely syntactic phenomena, their acquisition does not rely solely on exposure to linguistic data. Therefore, these phenomena serve as an effective means of determining whether LLMs can make linguistic generalizations and genuinely develop linguistic knowledge. However, relatively few studies have evaluated LLMs on these phenomena, and even fewer have done so using psycholinguistic methodologies or stimuli, or have compared LLMs' performance with that of human speakers.

This study aims to fill this gap by evaluating the ability of GePpeTto (De Mattei et al. 2020), and four models belonging to the Minerva family (-350M, -1B, -3B and -7B) (Orlando et al. 2024) to semantically recover Instruments from verb meaning in Italian, both when verbs appear in isolation and within a syntactic context. The models are tested through two tasks explicitly designed to assess the same capability in Italian speakers, enabling a comparison of the models' performance to that of human participants. We chose to focus on the Instrumental role because, in Italian, it is syntactically optional while still being present at the semantic level, making it particularly suitable for this investigation. For instance, one might say *Ho tagliato il pane* ∅ 'I cut the bread ∅' without explicitly mentioning the tool used (e.g., a knife), yet the knife still exists at the conceptual level. The ability to infer or recover the appropriate Instrument (in this case, the knife) from the verb meaning is not solely dependent on linguistic exposure, as Instruments are frequently omitted in discourse.

These pages are structured as follows: in Section 2, we provide an overview of the main studies addressing the acquisition of phenomena at the syntax-semantics interface by LLMs. In Section 3, we define the Instrument role and present our hypothesis on its syntactic optionality and semantic recoverability. Section 4 is dedicated to the presentation of our experiments. Experiment 1 (Section 4.1) focuses on verbs in isolation, while Experiment 2 (Section 4.2) examines verbs in a syntactic context. Finally, Section 5 presents the conclusions.

## 2. Large Language Models and the syntax-semantics interface

Linguistic phenomena situated at the syntax-semantics interface — such as verb argument structure, argument structure alternations, and implicit arguments — have been extensively studied in theoretical linguistics (Levin 1993; Resnik 1993; Croft 1998; Jezek 2003), psycholinguistic (McRae, Spivey-Knowlton, and Tanenhaus 1998; Ferretti, McRae, and Hatherell 2001; Kamide, Altmann, and Haywood 2003; McRae and Matsuki 2009; Cappelli 2022), and acquisitional linguistics (Medina 2007; Cappelli 2022).

Crucially, the mastering of these phenomena relies on the integration of multiple aspects of linguistic knowledge (i.e., lexical, semantic, and syntactic), which are not necessarily enhanced by mere exposure to large amounts of data. Investigating whether and how LLMs acquire such phenomena can offer valuable insights into the nature of their linguistic competence, which, as discussed in Section 1, remains a topic of ongoing debate.

Nonetheless, relatively few studies have been specifically designed to test LLMs' knowledge of syntax-semantics interface phenomena (Rogers, Kovaleva, and Rumshisky 2020). Among these exceptions, some research focuses on argument structure alternations, such as the Causative-Inchoative Alternation (Warstadt, Singh, and Bowman 2019; Kann et al. 2019; Veres and Sandblåst 2019; Warstadt et al. 2020; Seyffarth and Kallmeyer 2020; Thrush, Wilcox, and Levy 2020), the *Spray-Load* Alternation (Kann et al. 2019; Thrush, Wilcox, and Levy 2020; Veres and Sampson 2023; Samo et al. 2023; Wilson, Petty, and Frank 2023), and the Instrument-Subject Alternation (Seyffarth and Kallmeyer 2020). Others studies investigate thematic fit and verb selectional preferences (Vassallo et al. 2018; Ettinger 2020; Metheniti, Van de Cruys, and Hathout 2020; Kauf et al. 2022), or the assignment of thematic roles in optionally transitive English verbs (Tjuatja et al. 2023).

Most of these studies test the models employing tasks that have been widely used and have already proven to be effective in evaluating LLMs, including acceptability/plausibility judgment tasks, classification tasks and thematic fit estimation (Vassallo et al. 2018; Warstadt, Singh, and Bowman 2019; Kann et al. 2019; Seyffarth and Kallmeyer 2020; Kauf et al. 2022; Veres and Sampson 2023). Fewer works, however, adopt psycholinguistically inspired methodologies (Thrush, Wilcox, and Levy 2020; Metheniti, Van de Cruys, and Hathout 2020; Ettinger 2020; Samo et al. 2023; Wilson, Petty, and Frank 2023).

Even fewer studies directly compare LLMs' performance to human behavior. Notably, Ettinger (2020) assesses pre-trained BERT models on tasks targeting commonsense and pragmatic inference, semantic roles and event knowledge, category membership, and negation, using stimuli from psycholinguistic experiments and comparing the results to human data gathered through cloze tasks and N400 measurements. Another example is the work by Li et al. (2022), which examines the existence of argument structure constructions in LMs using adapted psycholinguistic tasks. Importantly, both studies point out that, although LLMs display some capacity for linguistic generalization, their performance on syntax-semantics interface phenomena does not fully align with human behavior — particularly regarding the semantic component, which appears more challenging than the syntactic one. Moreover, results suggest that both word frequency and the type of task significantly affect model performance.

### 2.1 Italian LLMs for the investigation of the Instrumental role in Italian

In this study, we consider different LLMs and compare their performance both to each other and to that of Italian speakers.

The evaluated pre-trained models are five Italian autoregressive models based on three distinct architectures (GPT-2, and Mistral). The set included GePpeTto (De Mattei et al. 2020), a GPT-2 small model trained from scratch on approximately 13 GB of Italian text (Italian Wikipedia and ItWac), as well as four sizes of the Minerva family (350M, 1B, 3B, 7B) (Orlando et al. 2024),trained on large-scale corpora containing at least 50% Italian. Differences within the Minerva family concern both the size and the composition of their training data. In terms of size, total token counts range from 70 B for Minerva-350M to 2.48 T for Minerva-7B. With respect to data composition, all models are trained on CulturaX; however, Minerva-7B is additionally trained on Italian Wikipedia, Wikisource, EurLex, the Gazzetta Ufficiale, the Gutenberg Project, and other sources (Orlando et al. 2024). As a result, this model is not only the largest but also the one exposed to the richest and most diverse set of Italian data.

The models selected for this study were chosen because at least 50% of their pretraining data consists of Italian text. This stands in contrast to larger and more commonly used multilingual models, which, although they do include Italian in their training corpora, typically do so in much smaller proportions. By focusing on models with substantial Italian-specific pretraining, we ensured a more controlled and linguistically relevant comparison, reducing potential confounds stemming from imbalanced multilingual training.

## 3. Instrumental role in Italian: an overview

### 3.1 Semantic definition and syntactic realization

The label *Instrument* refers to the semantic participant that contributes to the realization of the event denoted by the verb, after being manipulated by the Agent (Suozzi, Cardinaletti, and Lebani 2024). For instance, in a *cutting* event, an Agent manipulates a cutting object (e.g., scissors, a knife, etc.), causing it to come into contact with the Patient. The cutting object, in turn, brings about a change in the Patient, namely making it become cut. In an event of cutting, the cutting object is thus the Instrument, which plays a causal role in the event.

Other roles can by played by an Instrument in a given event (Marantz 1984; Jackendoff 1990; Schlesinger 1995; Koenig, Mauner, and Bienvenue 2003; Koenig et al. 2008; Rissman 2013; Rissman and Rawlins 2017). These instrumental sub-roles are: (i) necessary precondition, as in *Gianni entra nella stanza con la chiave* 'Gianni enters the room with the key', where the key (Instrument) enables Gianni to enter the room, but it does not cause him to step into it; and (ii) helping entity, as in *Gianni beve la Coca-Cola con la cannuccia* 'Gianni drinks the Coke with the straw', where the Instrument (the straw) only helps the realization of the event, which could take place without it as well (cf. Koenig, Mauner, and Bienvenue (2003) for the pragmatically-defined notion of "helping").

In Italian, at the syntactic level Instruments are primarily realized as prepositional phrases headed by the preposition *con* 'with' (*con*-PPs: see Example 1). Furthermore, *Con*-PPs can be anaphorically resumed by the coreferential clitic pronoun *ci* 'with it' (Example 2), which has an instrumental meaning in these cases. When realized through the *con*-PPs (Example 1) and clitic pronoun *ci* 'with it' (Example 2), Instruments are always syntactically optional, i.e., they can always be syntactically dropped without the resulting sentence being ungrammatical (Example 3).

**Example 1**

Luca rompe il vetro [con [la pietra]]
'Luca breaks the glass [with [the rock]]'

**Example 2**

(Con la pietra$_i$,) Luca [ci$_i$] rompe il vetro
LIT: (With the rock$_i$,) Luca with.it$_i$=breaks the glass
'(With the rock$_i$,) Luca breaks the glass [with it$_i$]'

**Example 3**

(Con la pietra$_i$,) Luca (ci$_i$) rompe il vetro
LIT: (With the rock$_i$), Luca with.it$_i$=breaks the glass
'(With the rock$_i$), Luca breaks the glass (with it$_i$)'

Another way to syntactically realize this semantic role in Italian is through *use*-structures, where the Instrument is the internal complement of the verb *usare* 'to use' (Example 4). In these constructions, Instruments must be syntactically realized, as illustrated by the ungrammaticality of Example 5.

**Example 4**

Luca usa [la pietra] per rompere il vetro
'Luca uses [the rock] to break the glass'

**Example 5**

Luca usa *(la pietra) per rompere il vetro
'Luca uses *(the rock) to break the glass'

In this study, we focus exclusively on the *con*-PPs strategy, for two main reasons. First, *con*-PPs represent the most prototypical syntactic realization of Instruments in Italian. Second, they provide greater insight into the semantic relationship between a verb and the lexical items that can function as its potential Instruments. The instrumental clitic *ci* always resumes a previously mentioned *con*-PP, making it less informative in this respect. Similarly, in *use*-structures, although the Instrument appears as an argument of the verb *usare* 'to use' at the syntactic level, its selection is semantically determined by the verb of the infinitival *to*-clause. For instance, in Example 4, it is the verb *rompere* 'to break' that selects the instrumental lexical item *pietra* 'rock', rather than the verb *usare* 'to use'.

### 3.2 Syntactic optionality and semantic recoverability

Syntactic optionality of Instruments has often been used as a case for considering them adjuncts (Dowty 1982; Jackendoff 1990; Rissman and Rawlins 2017). On the contrary, following Suozzi, Cardinaletti, and Lebani (2024), we claim that Instruments, despite being syntactically optional, are not adjuncts. Instead, we maintain that their omission is an instance of argument omission, similar to the widely-known phenomenon of indefinite null object (Example 6), and that, as such, it is ruled by semantic recoverability (Suozzi, Cardinaletti, and Lebani 2024).

**Example 6**

Luca ha già parcheggiato ∅
'Luca has already parked ∅'

Among the several factors that license or facilitate argument omission, semantic recoverability is recognized as the primary one (Kardos 2010; Hickman, Taylor, and Raskin 2016; Cappelli and Lenci 2020). Semantic recoverability is defined as the possibility for an omitted argument to be semantically recovered from the verb meaning,

either alone or in combination with its internal argument(s). This notion is linked to that of verbal semantic selectivity: the more a verb is selective (i.e., the fewer lexical items function as fillers of a given argument slot), the more an argument is recoverable.

Semantic recoverability, as defined thus far, rules argument omission in that (i) an argument can be omitted if and only if it is semantically recoverable from the verb meaning and (ii) the greater an argument's recoverability from the verb's meaning, the more likely it is to be omitted. To illustrate, consider the contrast between Example 6 and Example 7.

**Example 7**

*Luca ha già sollevato ∅
'*Luca has already lifted ∅'

The contrast in grammaticality between examples (6) and (7) arises from the differing semantic selectivity of *parcheggiare* 'to park' and *sollevare* 'to lift'. The former is highly selective, allowing only a limited set of semantically similar lexical items (e.g., car, van, etc.) as fillers of its internal argument slot. In contrast, *sollevare* 'to lift' is not selective, as a wide range of lexical items can serve as its internal argument. As a result, the internal argument of *parcheggiare* 'to park', is highly recoverable and, therefore, omissible, whereas that of *sollevare* is not.

Our main claim is that Instrument omission is ruled by semantic recoverability, as well. That is, Instruments are syntactically omitted because they are recoverable from the verb meaning. Moreover, the semantic recoverability of Instruments varies in degree. Building upon the classification of arguments into Shadow, Default, and True arguments, as proposed by Pustejovsky (1995b, 1995a) and later adapted to Italian by Jezek (2017), three types of Instruments that differ in their semantic recoverability are identified (Suozzi, Cardinaletti, and Lebani 2024): Shadow, Default and Open Instruments.

**Shadow Instruments** (SI) are maximally recoverable from the verb, which selects for an extremely narrow set of instrumental lexical items that in some cases consists of a singleton (maximum semantic selectivity). Their maximum recoverability holds for both verbs in isolation and within a syntactic context (Example 8). Verbs that license SI in Italian (henceforth, shadow-verbs) are, e.g., *pettinare* 'to comb', *sciare* 'to ski', *segare* 'to saw', etc.

**Example 8**

Spazzolare 'to brush' (⇒ INST: spazzola 'brush')

Il bambino spazzola la bambola (⇒ INST: spazzola)
'The child brushes the doll's hair (⇒ INST: brush)'

Since they are maximally recoverable, SI are always syntactically omitted, unless they are further modified ( *Gianni imburra il pane con il burro alle erbe* 'Gianni butters the bread with the herb butter') or subtyped (*Gianni imburra il pane con la margarina* 'Gianni butters the bread with margarine'). Otherwise, the sentence is redundant and less acceptable (??*Gianni imburra il pane con il burro* '??Gianni butters the bread with butter').

**Default Instruments** (DI) are recoverable from the verb alone, which selects for a restricted and semantically coherent set of instrumental lexical items in isolation (Example 9). In addition, DIs' semantic recoverability can be increased (i.e., they can

be can be shadowed) by the semantic interaction between the verb and its internal argument (Example 10). Verbs that license DI in Italian (henceforth, default-verbs) are, e.g., *tagliare* 'to cut', *sparare* 'to shoot', *pescare* 'to fish', etc.

**Example 9**

Tagliare 'to cut' (⇒ INST: Cutting Objects)

**Example 10**

Gianni taglia il prato$_1$/il pane$_2$ (⇒ INST: tagliaerba$_1$/coltello$_2$)
'Gianni cuts the lawn$_1$/the bread$_2$ (⇒ INST: lawnmower$_1$/knife$_2$)'

DI are usually omitted syntactically, because they are highly recoverable thanks to the interaction between the meaning of the verb and that of its internal argument. They are only realized syntactically when a single entity belonging to the selected set must be mentioned, or when they are further modified or subtyped.

Finally, **Open Instruments** (OI) are minimally recoverable from the verb in isolation, which selects for a broad and not (necessarily) semantically coherent set of entities. The semantic recoverability of certain OI is increased by the semantic interaction between the verb and its internal argument(s), but this is not always the case. To illustrate, consider Example 11 and Example 12: when the verbs are in isolation, OI are unrecoverable from the verb meaning. On the contrary, while the semantic interaction between the verb *distruggere* 'to destroy' and its internal arguments does not increase the OI semantic recoverability, this happens when the verb *aprire* 'to open' interacts with the internal argument *bottiglia* 'bottle'. Verbs that entail OI in Italian (henceforth, open-verbs) are *rompere* 'to break', *sporcare* 'to dirt', *aprire* 'to open', etc.

**Example 11**

Distruggere 'to destroy' (⇒ INST: martello 'hammer', dinamite 'dynamite', fuoco 'fire', etc.)

Gianni ha distrutto la parete (⇒ INST: martello, dinamite, fuoco, etc.)
'Gianni destroyed the wall' (⇒ INST: 'hammer', 'dynamite', 'fire', etc.)

**Example 12**

Aprire 'to open' (⇒ INST: chiave 'key', mani 'hands', cavatappi 'corkscrew', etc.)

Gianni ha aperto la bottiglia (⇒ INST: mani/cavatappi)
'Gianni opened the bottle' (⇒ INST: 'hands'/'corkscrew')

OI tend to be syntactically realized more frequently, as they are less easily recoverable from the meaning of the verb (± internal argument), and therefore not redundant.

Shadow-, default-, and open-verbs are progressively less semantically selective, which, in turn, makes SI, DI, and OI progressively less semantically recoverable. The presence of a syntactic context also influences their semantic recoverability, often enhancing it through the interaction between the verb's meaning and that of its internal argument.

The semantic recoverability of SI, DI, and OI is reflected in their syntactic realization: SI are maximally recoverable and therefore minimally realized at the syntactic

level; OI are minimally recoverable and thus more frequently syntactically realized. DI occupy an intermediate position with respect to semantic recoverability, but syntactic context plays a significant role in increasing it, to the extent that they are rarely realized syntactically (cf. Suozzi, Cardinaletti, and Lebani (2024), which investigates the production/omission patterns of Instruments through a comprehensive corpus analysis).

## 4. The capability of recovering Instruments: humans and LLMs

The primary research question addressed in this study is whether pre-trained language models assign relative probabilities to potential instrumental fillers (INST-lexical items) for shadow-, default- and open-verbs that are comparable to the frequencies of production found in Italian speakers. Specifically, we test whether LLMs semantically recover Instrumental meanings from shadow-, default-, and open-verbs, either presented in isolation (Experiment 1) or within richer syntactic contexts (Experiment 2), in a manner that parallels human sensitivity to verb selectivity.

Sensitivity to verb selectivity and Instrument recoverability are assessed differently in humans and LLMs. As for humans, we ask them to produce all the lexical items that they can be fillers of the instrumental slot for a given verb. LLMs are subsequently asked to assign a probability to each INST-lexical item. We then evaluate their performance in terms of alignment to human responses. Namely, we observe whether LLMs assign the higher probability to INST-lexical item most frequently produced by human, and whether they rank all the INST-lexical items produced by humans in a manner that parallels their frequency of production found in humans.

Both humans and models are expected to display graded probabilistic patterns across these verb types. In humans, this is reflected in the distribution of produced INST-lexical items; in models, in the distribution of log-probabilities assigned to human-produced INST-lexical items.

More specifically, under the hypothesis that shadow-, default-, and open-verbs are progressively less selective, we expect humans to produce a progressively increasing number of INST-lexical items for shadow-, default-, and open-verbs (shadow < default < open). Further, we expect the frequency of production to be progressively less skewed for shadow-, default-, and open-verbs: for shadow-verbs, one INST-lexical item is expected to be highly frequent, while the others less frequent. For default- and open-verbs, the frequency distribution is expected to be progressively more flat, with more INST-lexical items having similar frequency of production.

Concerning models, higher selectivity (shadow-verbs) is expected to result in a narrow probability distribution, with a few highly probable Instrumental items and many low-probability ones. Conversely, lower selectivity (open-verbs) should yield a flatter distribution, with a larger number of moderately probable Instruments. Default-verbs are expected to fall in between these two extremes.

### 4.1 Experiment 1: semantic recoverability with verbs in isolation

#### 4.1.1 Materials and Procedure

In order to assess the ability of Italian speakers to semantically recover the appropriate Instrument(s) from the verb meaning alone, a questionnaire named "The top 10 Instruments..." was created, inspired by an unpublished norming experiment by Annie Ledered (as cited in Resnik (1993, 86)).

For the questionnaire, 71 high-frequency verbs were selected from the Italian section of the CHILDES talkbank (MacWhinney 2000) and the *Nuovo Vocabolario di Base* (Chiari

and De Mauro 2012). This ensures that all verbs are known by all participants. Of these, 23 are shadow-, 26 are default-, and 22 are open-verbs. Some examples of the verbs used in the experiment are:

> **Shadow-verbs**: telefonare 'to phone', incollare 'to glue', sciare 'to ski', spazzolare 'to brush', pattinare 'to skate', salare 'to salt', etc.
>
> **Default-verbs**: scrivere 'to write', mangiare 'to eat', tagliare 'to cut', stirare 'to iron', pulire 'to clean', asciugare 'to dry', etc.
>
> **Open-verbs**: chiudere 'to close', andare 'to go', rompere 'to break', distruggere 'to destroy', costruire 'to build', lavorare 'to work (on)', etc.

Human participants were asked to write down the Instruments they thought could be used to perform the action denoted by each verb, from a minimum of 1 to a maximum of 10. An example of the instructions they saw for each verb is shown in Example 13.

**Example 13**

VERBO: **MARTELLARE**
Scrivi tutti gli strumenti che ti vengono in mente pensando all'azione descritta dal verbo («Cosa si usa per martellare?»), da un minimo di 1 a un massimo di 10.

VERB: **TO HAMMER**
Write down all the instruments that come to mind when thinking about the action described by the verb (e.g., «What is used for hammering?»), from a minimum of 1 to a maximum of 10.

The questionnaire was administered online through the *Qualtrics* platform[1]. Participants had no time limit to carry out the task, and each participant saw a balanced subset of verbs in a randomized order.

Based on human-generated data, we evaluated LLMs. We used four phrasal patterns that serve to syntactically express the Instrumental role in Italian. That is, pattern containing *con*-PPs, the clitic pronoun *ci*, and the verb *usare*. Within these patterns, placeholders (i.e., [VERB] and [INST-lexical item]) were systematically replaced with each verb and each corresponding human-made INST-lexical item associated with that verb. An example of the four patterns and their versions filled with the INST-lexical item *carte* 'cards' is reported in Table 1.
We extracted log-probabilities for each pattern filled with each INST-lexical item (Table 1) of a given verb. Therefore, for each INST-lexical item of each verb, we extracted four log-probabilities, which were then averaged to obtain a single log probability value for each INST-lexical item. All models were queried through the Hugging Face transformers library for Python. We opted to use predefined sentence frames with different syntactic configurations, combined with probability extraction from the LLMs' internal next-token distributions rather than prompt-based elicitation. This methodological choice was intended to minimize any dependence of the output on the specific phrasing of the query or on explicit task instructions. By avoiding direct prompts and instead evaluating the likelihood of full-sentence completions within controlled frames,

1 Qualtrics software, Version September-December, 2021 of Qualtrics. Copyright © 2020 Qualtrics. Provo, Utah, USA. Available at `https://www.qualtrics.com`.

**Table 1**
Italian patterns for Instrument encoding (left) and their instantiations with the verb *giocare* 'to play' and the INST-lexical item *carte* 'cards' (right).

| **Pattern** | **Pattern and filler** |
|---|---|
| [VERBO] con [item lessicale-STRUM]<br>'[VERB] with [INST-lexical item]' | Giocare con le carte.<br>'Play with cards.' |
| Con [item lessicale-STRUM] ci si può [VERBO]<br>'With [TOOL], with it one can [VERB]' | Con le carte ci si può giocare.<br>'With cards, with them one can play.' |
| Con [item lessicale-STRUM] si può [VERBO]<br>'With [INST-lexical item], you can [VERB]' | Con le carte si può giocare.<br>'With cards, you can play.' |
| Si può usare [item lessicale-STRUM] per [VERBO]<br>'You can use [INST-lexical item] to [VERB]' | Si può usare le carte per giocare.<br>'You can use cards to play.' |

we aimed to obtain responses that reflect the models' underlying lexical–semantic representations rather than prompt-specific behavior.

#### 4.1.2 Participants

89 Italian-speaking adults took part in the experiment, aged from 18 to more than 50 years.

Participants were recruited voluntarily through social networks. To take part in the experiment, each prospective participant read and signed an informed consent form, in compliance with Regulation (EU) 2016/679. All experimental procedures were approved by the Ethics Committee of Ca' Foscari University of Venice.

As for pre-trained models, we evaluated GePpeTto, as well as four sizes of the Minerva family (350M, 1B, 3B and 7B), all presented in Section 2.1.

#### 4.1.3 Response Coding

Prior to analysis, responses provided by Italian speakers were normed. First, non-instrumental lexical items - such as *destinatario* 'addressee', produced for the verb *scrivere* 'to write' - were excluded. Second, for shadow-verbs, instrumental lexical items denoting subtypes of the same instrument were merged and counted as a single item. For instance, *telefono fisso* 'landline', *telefono cellulare* 'mobile phone', and *smartphone* 'smartphone', produced for the verb *telefonare* 'to phone', were grouped under a single instrumental lexical item: *telefono* 'phone'.

#### 4.1.4 Metrics

Human responses served to obtain an overview of the semantic selectivity of the three verb classes (shadow, default, open). We indeed computed the *Unique Response Ratio*, defined as the number of distinct instruments produced by participants for each verb divided by the total number of collected responses.

To compare the performance of LLMs with that of human participants, we correlated model log-probabilities and human production frequencies. To this end, we used two metrics.

The first metric measures the proportion of cases in which the instrument most frequently produced by humans is also the one assigned the highest probability by the model. This provides an estimate of how often the model's top prediction aligns with the dominant human response. We refer to this proportion as *accuracy*. The chance level is computed by taking the average probability that the model assigns the first-rank position to the word that humans place in the first position.

The second metric involves calculating the Pearson's correlations between the tie-corrected ranks of human production frequencies and those derived from the models' probabilities. This allows us to assess the degree to which models and humans share a similar ranking structure over the set of possible INST-lexical items, beyond agreement on only the top choice.

#### 4.1.5 Results

Figure 1 displays the diversity of human responses across verb types, measured using the Unique Response Ratio for each verb. Shadow-verbs exhibit the lowest values, with a median of 0.140, a mean of 0.162, and a relatively narrow spread (IQR = 0.100, SD = 0.083). Default-verbs show intermediate values, with a median of 0.300 and a mean of 0.302, and a similar variability (IQR = 0.090, SD = 0.082). Open-verbs yield the highest ratios, with a median of 0.380, a mean of 0.377, and slightly larger variability (IQR = 0.105, SD = 0.111). Overall, the data indicate a gradual increase in the Unique Response Ratio from shadow- to default- and finally to open-verbs, as reflected in both central tendency and spread. This confirms that - for human speakers - shadow-verbs are maximally selective (and SI maximally recoverable), and that semantic selectivity and recoverability progressively decrease for default-verbs (and DI) and open-verbs (and OI).

Turning to the LLM's performance, (cf. Section 4.1.4), Figure 2 displays their accuracy scores.

For shadow-verbs (chance = 0.36), all models performed substantially above chance, with peak accuracy observed for Minerva-1B and Minerva-3B (0.692, a 0.332 advantage over chance). Notably, performance did not increase monotonically with model size: Minerva-7B showed a decline to 0.577, though still well above chance ($\Delta$ = 0.217). For default-verbs (chance = 0.07), models also exceeded chance, but the gains were smaller in absolute terms. Accuracy ranged from 0.148 (GePpeTto) to 0.444 (Minerva-3B), corresponding to improvements over chance of 0.078–0.374. The trend within this category suggests a generally positive effect of model size, peaking at Minerva-3B. Open-verbs (chance = 0.06) showed the most variability relative to chance. While GePpeTto, Minerva-350M, and Minerva-1B were only modestly above chance ($\Delta$ = 0.199, 0.199, 0.162), Minerva-7B reached 0.370 ($\Delta$ = 0.310), representing the largest distance from chance for this category. Overall, the results highlight that model improvements manifest differently within each verb type, and raw accuracy comparisons across categories can be misleading without accounting for category-specific chance levels.

Correlation analyses confirm that the relationship between human and model rankings is moderate or medium to strong. As reported in Table 2, coefficient scores range from approximately 0.43 to 0.69 across the different model–verb combinations, with the highest value obtained by Minerva-3B on shadow verbs ($\rho = 0.694$). Notably, Minerva-3B achieves the highest correlation in *all* verb categories, consistently outperforming the larger Minerva-7B. These correlations suggest that models capture some broad tendencies in human ranking distributions but do not reproduce human probabilistic structure in a fine-grained manner.

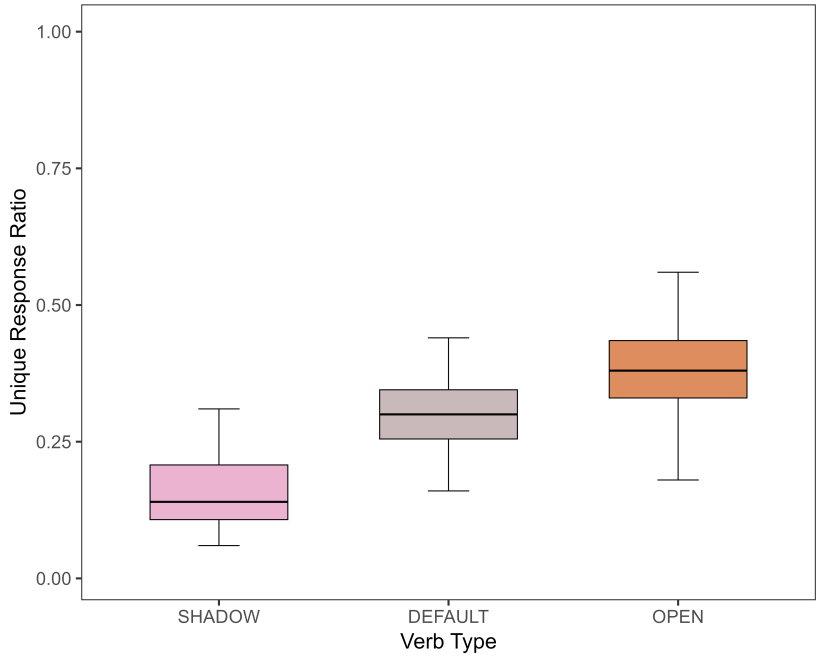


**Figure 1**
Boxplots showing the ratio between the number of unique instruments and the total number of responses for each verb type (shadow, default, and open).

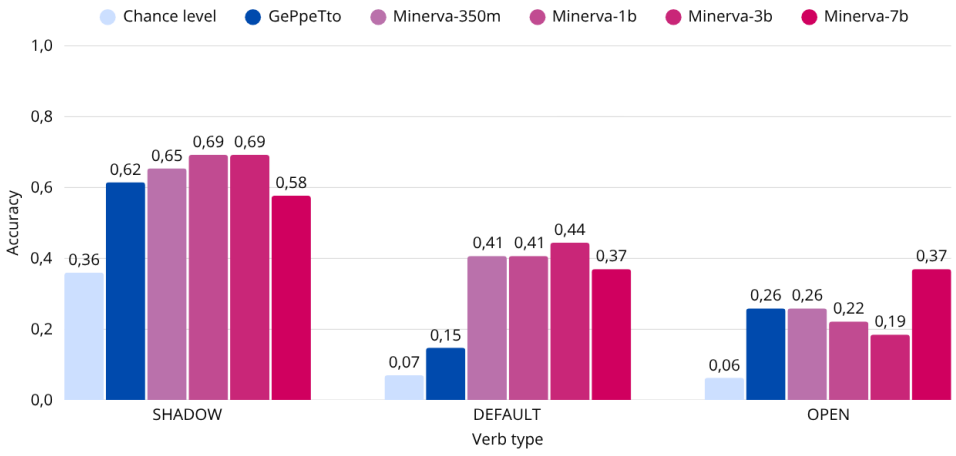


**Figure 2**
Accuracy scores for each model across the three verb types (Shadow, Default, and Open).

**Table 2**
Correlation between the by-verb ranking derived from human production frequency and the by-verb ranking derived from the model's log-probability. All correlations reported were statistically significant with $p < 0.001$.

| | **Shadow** | **Default** | **Open** |
|---|---|---|---|
| GePpeTto | 0.577 | 0.430 | 0.494 |
| Minerva-350M | 0.650 | 0.482 | 0.549 |
| Minerva-1B | 0.679 | 0.462 | 0.530 |
| Minerva-3B | **0.694** | **0.538** | **0.565** |
| Minerva-7B | 0.675 | 0.507 | 0.553 |

#### 4.1.6 Discussion

This first experiment aimed at measuring the ability to semantically recover the appropriate Instrument(s) from the verb meaning, when verbs are presented in isolation.

Our findings show that SI are more easily recovered than DI and OI by all LLMs under investigation. This "ease" probably is a byproduct of the human-generated data: Italian speakers produced fewer INST-lexical items for shadow-verbs than for default- and open-verbs. Among these items, one is by far the most frequently produced - typically the noun incorporated into the verb (e.g., *sega* 'saw' for *segare* 'to saw') - whereas the remaining items are produced much less frequently, i.e., they are less likely fillers of the Instrumental slot for these verbs. Since only a few INST-lexical items need to be ranked for these verbs, and one of them is also perceived by humans as the most likely filler, the ranking procedure is easier for all models. This trend holds for both accuracy and the correlation between by-verb rankings. For SI, moreover, no clear advantage is observed in the performance of Minerva-7B, despite it being the largest model and being trained on data that differ quantitatively and qualitatively from those of the other models (Orlando et al. 2024). On the other hand, the task becomes more difficult with DI and OI: for default-, and especially for open-verbs, Italian speakers produce a larger number of INST-lexical items, often with similar frequencies. For these verbs, several INST-lexical items are almost equally plausible candidates for filling the Instrument slot, particularly in the case of open-verbs. Indeed, all LLMs show a drop in both accuracy and the correlation between by-verb rankings, but a slight advantage of larger models is observed. When focusing on the accuracy for open-verbs, Minerva-7B displays a clear advantage. Its size and training likely allow it to better identify the INST-lexical item most frequently produced by human speakers. However, this advantage is not reflected in the correlation of the by-verb rankings, signaling that while it manages to identify the most likely candidates, the ranking it provides does not fully aligns with human perceptions. This further supports the view that larger scale does not necessarily yield closer alignment with human behavior.

Overall, our results suggest that all models perform well on easier tasks (i.e., with shadow-verbs), but they continue to struggle in semantically recovering the appropriate INST-lexical items when verbs are less selective and a larger number of candidates must be considered. In these cases, larger models trained on quantitatively and qualitatively richer input exhibit a clear advantage concerning top-rank predictions.

### 4.2 Experiment 2: semantic recoverability with verbs in a syntactic context

Oue second experiment was designed to explore the ability of pre-trained models to semantically recover the appropriate Instrument(s) from the verb's meaning when presented within a syntactic context, by comparing them to both Italian speakers. As suggested in Section 3.2, the semantic interaction between a verb and its internal argument enhances the verb's semantic selectivity, which, in turn, makes the Instrument(s) more semantically recoverable.

#### 4.2.1 Materials

The verbs used in the task(s) created for this experiment are a subset of those used in Experiment 1. Namely, we selected 8 shadow-, 8 default- and 8 open-verbs, each appearing twice in the task, resulting in a total of 48 experimental items.

The task administered to Italian-speaking adults is an elicited production task. Each item consists of a sentence where the Instrument is omitted (stimulus), and a *wh*-question that elicits the production of a *con*-PP containing the appropriate instrumental lexical item. Crucially, no visual stimuli are used for this task, so as not to provide any additional cue about the instrument used to perform the action denoted by the verb. Three example items can be appreciated in Example 14, one for the shadow-verb *incollare* 'to glue', one for the default-verb *tagliare* 'to cut', and one for the open-verb *rompere* 'to break'.

**Example 14**

La bambina incolla la foto. Con cosa incolla la foto?
'The girl glues the photo. What does she glue the photo with?'

Il nonno taglia il pane. Con cosa taglia il pane?
'The grandfather cuts the bread. What does he cut the bread with?'

La bambina rompe la finestra. Con cosa rompe la finestra?
'The girl breaks the window. What does she break the window with?'

In this experiment, to test the models we extracted, for each verb, the probability assigned to every human-produced instrument within three sentence patterns (as illustrated in Table 3). These patterns were designed to represent the range of syntactic configurations with which instrumental roles are expressed in Italian.

**Table 3**
Italian patterns for Instrument encoding (left) and their instantiations with the verb *tagliare* 'to cut' and the INST-lexical item *forbici* 'scissors' (right).

| Pattern | Pattern and filler |
|---|---|
| Il nonno taglia il pane con [item lessicale-STRUM]<br>'Grandpa cuts the bread with [INST-lexical item].' | Il nonno taglia il pane con le forbici.<br>'Grandpa cuts the bread with scissors.' |
| Il nonno taglia il pane usando [item lessicale-STRUM]<br>'Grandpa cuts the bread using [INST-lexical item].' | Il nonno taglia il pane usando le forbici.<br>'Grandpa cuts the bread using scissors.' |
| Con [item lessicale-STRUM] il nonno ci taglia il pane<br>'With [INST-lexical item], Grandpa cuts the bread with it.' | Con le forbici il nonno ci taglia il pane.<br>'With scissors, Grandpa cuts the bread with them.' |

#### 4.2.2 Participants and Procedure

The experiment involved (i) 25 Italian-speaking adults; (ii) five models: GePpeTto, Minerva-350M, Minerva-1B, Minerva-3B, and Minerva-7B.

Participants were recruited voluntarily through social networks. The elicited production task was administered online, using the *Gorilla Experiment Builder* (www.gorilla.sc) to create and host the experiment (Anwyl-Irvine et al. 2020). The oral administration was maintained, as the experimenter recorded herself while reading the stimuli, which were imported into the platform and orally presented. Participants answered orally, their responses were recorded and subsequently transcribed by the experimenter.

To take part in the experiment, each prospective participant read and signed an informed consent form, in compliance with Regulation (EU) 2016/679. All experimental procedures were approved by the Ethics Committee of Ca' Foscari University of Venice.

#### 4.2.3 Metrics

In Experiment 2, we employed the same metrics as in Experiment 1, described in detail in Section 4.1.4 above.

#### 4.2.4 Results

Figure 3 compares the distribution of Unique Response Ratios across shadow-, default-, and open-verbs. As in Figure 1, where verbs were presented in isolation, the current data again reveal a clear progression across categories: shadow-verbs show the lowest ratios, with a median of 0.065 and a narrow interquartile range (IQR = 0.025), indicating very limited variability; dfault-verbs present intermediate values, with a median of 0.125 and greater dispersion (IQR = 0.073); open-verbs display the highest ratios, with a median of 0.170 and the widest overall range (0.250, from 0.110 to 0.360). Overall, this boxplot confirms the same upward trend observed previously: unique response ratios increase consistently from shadow- to default- to open-verbs, both in central tendency and variability. This confirms that - for human participants - a difference in semantic selectivity of shadow-, default-, and open-verbs (as well as in the semantic recoverability of the corresponding Instruments - SI, DI, OI) remains. The role of the linguistic context in increasing both is more evident for default-verbs and DI: shadow-verbs are maximally selective in isolation (SI thus being maximally recoverable), while the context has a reduced impact on the semantic recoverability and selectivity of open-verbs and OI than on default-verbs and DI.

Model performance in this task exhibited distinct patterns across verb types, with differences in how far accuracy exceeded (or fell below) chance levels (Figure 4). For shadow-verbs (chance = 0.44), performance varied substantially across models. GePpeTto and Minerva-350M were slightly below chance (0.333, $\Delta$ = –0.107), whereas Minerva-1B and Minerva-3B exceeded chance (0.667, $\Delta$ = 0.227; 0.500, $\Delta$ = 0.060, respectively). The largest model, Minerva-7B, showed the highest accuracy (0.833, $\Delta$ = 0.393), indicating a strong within-category improvement for larger models, though gains were not strictly linear. For default-verbs (chance = 0.19), most models performed at or below chance, with GePpeTto, Minerva-350M, and Minerva-1B all at 0.111 ($\Delta$ = –0.079). Performance improved for Minerva-3B (0.333, $\Delta$ = 0.143) and remained moderate for Minerva-7B (0.222, $\Delta$ = 0.032), suggesting that only some models reliably exceed chance for this less constrained verb category. Open-verbs (chance = 0.12) showed a strikingly different pattern. Smaller models substantially outperformed chance (GePpeTto = 0.875, $\Delta$ = 0.755; Minerva-350M = 0.750, $\Delta$ = 0.630), while intermediate models decreased toward

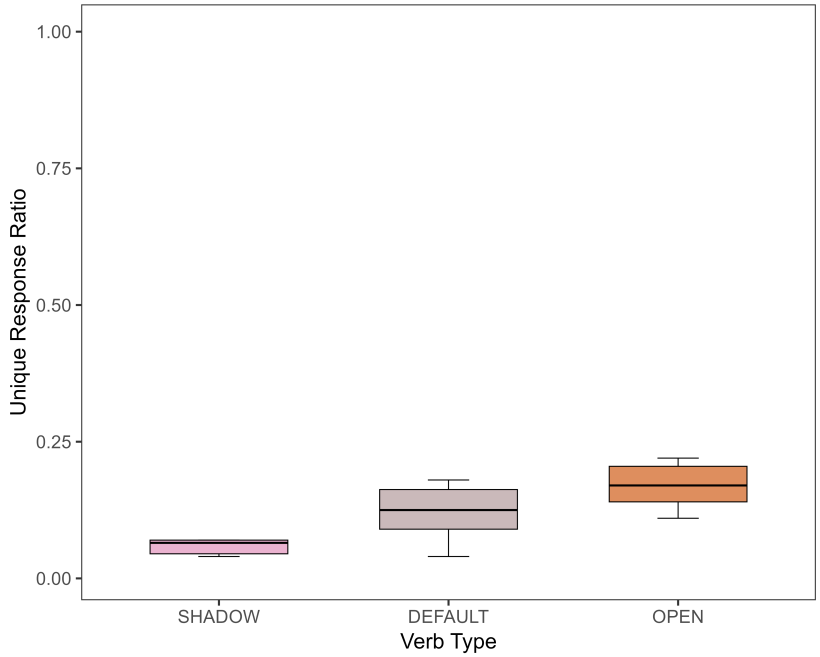


**Figure 3**
Boxplots showing the ratio between the number of unique instruments and the total number of responses for each verb type (Shadow, Default, and Open).

chance (Minerva-1B = 0.500, $\Delta$ = 0.380; Minerva-3B = 0.375, $\Delta$ = 0.255), and Minerva-7B plateaued at 0.375 ($\Delta$ = 0.255). This suggests that for open verbs, larger models do not necessarily improve performance and may even regress relative to smaller models, highlighting non-linear effects of model size.

Overall, these results emphasize that trends within each verb category are more informative than raw cross-category comparisons. Distance from chance provides a clearer picture of model competence for each type of semantic constraint.

The correlations reported in Table 4 reflect the relationship between the ranking derived from human production frequencies and the ranking based on the mean log-probabilities assigned by the LLMs to the four phrasal patterns. For shadow-verbs, correlations ranged from modest to strong, with GePpeTto and Minerva-350M showing moderate alignment with human responses ($\rho$ = 0.427), while Minerva-7B exhibited the strongest correspondence ($\rho$ = 0.739, $p < 0.001$). Default-verbs generally showed weaker correlations overall, with only Minerva-3B reaching a moderate, significant correlation ($\rho$ = 0.384, $p < 0.001$), indicating that models capture human preferences less reliably for this less constrained category. Open-verbs consistently showed high correlations across most models, particularly GePpeTto, Minerva-350M, and Minerva-1B ($\rho$ = 0.627–0.673, all $p_s < 0.001$), suggesting that model predictions align closely with human choices for verbs that allow multiple instrument options. These results highlight that model-human agreement is strongly dependent on verb type: shadow- and open-verbs tend to produce higher correlations, whereas default verbs remain challenging. Moreover,

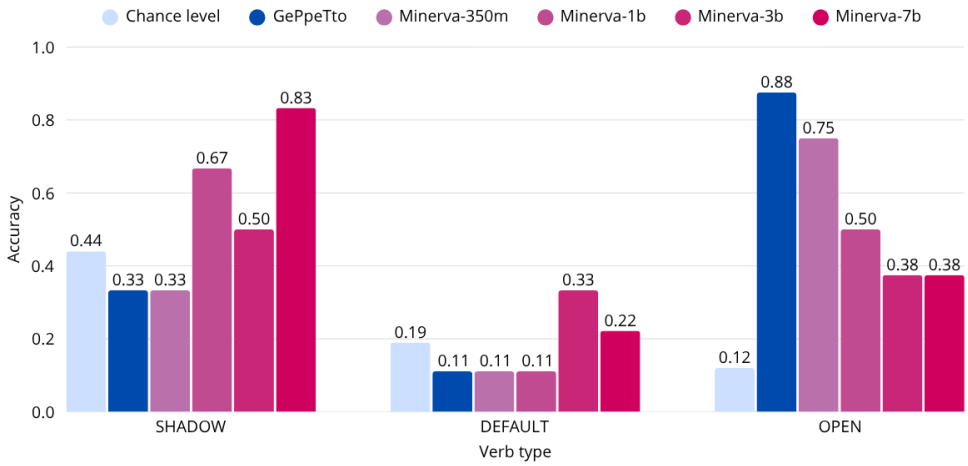


**Figure 4**
Accuracy scores for each model across the three verb types (Shadow, Default, and Open).

the strength of correlations does not scale linearly with model size, reflecting non-monotonic effects of capacity on capturing human-like preferences.

**Table 4**
Correlation between the by-verb ranking derived from human production frequency and the by-verb ranking derived from the model's log-probability (p-values indicated as $p < 0.05$ *, $p < 0.01$ **, and $p < 0.001$ ***).

| Model | Shadow | Default | Open |
|---|---|---|---|
| GePpeTto | 0.427 | 0.224 | 0.627*** |
| Minerva-350M | 0.427 | 0.307 * | **0.673**** |
| Minerva-1B | 0.538* | 0.262 | 0.632*** |
| Minerva-3B | 0.222 | **0.384**** | 0.588*** |
| Minerva-7B | **0.739**** | 0.304* | 0.559*** |

### 4.2.5 Discussion

Experiment 2 aimed to assess the ability to semantically recover the appropriate Instrument(s) from a verb when interacting with its internal argument. As hinted at in Section 3.2, and as confirmed by human-generated data, semantic recoverability of SI, DI and OI is increased by the presence of a syntactic context. This holds particularly for DI (cf. Figures 1 and 3), while for SI and SI the effect is smaller: indeed, shadow-verbs are already maximally selective in isolation, while open-verbs remain less semantically selective even when inserted in a syntactic context.

Turning now to LLMs, a different pattern of accuracy emerges in this experiment. While in Experiment 1 shadow-verbs were the easiest to manage for smaller and larger models, in this experiment the scenario is different. The only models whose accuracy is above chance are Minerva-1B, -3B and especially Minerva-7B, which clearly outper-

forms all other models. Concerning the correlation of by-verb rankings, only Minerva-1B and Minerva-7B exceed a 0.5 value, with Minerva-7B diplaying a significantly higher correlation.

For DI, only the accuracy achieved by Minerva-3B and Minerva-7B are above chance, and the correlation of by-verb rankings also display lower values. A reverse pattern is observed for open-verbs. Here, all models perform significantly above chance - with smaller models outperforming larger ones. This also holds for correlations of by-verb rankings.

We interpret these results in light of two factors: (i) the role played by the interaction between the meaning of the verb and that of its syntactic context in determining - for each experimental item - the most suitable INST-lexical item to fill the instrumental argument slot; and (ii) the pattern of syntactic production/omission of the Instrument in spontaneous speech, depending on its semantic recoverability (introduced in Section 3.2).

For shadow-verbs, the Instrument is almost always omitted in spontaneous speech, and the linguistic context plays a (albeit limited) role in specifying the most appropriate INST-lexical item. In this case, larger models appear to be advantaged: this advantage reflects their ability to semantically recover SI even though they occur very rarely in the training input. In other words, this demonstrates the development of a linguistic capacity that goes beyond what the model observes in the input.

For open-verbs, the effect of syntactic context is greater than for SI but smaller than for DI. Conversely, in spontaneous speech, OI are the Instruments most frequently produced. For these verbs, the best performance is achieved by smaller models, although all models perform above chance. This pattern likely reflects the fact that while all models tend to do so, smaller models rely more heavily on the training input, in which OI actually occur more frequently, making them more likely to identify the appropriate Instrument for these verbs.

Finally, DI appear to be the most difficult to semantically recover: for these Instruments, the linguistic context plays a crucial role in increasing their semantic recoverability, and the Instruments themselves occur very rarely in spontaneous speech (at a rate comparable to SIs). These verbs are therefore the best candidates to observe the model's ability to simultaneously handle semantic interactions between the verb and its internal argument and the lack of evidence in the input. The data for these verbs indicate that they are the most challenging overall and reveal a slight advantage of larger models over smaller ones.

This experiment demonstrates that the simultaneous management of scarce input evidence and internal sentence-level semantic interactions remains a complex task for LLMs. Nonetheless, it also highlights that larger models trained on more diverse input show a tendency to develop enhanced generalization abilities and more refined semantic knowledge.

## 5. Conclusions

In this study, we examined the extent to which language models can semantically recover Instruments from verb meaning, both in isolation (Experiment 1) and within a syntactic context (Experiment 2). To this end, we compared the performance of GePpeTto and four Minerva models of increasing size (350M, 1B, 3B, and 7B) with that of Italian speakers. We assessed model performance using (i) accuracy and (ii) correlations of by-verb rankings, in order to determine whether LLMs align with humans in their ability to recover Instrument meanings.

Experiment 1 showed that LLMs benefit from highly selective verbs when they must rank a small set of lexical items, one of which is strongly favored. Their performance, however, decreases for less selective verbs. The only exception is the relatively high accuracy achieved by Minerva-7B for open-verbs, which are challenging for all other models. Yet this advantage - likely reflecting the model's larger size and richer training corpus - does not extend to correlations of by-verb rankings. This pattern supports the view that increased scale does not necessarily yield closer alignment with human behavior.

Experiment 2 specifically tested models' ability to recover appropriate Instruments when verb meaning interacts with the meaning of the internal argument. Here, shadow- and especially default-verbs are more challenging for smaller models, whereas open-verbs are the easiest. We interpret these outcomes in light of (i) the role of linguistic context in shaping the semantic recoverability of SI, DI, and OI, and (ii) the patterns of production and omission of SI, DI, and OI in the input. For smaller models, open-verbs are easier: OI appear more frequently in spontaneous speech and their semantic recoverability is only moderately influenced by linguistic context. In contrast, SI are difficult because, although linguistic context does not substantially increase their recoverability, they tend to be omitted in spontaneous speech. DI are the most challenging, as they are frequently omitted and their recoverability strongly depends on the interaction between verb meaning and internal-argument meaning. These findings suggest that smaller models rely more heavily on distributional properties of the training input to recover Instruments. Larger models, by contrast, appear to develop a more robust capacity for generalization and thus greater independence from the input.

Considering Experiments 1 and 2 together, we observe that semantic recoverability is easier for LLMs when verbs are presented in isolation than when they must integrate verb meaning with the meaning of the internal argument.

Finally, it is worth noting a consistent discrepancy between accuracy and by-verb ranking correlations. When correlations are considered, LLMs' alignment with humans is systematically lower than when accuracy is considered. In other words, even when models' top predictions match those of human participants, they do not faithfully reproduce the fine-grained probabilistic structure of human judgments.

In conclusion, our results show that LLMs exhibit a degree of ability to semantically recover appropriate Instruments, and that this ability depends on verb selectivity, the presence or absence of linguistic context, and model size. In particular, however, LLMs are able at predicting the likeliest candidates for a given Instrumental slot, but do not strongly align with humans' probabilistic structure.

**References**

Anwyl-Irvine, Alexander L., Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52(1):388–407.

Blank, Idan A. 2023. What are large language models supposed to model? *Trends in Cognitive Sciences*, 27:987–989.

Cappelli, Giulia. 2022. *Implicit indefinite objects at the syntax-semantics-pragmatics interface: a probabilistic model of acceptability judgments*. Ph.D. thesis, Scuola Normale Superiore.

Cappelli, Giulia and Alessandro Lenci. 2020. PISA: A measure of preference in selection of arguments to model verb argument recoverability. In Iryna Gurevych, Marianna Apidianaki, and Manaal Faruqui, editors, *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics (SEM* 2020)*, pages 131–136, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Chiari, Isabella and Tullio De Mauro. 2012. The new basic vocabulary of Italian: problems and methods. *Statistica Applicata - Italian Journal of Applied Statistics*, 22:21–35.

Croft, William. 1998. Event structure in argument linking. In Miriam Butt and Wilhelm Geuder, editors, *The projection of arguments: Lexical and compositional factors*. Center for the Study of Language and Information, Stanford, CA, pages 21–63.

De Mattei, Lorenzo, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 97–104, Bologna, Italy, March 1–3, 2021. CEUR Workshop Proceedings.

Dowty, David. 1982. Grammatical relations and Montague grammar. In Pauline Jacobson and Geoffrey K. Pullum, editors, *The Nature of Syntactic Representation. Synthese Language Library, vol 15*. Springer, pages 79–130.

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ferretti, Todd R., Ken McRae, and Andrea Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44:516–547.

Hickman, Louis, Julia M. Taylor, and Victor Raskin. 2016. Direct object omission as a sign of conceptual defaultness. In *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2016)*, pages 516–521, Hilton Key Largo Resort, Key Largo, Florida, USA, May 16–18, 2016. Proceedings available online at http://www.aaai.org/Press/Proceedings/flairs16.php and http://aaai.org/Library/FLAIRS/flairs16contents.php.

Jackendoff, Ray. 1990. *Semantic structures*. MIT Press, Cambridge, MA.

Jezek, Elisabetta. 2003. *Classi di verbi tra semantica e sintassi*. ETS Edizioni, Pisa, IT.

Jezek, Elisabetta. 2017. Generative lexicon theory and lexicography. In Patrick Hanks and Gilles-Maurice De Schryver, editors, *International Handbook of Modern Lexis and Lexicography*. Springer, pages 1–21.

Kamide, Yuki, Gerry T.M Altmann, and Sarah L. Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, 49:133–156.

Kann, Katharina, Alex Warstadt, Adina Williams, and Samuel R. Bowman. 2019. Verb argument structure alternations in word and sentence embeddings. In *Proceedings of the Society for Computation in Linguistics (SCiL 2019)*, pages 287–297, New York City, New York, USA, January 3–6, 2019.

Kardos, Eva. 2010. The argument expression of change-of-state verbs and pseudo-transitive verbs. *Bergen Language and Linguistics Studies*, 1.

Kauf, Carina, Anna A. Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2022. Event knowledge in large language models: The gap between the impossible and the unlikely. *Cognitive Science*, 47:1–40.

Kodner, Jordan, Sarah Payne, and Jeffrey Heinz. 2023. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). *arXiv preprint arXiv:2308.03228*.

Koenig, Jean-Pierre, Breton Bienvenue, Gail Mauner, and Kathy Conklin. 2008. What with? The anatomy of a (proto)-role. *Journal of Semantics*, 25(2):175–220.

Koenig, Jean-Pierre, Gail Mauner, and Breton Bienvenue. 2003. Arguments for adjuncts. *Cognition*, 89:67–103.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Li, Bai, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 7410–7423, Dublin, Ireland, May. Association for Computational Linguistics. May 22–27, 2022.

Linzen, Tal and Marco Baroni. 2022. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition*. Lawrence Erlbaum Associates.

Marantz, Alec. 1984. *On the Nature of Grammatical Relations*. MIT Press, Cambridge, MA.

McRae, Ken and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistic Compass*, 3:1417–1429.

McRae, Ken, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312.
Medina, Tamara N. 2007. *Learning Which Verbs Allow Object Omission: Verb Semantic Selectivity and the Implicit Object Construction*. Ph.D. thesis, John Hopkins University.
Metheniti, Eleni, Tim Van de Cruys, and Nabil Hathout. 2020. How relevant are selectional preferences for transformer-based language models? In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 1266–1278, Barcelona, Spain (Online), December 8–13, 2020. International Committee on Computational Linguistics.
Orlando, Riccardo, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva llms: The first family of large language models trained from scratch on italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy, December 4–6, 2024. CEUR Workshop Proceedings.
Piantadosi, Steven. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz Preprint. url: https://lingbuzz.net/lingbuzz/007180*.
Pustejovsky, James. 1995a. *The Generative Lexicon*. MIT Press, Cambridge, MA.
Pustejovsky, James. 1995b. Linguistic constraints on type coercion. In Patrick Saint-Dizier and EvelynEditors Viegas, editors, *Computational Lexical Semantics*, Studies in Natural Language Processing. Cambridge University Press, page 71–97.
Resnik, Philip. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships*. Ph.D. thesis, University of Pennsylvania.
Rissman, Lilia. 2013. *Event Participant Representations and the Instrumental Role: A Cross-Linguistic Study*. Ph.D. thesis, John Hopkins University, Baltimore.
Rissman, Lilia and Kyle Rawlins. 2017. Ingredients of instrumental meaning. *Journal of Semantics*, 34(3):507–537.
Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
Samo, Giuseppe, Vivi Nastase, Chunyang Jiang, and Paola Merlo. 2023. BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 12276–12287, Singapore, December 10–14, 2023. Association for Computational Linguistics.
Schlesinger, Izchak M. 1995. *Cognitive Space and Linguistic Case: Semantic and Syntactic Categories in English*. Cambridge University Press, Cambridge, UK.
Seyffarth, Esther and Laura Kallmeyer. 2020. Corpus-based identification of verbs participating in verb alternations using classification and manual annotation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 4044–4055, Barcelona, Spain (Online), December 8–13, 2020. International Committee on Computational Linguistics.
Suozzi, Alice, Anna Cardinaletti, and Gianluca E. Lebani. 2024. On the argument status of instruments in italian. In Mia Batinić Angster and Marco Angster, editors, *The verbal kaleidoscope: perspectives on the syntax and semantics of verbs*. Morepress, Unieversity of Zadar, chapter 10, pages 261–301.
Thrush, Tristan, Ethan Wilcox, and Roger Levy. 2020. Investigating novel verb learning in BERT: Selectional preference classes and alternation-based syntactic generalization. In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 265–275, Online (in conjunction with EMNLP 2020), November 19, 2020. Association for Computational Linguistics.
Tjuatja, Lindia, Emmy Liu, Lori Levin, and Graham Neubig. 2023. Syntax and semantics meet in the "middle": Probing the syntax-semantics interface of LMs through agentivity. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 149–164, New York City, NY, USA, June 1–3, 2023. Association for Computational Linguistics.
Vassallo, Paolo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2018. Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of

Argument Typicality. In *Proceedings of the LREC 2018 Workshop on Linguistic and Neurocognitive Resources (LiNCR)*, Miyazaki, Japan, May.

Veres, Csaba and Jennifer Sampson. 2023. Self supervised learning and the poverty of the stimulus. *Data & Knowledge Engineering*, 147.

Veres, Csaba and Bjørn Helge Sandblåst. 2019. A machine learning benchmark with meaning: Learnability and verb semantics. In Jing Liu and John Bailey, editors, *AI 2019: Advances in Artificial Intelligence*, volume 11919 of *Lecture Notes in Computer Science*, pages 435–447. Springer, Cham.

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Wilcox, Ethan G., Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Wilson, Michael, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.