Evaluating Lexical Proficiency in Neural Language Models

Cristiano Ciaccio, Alessio Miaschi, Felice Dell'Orletta
Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC)
ItaliaNLP Lab, Pisa
{name.surname}@ilc.cnr.it

Abstract

We present a novel evaluation framework designed to assess the lexical proficiency and linguistic creativity of Transformer-based Language Models (LMs). We validate the framework by analyzing the performance of a set of LMs of different sizes, in both mono- and multilingual configuration, across tasks involving the generation, definition, and contextual usage of lexicalized words, neologisms, and nonce words. To support these evaluations, we developed a novel dataset of lexical entries for the Italian language, including curated definitions and usage examples sourced from various online platforms. The results highlight the robustness and effectiveness of our framework in evaluating multiple dimensions of LMs' linguistic understanding and offer an insight, through the assessment of their linguistic creativity, on the lexical generalization abilities of LMs[1].

1 Introduction

Recent advancements in Natural Language Processing have been significantly shaped by the Deep Learning tsunami (Manning, 2015) and the introduction of Transformer-based Language Models (Vaswani et al., 2017). As a result, several studies have been conducted to investigate the potential of such models in numerous tasks, downstream applications (Hendrycks et al., 2021b,a; Beeching et al., 2023) and their linguistic competencies (Gauthier et al., 2020; Waldis et al., 2024). On another front, some studies focused on investigating the abilities of these models in tasks related to lexical proficiency. Among these, Xu et al. (2024) and Aljaafari et al. (2024) tested LLMs on the reverse dictionary (RD) task, which involves generating words that match a given description D (Siddique and Beg, 2023), to probe their capacity for conceptual inference. To the best of our knowledge, only one work (Zheng et al., 2024, *Neo-Bench*) has examined the lexical proficiency of LMs in linguistic contexts that extend beyond commonly lexicalized words. Despite the efforts, lexical proficiency in LMs remains an underlooked topic. This is significant since linguistic generalization tasks such as defining and generating novel words require complex morphological understanding, linguistic creativity, commonsense knowledge and the ability to generalize over seen concepts (Fauconnier and Turner, 2003), thus serving as a key playground to assess the generalization capabilities of LMs. For example, defining *seismophony* as "the sound made by earthquakes", or vice versa, generating *seismophony* from the definition, entails that an LM learned and generalized the meaning shift caused by the *-phony* suffix. This also requires identifying the lexical base (*seism*), linking it correctly to the concept of *earthquake* and deriving a novel concept through the morphological combination of the two morphemes. Hence, to be executed efficiently, these generalization tasks require both a strong lexical and morphological grounding: **learning word formation rules plays an essential role in achieving abilities that can ease the understanding of linguistically motivated out-of-vocabulary** (OOV) **words**.

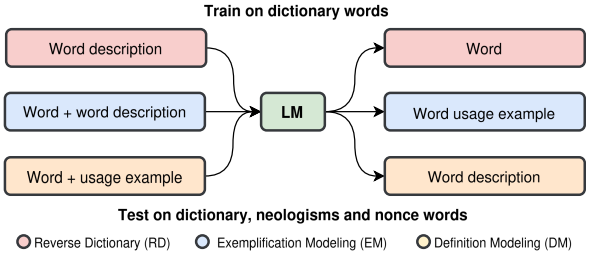

Figure 1: Overview of the proposed framework and the lexical proficiency tasks.

Building upon these premises, we propose a framework to (stress-)test the lexical abilities of

[1]The resources are available at the following repository: https://github.com/snizio/Lexical-Proficiency.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1267–1286

LMs. Specifically, we focus on the Italian language and employ T5 models (Raffel et al., 2019), both mono and multi-lingual and at different sizes, to assess their ability to: **generate**, **define**, and **coherently use lexicalized words** (see Figure 1), as well as to extend these capabilities to **neologisms** and **nonce words**, where nonce words should be interpreted as unattested linguistic artifacts (*H-Creative*, Boden, 2004). Furthermore, we assess the creative dimensions of the generated nonce words leveraging the Optimal Innovation Hypothesis (Giora et al., 2004), which states that the pleasure of an optimally innovative stimulus is in function of their degree of novelty and the automatic recoverability of a salient response related to that stimulus. Our approach is designed to answer the following questions: how do language models perform on lexical proficiency tasks? How do model size and pre-training language (mono- vs. multi-lingual) influence LMs' lexical abilities? To what extent do neologisms affect performance? Can language models approximate word-formation rules to successfully define, generate, and use nonce words that are meaningful and linguistically plausible?

To conduct the experiments, we **built a novel resource of lexical entries**, along with their definitions and usage examples, derived from different online sources.

To the best of our knowledge, this is the first time a unified framework has been proposed to test a language model's abilities in the generation, definition, and contextualized usage of words across various lexical settings, ranging from common words to neologisms and nonce words.

Contributions In this paper we: (i) propose a novel evaluation framework for assessing the lexical proficiency of LMs across tasks involving on the generation, definition and usage of words; (ii) develop a new lexical resource that supports such evaluations; (iii) evaluate the performance of different LMs focusing on the impact of model sizes, pre-training languages, linguistic settings and tasks; (iv) perform a human evaluation to quantify the degree of linguistic creativity exhibited by the LMs.

2 Related Work

In this section, we adopt the nomenclature from Veale and Butnariu (2006) when referring to the the creation of lexical innovations, *predictive lexicology*, and the definition of novel words, *explanatory lexicology*. We propose to distinguish between *open vocabulary* and *closed vocabulary* approaches for the Reverse Dictionary task, where the former refers to ranking a predefined vocabulary set of target words given a definition and the latter is not limited to a fixed vocabulary set and instead involves generation rather than ranking or prediction.

2.1 Reverse Dictionary (RD)

A reverse dictionary (Sierra, 2000) takes the description of a target lemma as input and outputs one or more lemmas that match the input description. Such systems have significant practical value in aiding for conditions such as anomia (Benson, 1979) and the *Tip of the Tongue* phenomenon (Brown and McNeill, 1966). Early approaches consisted mainly of string-matching methods with hand-crafted features (Zock and Bilac, 2004). The task evolved as a text-to-vector closed vocabulary retrieval problem, encoding the input query into the embedding space and rank a fixed set of vocabulary words whose embeddings are closest to the input representation (Hill et al., 2016; Zhang et al., 2020; Mickus et al., 2022). Recently, researchers started to leverage pre-trained transformer encoders with several predictors (Qi et al., 2020; Tian et al., 2024b,a; Yan et al., 2020; Siddique and Sufyan Beg, 2023). Mane et al. (2022) fine-tuned T5 on the RD task with a text-to-text approach and therefore in an open vocabulary fashion. Interestingly, Xu et al. (2024) and Aljaafari et al. (2024) exploit the task to probe the concept inference ability of generative LLMs.

Predictive lexicology Most works addressing neologisms come from the field of Computational Creativity and focus on the development of rule-based systems that leverage algorithms and various linguistic resources to generate new words, especially blends (Özbal and Strapparava, 2012; Stock and Strapparava, 2005; Smith et al., 2014; Mizrahi et al., 2020). More recently, word embeddings and character-based encoder-decoder LSTM architectures were used to generate blends given a pair of words (Simon, 2018; Das and Ghosh, 2017; Gangal et al., 2017). These systems can be used to generate neologisms, but they lack general linguistic knowledge, understanding of morphological restrictions, and they are confined to blends. Closest to our approach, Lencione et al. (2022) presented neologism generation as an extreme summarization task by training T5 to summarize definitions into single words. The model, when prompted with nonce glosses, is able to propose nonce words due to the

open vocabulary nature of this approach.

2.2 Definition Modeling (DM)

Originally proposed by Noraset et al. (2017) as a technique to explain distributed representations, Definition Modeling (DM) is the task of generating a gloss that describe the meaning of a word. As Mickus et al. (2019) suggested, due to the Distributional Hypothesis (Harris, 1954), the correct definition of a word can only be given when the linguistic context in which it occurs is known. Therefore, researchers started using contextual representations in order to account for polysemy (Ni and Wang, 2017; Ishiwatari et al., 2019). Similar to our approach, Generationary (Bevilacqua et al., 2020) used an encoder-decoder transformer by adding usage examples of the *definiendum* to the input text.

Explanatory lexicology Pinter et al. (2020), interestingly, studied blends as a class of OOV words and found that to recover the components of a blend is an extremely difficult task for BERT models. Malkin et al. (2021) discovered that GPT-3 (Brown et al., 2020) definitions of nonce words are sometimes preferred to those invented by humans . Contemporary to our work, NEO-BENCH (Zheng et al., 2024) evaluates state-of-the-art LLMs on several linguistic tasks involving neologisms (definition generation, translation, etc.) and discovered that performance is nearly halved, with bigger models achieving better results.

2.3 Exemplification Modeling (EM)

Proposed by Barba et al. (2021), Exemplification Modeling (EM) is the task of generating usage examples given a lemma paired with its gloss. Systems capable of performing this task have different practical applications being effective not only in WSD but also in the artificial generation of dictionary examples (He and Yiu, 2022).

3 Our Approach

We investigate the generation, definition, and usage of words, neologisms, and nonce words in the tasks of Reverse Dictionary (RD), Definition Modeling (DM) and Exemplification Modeling (EM). While attempting to solve these lexical tasks, the LMs are underneath exposed to a wide variety of phenomena that, for what concerns *motivated words*[2] (Grossmann and Rainer, 2013), exposes the morpholexical link that relates glosses to lemmas (i.e. *compounding*, *derivation*, *blending*, *acronyms* etc.). Acting according to these phenomena requires strong character knowledge (Liu et al., 2023, *Spelling Miracle*), morphemes identification, phonological understanding, etc.; rules that, when learned, ease the understanding of motivated words instead of solely relying on memorization.

Since a closed vocabulary approach would not work, as it is limited to a finite set of existing words $V = \{v_1, \ldots, v_{|V|}\}$, our approach leverages generative models pre-trained on a fixed set of subword tokens $S = \{s_1, \ldots, s_{|S|}\}$ (Kudo and Richardson, 2018), that can be combined to form the infinite set of all possible words $V \cup N$, where N indicates the infinite set of possible words[3].

For the purpose of our work, we used T5 Italian models (IT5) of different sizes, as well as the multilingual counterpart, to assess the impact of both model size and language capabilities. This choice is motivated by the fact that the IT5 models family is the only one for the Italian language available at different sizes and pre-trained on the same corpus.

While each model is **fine-tuned only on a set of words that are conventionally lexicalized** (i.e. present in a dictionary) and a small set of neologisms, **the evaluation is conducted across three linguistically different settings**: i) the **dictionary setting**, which consists in evaluating against an unseen split of the dictionary data as traditionally done in the RD and DM literature, ii) the **neologisms setting**, which involves evaluating against unseen neologisms that have zero to few occurrences in the models' pretraining data, iii) the **nonce words setting**, which consists of assessing the linguistically creative abilities in creating, defining, and using nonce words.

Lexical Proficiency Tasks Each task's *input* $\longrightarrow$ *output* pair is constructed starting from the following components: word, PoS, definition and, if available, label tags (e.g. *##medicine##*, *##chemistry##*, *figurative sense*, etc.), an etymology and a usage example:

> **RD**: "(PoS) + labels + definition + eventual etymology" $\longrightarrow$ "word", where the etymology is appended with a 1/5 chance if available. E.g. *"(adj) (figurative sense) that can be touched by hand,*

[2]Words where their lexical form and meaning results from a morphlogical combinations of morphemes (*photo* + *phobia*).

[3]We assume that N is infinite by extending the Chomskyan notion of linguistic creativity and recursion (Chomsky, 2002) beyond syntax to word formation rules: given a finite set of morphemes and phonemes, it is theoretically possible to construct an infinite set of word forms (e.g., *ex-ex-ex-wife*, *meta-meta-meta-language*, etc.).

and therefore evident to reason [ETYMOLOGY: from the late Latin tangibĭlis which derives from tangĕre meaning 'to touch']" ⟶ "tangible".

DM: "labels + word + (PoS) + eventual usage example" ⟶ "definition". E.g. *"##law## gallows (noun) [EXAMPLE: to send someone to the gallows or to be led to the gallows]" ⟶ "place where the death sentence is carried out".*

EM: "word + (PoS) + labels + definition" ⟶ "usage example". E.g. *"curiosity (noun) [DEFINITION: deep interest in learning new things]" ⟶ "curiosity is a strong stimulus to knowledge".*

Both the etymology and the labels are kept in order to further guide the model, during inference, towards a specific lexical and semantic field[4].

4 Datasets

In order to perform our experiments, we developed three new datasets, one for each lexical setting: the dictionary setting, the neologisms setting and the nonce words setting (see Appendix H for examples of lemmas in each dataset).

Dictionary dataset A set of words paired with definitions and usage examples essentially constitutes a dictionary. To the best of our knowledge, the only effort to develop an open and machine-readable dictionary for the Italian language is GLAW-IT (Calderone et al., 2016): built from the 2015 *Wikizionario*[5] dump, the resource is outdated. Therefore, in order to build our training data, we developed a new linguistic resource by parsing the April 2024 Wikizionario dump. The resource counts 370 786 Italian lexical entries, including forms, each paired with several linguistic metadata data when available (see coverage in Appendix A, Table 7). The extraction of structured data from the Wikizionario is a challenging task (Declerck et al., 2012) due to an unstructured markup language oriented to visualization and a lack of annotation constraints. Therefore the resource may naturally contain errors caused by both data and parsing inconsistencies. The resource covers 95.77% (with at least one definition) of the basic Italian vocabulary (De Mauro and Chiari, 2016) and contains 38 550 specialized terms, denoted by a semantic field label, offering a wide coverage of both common and technical lexicon.

To include some established neologisms in the training of our models, we also expanded our dataset with the ONLI neologisms database (*Osservatorio Neologico della Lingua Italiana*[6]), which contains 2 986 neologisms, up to 2019, paired with definitions, etymology, PoS and usage examples.

	Task	Train	Validation	Test
Dict.	RD	80,321	4,486	4,485
	DM	79,361	4,462	4,462
	EM	17,510	973	970
	Total	177,192	9,921	10,517

Table 1: Number of training, test and validation samples for each task for the dictionary setting.

Neologism dataset For the neologisms dataset we collected a list of neologisms from various online dictionaries (lexicalized after 2020 since the IT5-mT5 pre-training corpus reaches the end of 2020) that are not part of the Wikizionario or ONLI database. After selecting words with less then five occurrences in a 10% split of the IT5 pre-training data, we kept 100 neologisms and manually annotated them with definitions, PoS and usage examples. It should be noted that words in this setting mostly come from the news domain focusing on politics, COVID-19 social dynamics and contain several foreignerisms.

Nonce words dataset For the nonce words dataset, we used GPT-4o to obtain a list of 100 unattested nonce words paired with definitions, PoS and usage examples. This strategy builds on the findings that GPT-3's definitions of nonce words are sometimes preferred to those invented by humans (Malkin et al., 2021) (see Appendix E). The definitions are mostly about innovative and concrete objects, focusing on combining technology with several scientific disciplines or artefacts.

5 Experimental setting

We fine-tuned each model in a text-to-text multitask format (Raffel et al., 2019) only on the dictionary dataset by combining the Wikizionario with the ONLI neologisms database[7]. The obtained dataset was split into 90% training, 5% validation, and 5% test (see Table 1), stratifying by PoS tags. The models were tuned for a maximum of 15 epochs with early stopping based on validation loss (see Appendix B for training details).

For the RD task, our models don't rank words by default but rather rely on a decoder that autore-

[4] See Appendix I for examples.

[5] The Italian edition of Wiktionary.

[6] https://www.iliesi.cnr.it/ONLI/

[7] Proceprocessing steps applied to the obtained dataset are detailed in Appendix A.

Model	Lang	#P	#T	#T/#P
IT5-small	IT	60M	41B	683.33
IT5-base	IT	220M	41B	186.36
MT5-base	Multi	580M	6.3T	10,862.06
IT5-large	IT	738M	41B	55.55

Table 2: Models used in experiments along with the pre-training languages (*Lang*), number of parameters (*#P*), training tokens (*#T*) and tokens per parameter (*#T/#P*).

gressively generates each subtoken. Therefore, for each definition we produce 100 words (beam = 100) using the *diverse beam search* decoding strategy (Vijayakumar et al., 2016) with a diverse penalty of 0.8 (to ensure diversity) and rank them based on their probability. It's important to note that using beam search with such a high number of beams for creative lexical tasks as nonce word generation, may result in a very conservative, and possibly degenerate (Holtzman et al., 2020), approximation of the linguistically creative abilities of these models. That is, while a deterministic decoding strategy allows for a fair comparison between models, a sampling-based one would lead to the generation of more interesting candidates.

5.1 Models

The experiments were conducted on the Italian T5 models family (Sarti and Nissim, 2024): encoder-decoder Transformers pre-trained with masked span prediction on a cleaned version of the mC4's Italian split. Specifically, we finetuned and evaluated three different sizes of parameters: small (60M), base (220M) and large (738M). Furthermore, we included the MT5-base (580M) model (Xue et al., 2021) to assess the impact of multilingual pre-training on the tasks. Details about the models are reported in Table 2.

5.2 Evaluation strategies

We tested several metrics to assess the models' abilities across all tasks. Given the differences between them, some metrics are shared while others are task-specific. ROUGE-N (Lin, 2004) is used across RD and DM to measure the non-continuous overlap of n-grams (computed at a subtoken level[8]). Performances on the RD task are also evaluated using CER (Morris et al., 2004) (normalized Levenshtein distance) and, in compliance with previous work, using Acc@1/10/100[9]. For generative tasks, metrics of lexical overlap show a weak correlation with human judgment (Liu et al., 2016) and should be paired with a semantic-oriented score. Therefore, RD and DM are evaluated using SBERT (Reimers and Gurevych, 2019) representations: assuming that a word carries a similar representation to its definition, we discount the cosine similarity sim between target embedding **t** and prediction embedding **p** by a factor **Z** that measures how similar **t** is to the source input **s** with respect to **p**, penalizing predictions that are less similar to the source text even though they could be broadly semantically similar to the target. The measure m is:

$$Z = 1 - \max\left(0, sim(\mathbf{s}, \mathbf{t}) - sim(\mathbf{s}, \mathbf{p})\right)$$

$$m = Z \cdot sim(\mathbf{t}, \mathbf{p})$$

The EM task is evaluated using the median perplexity of the prediction versus the target usage examples, computed with Minerva-1B (Orlando et al., 2024)[10]. Since both DM and EM are tackled with a sampling decoding strategy and due to the small size of the evaluation datasets, the metrics reported in the neologisms and nonce words settings are the average across 5 runs.

Unlike the previous lexical settings, the evaluation of the nonce words setting is conducted in two ways: 1) for the DM and EM tasks we used the definitions and examples from the nonce dataset as targets, collected using GPT-4o; 2) **the RD task for nonce words is evaluated by human judges due to the creative aspects of this scenario.** Using the Prolific platform[11], we collected human judgments over 100 pairs of definitions (taken from the nonce words dataset) and nonce words (generated by our models). Building on the Optimal Innovation Hypothesis, 5 Italian native speakers are asked to read each definition-word pair and express two judgments about the nonce word on a 5-point Likert scale: the perceived **novelty** of the word w.r.t. their lexicon and a score of **adhesion** to the definition[12]. The intuition is that words that lie in the top-middle of the novelty scale but excel in adhesion should represent a lexical creative artefact that also entails the definition. To ensure that the candidate

[8] We used the trained IT5 SentencePiece unigram tokenizer (Kudo and Richardson, 2018) to split the text into tokens before computing Rouge-N.

[9] Given a set of generated words sorted by probability, *Acc@1/10/100* consists in measuring the model accuracy against the first/first ten/first hundred words.

[10] https://huggingface.co/sapienzanlp/Minerva-1B-base-v1.0.

[11] https://www.prolific.com/.

[12] Details on annotation in Appendix C.

		Reverse Dictionary					Definition Modeling				Exemplification Modeling	
		Acc@1/10/100	R-1	R-2	CER↓	SBERT	R-1	R-2	R-L	SBERT	PPL pred. ↓	PPL target
Dict.	**IT5-small**	*.29/.4/.53*	*41.33*	*31.19*	*50.58*	*0.68*	36.85	*23.98*	34.87	*0.61*	144.49	80.26
	IT5-base	.37/.52/.66	48	37.01	46	0.71	**39.58**	**26.54**	**37.42**	**0.65**	118.26	
	MT5-base	.33/.46/.57	43.64	33.73	47.95	0.7	*36.43*	24.58	*34.71*	0.62	*161.8*	
	IT5-large	**.39/.56/.69**	**49.7**	**38.8**	**43.83**	**0.73**	38.97	25.94	36.94	**0.65**	**112.66**	
	Avg	.34/.48/.61	45.67	35.18	47.09	0.7	37.96	25.26	35.98	0.63	134.3	
Neo.	**IT5-small**	*.06/.12/.13*	*25.39*	16.37	*71.95*	*0.55*	*18.36*	*3.44*	14.8	*0.45*	60.6	53.38
	IT5-base	.09/**.16**/.21	**33.06**	19.99	61.47	**0.6**	**21.21**	**5.36**	**16.92**	**0.53**	53.6	
	MT5-base	.08/.15/.18	26.82	*14.23*	**59.98**	0.59	18.43	3.66	*14.4*	0.48	*79.52*	
	IT5-large	**.1/.16/.27**	32.42	**20.64**	63.2	**0.6**	20.69	4.34	16.36	**0.53**	**43.44**	
	Avg	.08/.14/.19	29.4	17.8	64.05	0.58	19.67	4.2	15.62	0.5	59.15	
Nonce	**IT5-small**	—	—	—	—	—	18.91	*2.83*	15.13	*0.49*	68.35	64.28
	IT5-base	—	—	—	—	—	**21.79**	**4.19**	**17.13**	0.56	67.31	
	MT5-base	—	—	—	—	—	*18.1*	2.93	*14.15*	0.51	*84.33*	
	IT5-large	—	—	—	—	—	21.09	3.78	16.6	**0.58**	**48.05**	
	Avg	—	—	—	—	—	19.97	3.42	15.72	0.53	67.01	

Table 3: Results obtained by all the models for all the tasks (RD, DM and EM) and the three linguistically different settings: *Dict.*, *Neo.* and *Nonce*. R-1, R-2, and R-L stands for Rouge-1, Rouge-2 and Rouge-L (longest common subsequence).

	Noun	Adj.	Verb	Acron.
IT5-small	*.25/.5*	*.39/.57*	*.21/.5*	*.26/.30*
IT5-base	.37/**.56**	.51/.61	.35/**.54**	.68/.38
MT5-base	.29/*.5*	.47/*.55*	.29/*.48*	.52/.38
IT5-large	**.43/.56**	**.57/.62**	**.42/.54**	**.76/.44**
Avg.	.33/.53	.48/.58	.31/.51	.42/.38

Table 4: Reverse Dictionary Acc@10/Definition Modeling SBERT for full words and acronyms.

words are effectively new, we kept words that are not present in a lexicon of the Italian language (see Appendix G for details).

6 Results

First, we present the results obtained for the three different linguistic settings (Table 3), assessing the change in performance between model sizes and tasks. Subsequently, we analyze the results obtained from the human annotations of the generated nonce words[13].

Dictionary setting For the RD task, the metrics show a significant shift between IT5-small and other models, while IT5-large ranks first across all scores, closely followed by the monolingual base model. The multilingual counterpart, MT5-base, stays behind IT5-base, despite its bigger parameter size. ROUGE-1 and ROUGE-2 indicate that almost 50% of the time, the large model correctly predicts a single subtoken of the word, dropping to 38.8% when it comes to predicting at least two non-overlapping subtokens of the target word. Along with CER, this indicates that the models are morphologically flexible, to some extent, while still maintaining an overlap with the target word. The Acc@1/10/100 and the SBERT scores follow the same rank and add a semantic component to our evaluation. DM follows a similar trend, with ROUGE-N being slightly higher for IT5-base. Our SBERT metric follows the aforementioned rank, being equal between the Italian base and large models. In the EM task, the perplexity of the generated examples follows the tokens-per-parameters rank with the multilingual model ranking lowest. Table 4 shows that, among full PoS, adjectives are the easiest to retrieve and define, while verbs are the hardest: this is in line with the hypothesis that verbs, functioning as relational terms rather than discrete referents (Gentner, 2006), require more contextual anchoring. Nouns also show lower performance, which may be partly explained by the high proportion of specialized terms in the test set – 64% of nouns are marked with a domain-specific category label. Since the Wiktionary also contains acronyms, we were able to monitor acronym generation, which might leverage lexical memorization (i.e. memorize that *LLM* stands for *Large Language Model*), but it would also require character competence for unknown and low-frequency cases [14]. Interestingly, for acronyms, metrics increase drastically in the MT5-base, IT5-base and IT5-large. The plots in Figure 2 show the impact of several linguistic and non-linguistic factors in RD Acc@10 and DM SBERT (any point in the plots consist of at least 100 data instances). For the RD task, performance tends to decrease with increasing frequency rank[15] following a non-linear trend in which accuracy starts low, peaks for mid-frequency words, and then declines (ρ[16] = -0.09, p

[13] Examples of the generations obtained by the models are reported in Appendix I.

[14] see Appendix F for fictional acronyms generations.

[15] Words frequencies were obtained from a lemmatized version of the Italian Wikipedia.

[16] All the reported correlation are Spearman's rank correlation coefficients.

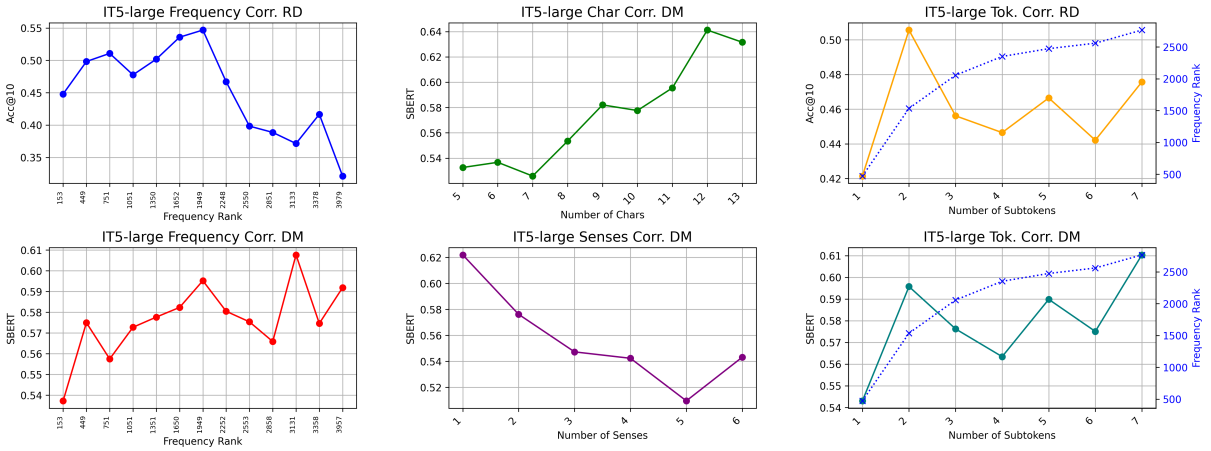

Figure 2: Variations of IT5-large performance according to frequency rank, number of characters, senses and subtokens. Scores are computed only against uninflected content words (nouns, adjectives and verbs).

< 0.001). This suggests that mid-frequency words, being less ambiguous yet still frequent, are easier to retrieve. Conversely, the performance trend for DM follows positively the frequency rank (although with almost no correlation, $\rho = 0.06$, $p < 0.001$). Consistent with Zipfian principles of word distribution, we observe a positive but weak correlation with the target lemma's number of characters, $\rho = 0.16$, $p < 0.001$, and a negative one with the number of senses, $\rho = -0.18$, $p < 0.001$. For RD, accuracy peaks when the target word is composed of exactly two subtokens, then gradually declines, yet remaining higher than for single-token words. A possible explanation is that single-token words tend to be high frequency[17] and have no compositional structure, token-wise, requiring pure lexical grounding. On the contrary, two token words are the ones located at the peak of the frequency rank-accuracy plot and may offer more compositional clues. This suggests that certain degrees of token compositionality could facilitate RD predictions. For DM, words with a higher number of subtokens reach higher levels of SBERT, again with a strong peak for words composed of exactly two subtokens, supporting the hypothesis that models may leverage token compositionality to decode meaning from the definiendum instead of relying solely on lexical memorization. Overall, the presented influencing factors show that the RD and DM tasks are, for some dynamics, inversely correlated: frequency has a positive impact on DM and a negative one on RD. Given the strong collinearity between frequency and the number of subtokens introduced by word segmentation algorithms, future work could disentangle this features dependence by leveraging partial correlation analysis to understand whether and how subtoken-level structure contributes to lexical generalization, especially in tasks that stress creativity and morphological productivity.

Neologisms setting The values reported for the neologisms setting on the RD task show a performance rank that is mostly consistent with the one reported for the dictionary setting. IT5-large and IT5-base tie in the reverse dictionary task while strongly outperforming other models according to ROUGE-N. The CER score, on the other hand, puts the MT5-base model as a top performer: several lemmas in the neologism dataset are foreignerisms (e.g. *booktoker*, *algospeak*), thus facilitating this task for the multilingual model. The reported metrics for the EM and DM tasks reflect the already observed rank in the dictionary setting. Overall, the results show that model **performance significantly drops when dealing with extremely low-frequency neologisms** according to both lexical and semantic metrics.

Nonce words setting Metrics for the nonce words automatic evaluation (DM and EM tasks) maintain the same rank between models found in the dictionary setting. Overall, the results are slightly better w.r.t. the neologisms evaluation with the SBERT score increasing across all models. This dynamic lies in the fact that words in this scenario are less related to the journalistic lexicon and the semantic field is less specific. Therefore, definitions in the nonce words dataset tend to be less related to real facts, which may be unknown to the mod-

[17] Due to BPE-like tokenization algorithms (Sennrich et al., 2016).

	Adhesion	Novelty	α
IT5-small	3.06±1.45	3.11±1.3	.51/.14
IT5-base	3.01±1.32	3.61±1.37	.29/.34
MT5-base	3.37±1.32	2.98±1.31	.37/.15
IT5-large	3.37±1.42	3.11±1.15	.41/.18
GPT-4o	3.86±1.09	3.32±1.15	.17/.07

Table 5: Mean and standard deviation for the adhesion and novelty scores given by human annotators. The column α reports Krippendorff's Alpha between annotators for adhesion/novelty.

els. The improvements over SBERT, especially for the IT5-large model (+5%), w.r.t. the neologisms setting, support the hypothesis.

6.1 Human evaluation: nonce generation

The average scores from the human annotation (see Table 5 and Figure 3) show several interesting results. We also reported scores for GPT-4o to provide a theoretical upper bound. The annotators' agreement is consistently higher for adhesion rather than novelty indicating that the latter is a more subjective judgment than the former. Nonetheless, annotators have an overall reasonable agreement considering the inherent personal variables that play a role in novelty perception and adhesion assessment of unseen words. The monolingual small and base models rank the lowest on the adhesion scale with their respective distributions of judgments being statistically different from all other models (see Appendix D) but similar to each other. Regarding novelty, IT5-base is the top performer but the worst in terms of adhesion. Along with IT5-small, they both have a higher degree of novelty than adhesion thus reversing the Optimal Innovation criterion which would require strong adhesion and medium novelty. In this regard, **GPT-4o scores represent candidates that better approximate our theoretical assumption**, with the majority of words being highly related to the definitions and with an average degree of novelty. On the other hand, the multilingual model ties with the IT5-large in terms of adhesion but has the lowest novelty score, while the **IT5-large is the only model**, among ours, **that can maintain a good degree of adhesion and a medium novelty**. Looking at the distributions in Figure 3, the trend emerges quite well: IT5-small produces a lot of incoherent and less innovative words, IT5-base received high values of novelty but failed to maximize adhesion and IT5-large produced judgments that, between our models, better approximate our theoretical assumption, with words mostly being perceived as quite novel and coherent with the definitions (see Table 6).

The variations in the agreement can be explained by the degree of obviousness and ambiguity of the generated words: for example, candidates that are clearly questionable in terms of novelty or adhesion would lead to a higher agreement, while more plausible words could enhance subjective factors. IT5-small candidates have a high adhesion agreement and scores are relatively low suggesting that the nonce words are evidently less related to the definitions. Interestingly, the IT5-base model has the highest score and agreement for novelty but the lowest one in terms of adhesion, thus indicating that the model produces highly innovative words which may be naturally less familiar and difficult to interpret in terms of adhesion. On the other hand, the multilingual and IT5-large models have the highest agreement in terms of adhesion and the lowest scores in terms of novelty, suggesting that medium-novelty words are easier to interpret in terms of adhesion to definitions. GPT-4o's low agreement, following this hypothesis, could mean that the generated words are linguistically more niche, refined, novel and less transparent thus fueling judgment ambiguity.

6.2 Discussion

The model's ranking is mostly consistent across all metrics showing that they tend to correctly capture a general trend in performance while offering a view on different linguistic axes. **Larger, monolingual models generally outperformed their multilingual counterparts** with MT5-base only gaining an edge when handling neologisms, particularly those containing foreignerism. These results align with previous findings (Sarti and Nissim, 2024), where IT5-models outperformed the multilingual ones on several downstream tasks, closely following our rank, suggesting that **proficiency on lexical tasks can be indicative of broader downstream performance** (Xu et al., 2024). Despite the notable drop in performance with low-frequency neologisms and nonce words (as observed in Zheng et al. 2024), the rank between models remained consistent. The same rank can be observed for the acronym accuracy in the RD task, where the sharp increase of larger models suggests that **larger models can store and access more sublexical information**, such as spelling patterns, even in character-blind models (Liu et al., 2023, *Spelling Miracle*).

Our models' ability to generate novel and co-

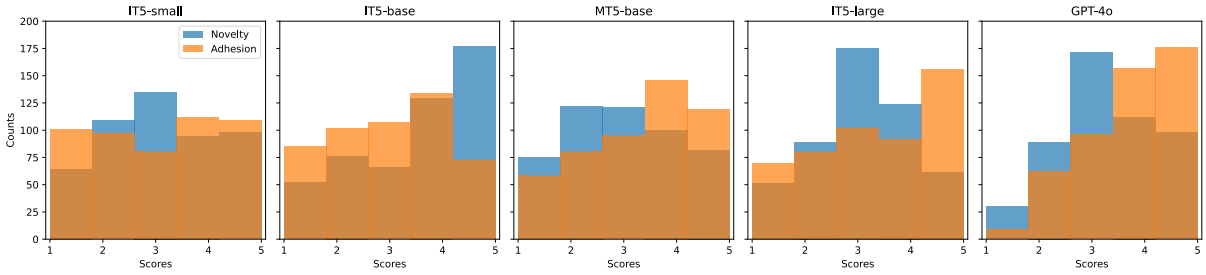

Figure 3: Distribution of novelty and adhesion human scores across the 5 values of the Likert scale for all models.

Definitions	Model	Predicted Word	Adhesion	Novelty
Veicolo progettato per esplorazioni su superfici planetarie, adatto a terreni extraterrestri. [transl. *Vehicle designed for exploration on planetary surfaces, suitable for extraterrestrial terrain.*]	IT5-small	planetaro	3.0	4.2
	IT5-base	elioplano [*helioplane*]	2.2	4.6
	MT5-base	cosmoplano [*cosmoplane*]	3.2	4.0
	IT5-large	astroveicolo [*astrovehicle*]	4.6	3.2
	GPT-4o	roverastro [*astrorover*]	3.6	3.4
Vela navigabile che raccoglie dati geologici mentre si sposta su laghi o mari, utilizzata in esplorazioni scientifiche. [*Navigable sail that collects geological data as it moves across lakes or seas, used in scientific exploration.*]	IT5-small	geonauta [*geonaut*]	4.6	2.4
	IT5-base	ecovela [*ecosail*]	4.4	1.8
	MT5-base	vettolaghiera	2.0	4.4
	IT5-large	idrovedetta [*hydropatrol*]	4.6	2.8
	GPT-4o	geonave [*geoship*]	4.0	3.2
Una tavola o superficie capace di mostrare visivamente il passare del tempo, evidenziando i cambiamenti avvenuti su di essa. [*A table or surface capable of visually showing the passage of time, highlighting the changes that have occurred on it.*]	IT5-small	cromatopompa	1.2	3.8
	IT5-base	cronopalestra [*chronogym*]	2.0	5.0
	MT5-base	retrotavola [*retrotable*]	2.2	3.0
	IT5-large	cronotavolo [*chronotable*]	4.4	3.0
	GPT-4o	cronotavola [*chronotable*]	3.6	3.6
Forma d'arte che utilizza nebbie artificiali e giochi di luce per creare installazioni immersive. [*An art form that uses artificial fog and light effects to create immersive installations.*]	IT5-small	immersivismo [*immersivism*]	3.8	2.4
	IT5-base	metacaduta [*metafall*]	2.0	4.6
	MT5-base	fotoart [*photoart*]	3.4	2.6
	IT5-large	nebbiografia [*foggraphy*]	4.4	3.0
	GPT-4o	nebbioarte [*fogart*]	3.6	3.6
Fenomeno in cui i movimenti delle placche terrestri generano onde sismiche che producono suoni dissonanti, studiato in geologia e acustica. [*Phenomenon in which the movements of the earth's plates generate seismic waves that produce dissonant sounds, studied in geology and acoustics.*]	IT5-small	biogeoacustica [*biogeoacoustics*]	4.4	3.4
	IT5-base	sismofonia [*seismophony*]	3.0	4.0
	MT5-base	sismismo [*seismism*]	3.0	4.0
	IT5-large	sismofonia [*seismophony*]	4.2	3.2
	GPT-4o	sismofonia [*seismophony*]	4.2	2.0

Table 6: Sample of generated nonce words (we tried to provide a translation when possible), along with adhesion and novelty average scores, for all the models. The definitions are those generated by GPT-4o.

herent nonce words further indicates that **LMs are capable of learning approximations of word formation rules, rather than relying solely on memorization, thus showing signs of generalization**. The human evaluation results, inspired by the Optimal Innovation Hypothesis, confirmed that larger models like IT5-large were more adept at producing creative and semantically consistent nonce words. Additionally, GPT-4o demonstrated potential for generating candidates aligned with the Optimal Innovation criteria.

7 Conclusion

In this paper, we presented a novel evaluation framework for assessing the lexical proficiency of Language Models. Additionally, we developed a tailored resource of Italian lexical entries suitable for a variety of lexical proficiency tasks. To the best of our knowledge, this is the first work to extend the Reverse Dictionary, Definition Modeling, and Exemplification Modeling tasks across commonly lexicalized words, recent neologisms and nonce words, with an emphasis on the creative aspects of this last setting. Our multifaceted evaluation framework showed that lexical proficiency tasks remain a significant challenge for LMs, particularly when stress-tested with neologisms and nonce words. On the other hand, the capability of bigger mono-lingual models to occasionally produce, use and define meaningful and unattested nonce words suggests that such models exhibit an understanding of the compositional nature of word formation rules, which is fundamental to master language understanding and to handle lexical innovation. Furthermore, the results obtained were consistent across models and settings, aligning with the results on downstream tasks found by Sarti and Nissim (2024), suggesting that lexical proficiency tasks are correlated to downstream performance (Xu et al., 2024). These findings, along with the insights on linguistic creativity, show the robustness and effectiveness of our framework in evaluating several aspects of LMs linguistic understanding. The proposed framework could provide a useful method that can be expanded to different languages and models. Moreover, our experiments also provide a first effort towards building RD, DM and EM tools for the Italian language.

Limitations

In this section, we discuss the limitations of our work. 1) **Tested models**: For our experiments, we relied on models based on the T5 family. We specifically chose these models to have a fair representation of model sizes, as well as the distinction between mono- and multi-lingual models, for the Italian language. Nevertheless, it would be beneficial to evaluate models with different architectures (e.g. encoder-decoder vs decoder-only), bigger in terms of parameters and including zero or few-shot evaluation of instruction-tuned models, in order to understand their abilities in the tasks we devised. 2) **Multilinguality**: Our results are limited to the Italian language, which constrains the generalizability of our findings to other languages. Since our framework is based on tasks that require access to sources with lexical entries, it would be relatively straightforward to port this approach to other languages. Therefore, in future work, it would be beneficial to test its generalizability and to explore potential cross-linguistic differences in LMs performance. 3) **Human evaluation**: Our selection of human raters was based on Italian as their native language with at least a BA/BSc degree. Future work could investigate whether the results would differ across different occupations and education levels. Moreover, extending the number of annotators per instance could provide stronger insights into their level of agreement and therefore about the generalizability of the results. 4) **Linguistic analysis**: Since our focus was to introduce a novel evaluation framework and to assess whether LMs can produce, explain and use lexical innovations, we did not conduct an in depth analysis on the specific phenomena related to the understanding of LMs' morphological productivity. On the other hand, our framework opens to future works that could investigate more particular linguistic aspects related to the generation and explanation of neologisms and nonce words.

Acknowledgments

The authors acknowledge the support of the project XAI-CARE-PNRR-MAD-2022-12376692 under the NRRP MUR program funded by the NextGenerationEU and the PNRR MUR project PE0000013-FAIR. Partial support was also received by the project "Advancing Italian Language Processing with Small-Scale Training and Preference Modeling" (IsCb8_AILP), funded by CINECA under the ISCRA initiative, for the availability of HPC resources and support.

References

Nura Aljaafari, Danilo S Carvalho, and André Freitas. 2024. The mechanics of conceptual interpretation in gpt models: Interpretative insights. *arXiv preprint arXiv:2408.11827*.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification modeling: Can you give me an example, please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3779–3785. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

D Frank Benson. 1979. Neurologic correlates of anomia. In *Studies in neurolinguistics*, pages 293–328. Elsevier.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "how we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Margaret A Boden. 2004. The creative mind: Myths and mechanisms.

Roger Brown and David McNeill. 1966. The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, 5(4):325–337.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Basilio Calderone, Franck Sajous, and Nabil Hathout. 2016. Glaw-it: A free large italian dictionary encoded in a fine-grained xml format. In *49th Annual Meeting of the Societas Linguistica Europaea (SLE 2016)*, pages 43–45.

Noam Chomsky. 2002. *Syntactic structures*. Mouton de Gruyter.

Kollol Das and Shaona Ghosh. 2017. Neuramanteau: A neural network ensemble model for lexical blends. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 576–583, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tullio De Mauro and I Chiari. 2016. Il nuovo vocabolario di base della lingua italiana. *Internazionale.[28/11/2020]. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana*.

Thierry Declerck, Stefania Racioppa, and Karlheinz Mörth. 2012. Automatized merging of italian lexical resources. In *LREC 2012 Workshop on Language Resource Merging Workshop Programme*, page 45.

Gilles Fauconnier and Mark Turner. 2003. Conceptual blending, form and meaning. *Recherches en communication*, 19:57–86.

Varun Gangal, Harsh Jhamtani, Graham Neubig, Eduard Hovy, and Eric Nyberg. 2017. Charmanteau: Character embedding models for portmanteau creation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2917–2922, Copenhagen, Denmark. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Dedre Gentner. 2006. Why verbs are hard to learn. *Action meets word: How children learn verbs*, pages 544–564.

Rachel Giora, Ofer Fein, Ann Kronrod, Idit Elnatan, Noa Shuval, and Adi Zur. 2004. Weapons of mass distraction: Optimal innovation and pleasure ratings. *Metaphor and symbol*, 19(2):115–141.

Maria Grossmann and Franz Rainer. 2013. *La formazione delle parole in italiano*. Walter de Gruyter.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Xingwei He and Siu Ming Yiu. 2022. Controllable dictionary example generation: Generating example sentences for specific targeted audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Gabriel R Lencione, Rodrigo F Nogueira, and Paula Y Pasqualini. 2022. Nameling: Creative neologism generation with transfer learning.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, Rj Mical, Mohammad Norouzi, and Noah Constant. 2023. Character-aware models improve visual text

rendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16270–16297, Toronto, Canada. Association for Computational Linguistics.

Nikolay Malkin, Sameera Lanka, Pranav Goel, Sudha Rao, and Nebojsa Jojic. 2021. GPT perdetry test: Generating new meanings for new words. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5542–5553, Online. Association for Computational Linguistics.

Sunil B Mane, Harshal Navneet Patil, Kanhaiya Balaji Madaswar, and Pranav Nitin Sadavarte. 2022. Wordalchemy: a transformer-based reverse dictionary. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pages 1–5. IEEE.

Christopher D. Manning. 2015. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Timothee Mickus, Denis Paperno, and Matthieu Constant. 2019. Mark my word: A sequence-to-sequence approach to definition modeling. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 1–11, Turku, Finland. Linköping University Electronic Press.

Timothee Mickus, Kees Van Deemter, Mathieu Constant, and Denis Paperno. 2022. Semeval-2022 task 1: CODWOE – comparing dictionaries and word embeddings. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1–14, Seattle, United States. Association for Computational Linguistics.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Riccardo Orlando, 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 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719.

Gözde Özbal and Carlo Strapparava. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 703–711, Jeju Island, Korea. Association for Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. Will it unblend? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.

Fanchao Qi, Lei Zhang, Yanhui Yang, Zhiyuan Liu, and Maosong Sun. 2020. WantWords: An open-source online reverse dictionary system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–181, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gabriele Sarti and Malvina Nissim. 2024. IT5: Text-to-text pretraining for Italian language understanding and generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia. ELRA and ICCL.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Bushra Siddique and M. M. Sufyan Beg. 2023. Reverse dictionary formation: State of the art and future directions. *SN Comput. Sci.*, 4(2):168.

Bushra Siddique and MM Sufyan Beg. 2023. Reverse dictionary formation: State of the art and future directions. *SN Computer Science*, 4(2):168.

Gerardo Sierra. 2000. The onomasiological dictionary: a gap in lexicography. In *Proceedings of the ninth Euralex international congress*, pages 223–235.

Jonathan A Simon. 2018. Entendrepreneur: Generating humorous portmanteaus using wordembeddings. In *Second Workshop on Machine Learning for Creativity and Design (NeurIPS 2018)*.

Michael R. Smith, Ryan S. Hintze, and Dan Ventura. 2014. Nehovah: A Neologism Creator Nomen Ipsum. In *Proceedings of the Fifth International Conference on Computational Creativity*, pages 173–181, Ljubljana, Slovenia.

Oliviero Stock and Carlo Strapparava. 2005. HAHAcronym: A computational humor system. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 113–116, Ann Arbor, Michigan. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Sicheng Tian, Shaobin Huang, Rongsheng Li, and Chi Wei. 2024a. A prompt construction method for the reverse dictionary task of large-scale language models. *Engineering Applications of Artificial Intelligence*, 133:108596.

Sicheng Tian, Shaobin Huang, Rongsheng Li, Chi Wei, and Liu Ye. 2024b. Rdmtl: Reverse dictionary model based on multitask learning. *Knowledge-Based Systems*, page 111869.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tony Veale and Cristina Butnariu. 2006. Exploring linguistic creativity via predictive lexicology. In *At the ECAI'2006 Joint International Workshop on Computational Creativity. Italy*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: Benchmark the linguistic competence of language models. *arXiv preprint arXiv:2404.18923*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Menghan Zhang, Peng Qian, and Xuanjing Huang. 2024. On the tip of the tongue: Analyzing conceptual representation in large language models with reverse-dictionary probe. *arXiv preprint arXiv:2402.14404*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hang Yan, Xiaonan Li, Xipeng Qiu, and Bocao Deng. 2020. BERT for monolingual and cross-lingual reverse dictionary. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4329–4338, Online. Association for Computational Linguistics.

Lei Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):312–319.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. NEOBENCH: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

Michael Zock and Slaven Bilac. 2004. Word lookup on the basis of associations : from an idea to a roadmap. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 29–35, Geneva, Switzerland. COLING.

A Datasets statistics and preprocessing

Table 7 reports the total number and coverage distribution of linguistic metadata of the Wikizionario dataset.

In order to obtain the training set, the following processing steps are followed: a training instance is created for each sense, therefore if a lemma has 2 PoS and 2 senses for each PoS we create 4 instances of training; we skipped 97% of verb forms, 70% of noun forms and 70% of adjective forms, we kept all lemmas longer than 1 character that are paired with a gloss of at least 20 characters, etymology is appended to the definition when available with a 1/5 chance. Other small processing steps are employed to clean empty glosses, too long examples, too short etymologies, glosses without meaning (as proper nouns), and glosses with the definiendum in the definiens. The obtained dataset is split into 90% training, 5% validation, 5% test stratifying by PoS (see Table 4 for PoS distribution).

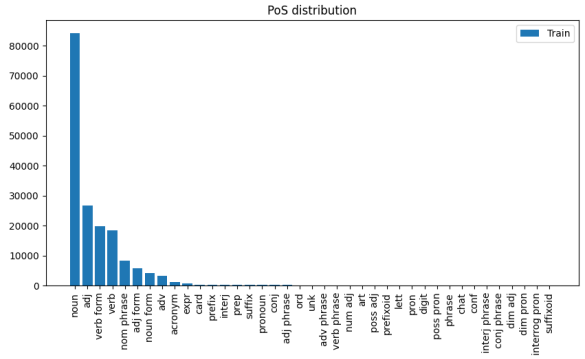

Figure 4: Distribution of PoS in the training set.

B Training details

The experiments were carried out using two NVIDIA GeForce RTX 4090 GPUs. All models share the following hyperparameters and tuning strategy: max length of the input and output text is 128 and 64 tokens, finetuning is performed for max 15 epochs with an early stopping (patience 3) based on the validation loss, the first 5% of the total steps have a linear warm up in the learning rate and, after that, the learning rate drops following a half cosine function until it reaches 5e-7; label smoothing (Szegedy et al., 2016) is used to introduce noise in the one-hot encoded target logits making the model less confident during generation; Adafactor (Shazeer and Stern, 2018) is used along with a weight decay of 1e-3. In order to perform the tuning with the batch sizes reported in table 8 we used gradient accumulation for the IT5-base, MT5-base and IT5-large models. All models maintained a decreasing validation loss across all the finetuning steps except for the IT5-base, which was stopped at the 13 epoch due to an increasing validation loss for the last three epochs.

Inference for the RD task uses diverse beam search with 100 beams, 100 beam groups and a diversity penalty of 0.8. Inference for the RD and EM tasks uses nucleus sampling with top_k = 50, top_p = 0.9 and a repetition penalty (Keskar et al., 2019) of 1.3 since outputs often contain word repetitions; 5 candidates are generated for each source text, then sorted by probability, the highest one is chosen. The sentence-transformer model used for computing similarity metrics, SBERT (Reimers and Gurevych, 2019), is "paraphrase-multilingual-mpnet-base-v2"[18] (278M) (the model is loaded through the Sentence Transformers library[19]). All experiments are conducted using the Hugging Face transformers library (Wolf et al., 2020).

C Human annotation

Human annotation was performed on the Prolific platform[20]. We recruited a total of 83 Italian native speakers with at least a BA/BSc degree and no language-related disorders. We performed the annotation for 100 definition-word pairs for each model (including GPT-4o), for a total of 500 samples. Each task was formulated as a questionnaire composed of a set of 27 definition-word pairs (25 + 2 control questions) and, for each of them, we collected the scores of 5 annotators. Each annotator was paid 1.30£ (7.80£ per hour). The annotators were asked to provide a score on a 5-value Likert scale according to the perceived **novelty** of the word and its **adhesion** to the definition. For instance, given the pair:

- **Definition**: Formaggio talmente buono che sembra venire dallo spazio [transl. *A cheese which tastes so good that it seems to have come from outer space*]
- **Word**: astrocacio [*astrocheese*]

The following questions were asked in the questionnaire:

[18] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.
[19] https://www.sbert.net/index.html.
[20] https://www.prolific.com/

PoS	total	def%	etym%	example%	syn%	ant%	ipa%	morpho%	syll%
verb form	289013	99.99	99.61	0.21	3.36	1.88	1.96	0.38	4.36
noun	34763	91.43	75.82	13.31	30.88	14.13	52.41	83.01	75.71
adj form	15332	99.32	61.38	2.30	15.45	12.95	7.02	94.73	57.72
noun form	15329	99.44	87.64	2.97	17.12	7.85	7.78	95.79	82.81
adj	13904	85.31	44.99	13.03	20.12	15.83	31.54	80.97	44.96
verb	6795	87.86	88.67	29.17	68.98	45.03	73.16	0.04	94.33
name	4673	98.29	24.59	1.58	2.10	0.11	8.20	50.61	16.46
adv	3116	96.05	84.40	6.74	35.17	21.92	41.91	0.22	93.42
noun phrase	3100	99.00	61.71	4.13	7.55	1.48	4.65	22.52	30.55
acron	567	96.30	34.57	3.70	2.12	0.00	6.00	7.58	3.00

Table 7: Linguistic metadata coverage over PoS with more than 500 words. Verb forms have no annotated morphology because it's usually fully written in the definitions. Forms in general are generated automatically by bots and are not fully annotated.

	lr	batch
IT5-small	1e-3	128
IT5-base	7e-4	128
MT5-base	1e-4	64
IT5-large	1e-4	32

Table 8: Hyperparameters that differs between models.

- **Novelty**: Rispetto al lessico che già conosci, quanto percepisci come nuova la parola che stai vedendo? La novità va intesa sia a livello della parola in sé sia a livello delle parti che la compongono. [*Compared to your vocabulary, how new do you perceive the word you are reading? Newness should be understood both at the level of the word itself and at the level of its component.*]
- **Adhesion**: Quanto è plausibile che questa parola sia descritta dalla definizione associata? [*How plausibly is this word described by the associated definition?*]

Before the annotation task, a set of 3 solved examples were presented to the annotators.

D Wilcoxon test

Figures 5 and 6 report the results obtained performing the Wilcoxon significance test between the distributions of adhesion and novelty human scores between each model pair.

E Nonce words dataset creation

In order to obtain a set 100 fictional dictionary entries of nonce words we prompted GPT-4o with the following text, where n is the number of requested entries and *PoS* is the part of speech:

Prompt "Inventa n parole italiane accompagnate da Part Of Speech (*PoS*), una definizione e un esempio d'uso. Fai in modo che (sia per questioni di motivazione morfologica sia per altre dinamiche linguistiche) le parole appaiano correlate con la definizione. Assicurati che le parole e i rispettivi sensi si riferiscano ad artefatti (concreti o astratti) inesistenti. Infine, le entrate lessicali fittizie de-

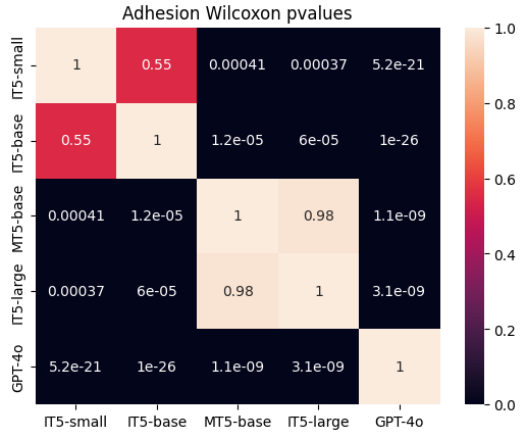

Figure 5: Wilcoxon significance test computed on the distribution of adhesion human scores between all models.

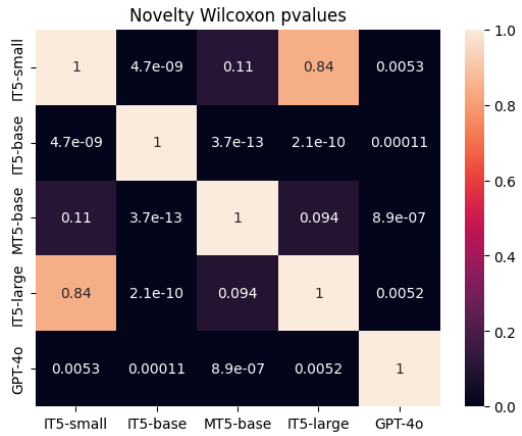

Figure 6: Wilcoxon significance test computed on the distribution of novelty human scores between all models.

vono sembrare reali, quindi ometti termini come "magico" o "immaginario" dalle definizioni. Varia il dominio". [transl. *"Invent n Italian words accompanied by Part Of Speech (PoS), a definition and an example of use. Make sure that (both for reasons of morphological motivation and for other linguistic dynamics) the words appear correlated with the definition. Make sure the words and their meanings refer to non-existent (concrete or abstract) artifacts. Finally, fictional lexical entries must sound real, so omit terms like "magical" or "imaginary" from the definitions. The domain varies."*].

We collected entries in batches of 10, trying to maintain a ratio of 70% nouns, 15% adjectives and 15% verbs. For each generated word we checked the existence of the generated nonce word in a huge lexicon of 4 592 345 words extracted form the WIkizionario, Italian Wikipedia (March 2019 dump) and the English Wiktionary (Sept. 2024 dump). Since GPT-4o often struggled to generate H-Creative linguistic artifacts, especially for adjectives and verbs, we had to perform several iterations to finally obtain a sample of 100 samples composed of 70 nouns, 16 verbs and 14 adjectives.

F Fictional acronym generation

In order to investigate whether character knowledge is actually leveraged by the models to perform RD on acronyms, we report in Table 9 ten examples of acronyms generated by the models starting from fictional definitions: such cases would require to understand the mechanism behind acronym generation and to rely on character knowledge (Liu et al., 2023, *Spelling Miracle*) since there is no prior association between the acronym and the definition. Only the large and, partially, the monolingual base model were able to effectively generate some correct acronyms, with IT5-large producing almost always the correct acronym. On the other hand, the smallest model seems to rely on existing acronyms failing to address novel cases. This supports the hypothesis that bigger and monolingual models generalized the task by effectively leveraging some kind of stored character information.

G Further details on the nonce words filtering process

In the RD task, given a set of 100 candidate words generated for each definition in the nonce words dataset, we introduced a filter that excludes candidates present in a lexicon extracted from the WIkizionario, Italian Wikipedia and the English Wiktionary. We employed this filter since all models, including GPT-4o, struggled to consistently generate a novel word as the most probable. On the other hand, each model was able to generate at least one unattested word in the 100 candidates set. Since words are ranked by probability we extracted the mean and standard deviation, reported in Table 10, of the first index where an unattested nonce word appeared.

All models were able to produce unattested words in the first ranks of the 100 candidates set, with bigger and monolingual models being the most "conservative" in terms of lexical innovation, while the smallest and multilingual ones are more prone to produce novel words.

H Datasets examples

To give a better understanding of each linguistic setting (i.e. dictionary, neologisms and nonce words) we report in Table 11 a small sample of lemmas taken from each dataset.

Definition	IT5-small	IT5-base	IT5-large	MT5-base
(acron) **e**nte **t**utela **l**ampade	eni	epli	**etl**	ttl
(acron) **a**ssociazione **x**ilofoni **i**talia	ais	api	**axi**	asip
(acron) **i**stituto di **c**ucina **c**omputazionale	iso	**icc**	**icc**	icpc
(acron) **c**entro **i**nternazionale di **p**ittura **a**nalitica	ici	**cipa**	**cipa**	ipa
(acron) **l**ezioni **i**taliane di **s**tile e **a**tteggiamento	ita	lisi	**lisa**	iat
(acron) **e**nte **n**azionale **m**atite spesse	eni	enis	**enms**	nms
(acron) **s**ocietà **a**nonima **c**amionisti **n**otturni	aci	sat	satn	sno
(acron) **c**ircolo **r**obot **i**ndipendenti	cdc	**cri**	**cri**	ccrp
(acron) **f**ictional **c**haracter **a**nalysis	sms	lca	**fca**	fbi
(acron) **s**ocietà **a**nonima **l**avanda **e**uropea	cep	soe	**sale**	snel

Table 9: Some examples of acronyms generated by the models from fictional definitions. Correct predicitons are in bold.

IT5-small	IT5-base	MT5-base	IT5-large
2.5±3.02	3.82±6.3	2.39±5.53	3.73±6.56

Table 10: Mean and standard deviation of the first index where an unattested nonce word appeared.

Setting	Lemmas
Dict.	emissario (*emissary*), sorpresa (*surprise*), demineralizzare (*demineralize*), webdesigner, malmostoso (*malmostous*)
Neos.	deplastificazione (*deplasticization*), dop economy, solarpunk, antropausa (*anthropause*), domicidio (*domicide*)
Nonce	idrotex (*hydrotex*), cronodinamica (*cronodynamics*), crepuscoloscopio (*duskscope*), fitocrazia (*phytocracy*), neurovisore (*neurovisor*)

Table 11: Some examples of lemmas taken from each linguistic setting.

I Models generations

The following tables report samples of the generated outputs for the tasks of RD, DM and EM.

RD	source	target	models	predictions
Dict.	(sost) ##diritto## giudizio espresso da un giudice in un processo che può essere di assoluzione o di condanna [transl. *(noun) ##law## judgment rendered by a judge in a trial, which may be of acquittal or conviction*]	sentenza [*sentence*]	IT5-small	'giudizio' [*judgment*], 'condanna' [*conviction*], 'esecrazione' [*execration*]
			IT5-base	'**sentenza**' [*sentence*], 'sentenza', 'verdetto' [*verdict*]
			MT5-base	'**sentenza**', 'provvedimento' [*provision*], 'giurisdizione' [*jurisdiction*]
			IT5-large	'**sentenza**', 'sentenza', 'pronunzia' [*pronouncement*]
	(verb) avere un'opinione diversa da quella di altre persone e manifestare ciò [*(verb) to have a different opinion from others and express it*]	obiettare [*to object*]	IT5-small	'adeguarsi' [*to conform*], 'opinionizzare' [*to opinionize*], 'biasimare' [*to blame*]
			IT5-base	'smentirsi' [*to contradict*], 'diffidare' [*to distrust*], 'discolpare' [*to exonerate*]
			MT5-base	'divergere' [*to diverge*], 'divergere', 'diffondersi' [*to spread*]
			IT5-large	'dissentire' [*to dissent*], 'opporsi' [*to oppose*], 'discostarsi' [*to differ*]
	(agg) che fa bene al proprio benessere [*(adj.) that promotes well-being*]	salutare [*healthy*]	IT5-small	'benevolo' [*benevolent*], 'sano' [*healthy*], 'benefico' [*beneficial*]
			IT5-base	'benefico' [*beneficial*], 'sano', 'salubre' [*healthful*]
			MT5-base	'benevolo' [*benevolent*], 'buono' [*good*], 'rispettoso' [*respectful*]
			IT5-large	'sano' [*healthy*], 'salubre' [*healthful*], '**salutare**'
	(agg) che è assurdo, inimmaginabile, inaccettabile, palesemente sciocco [*(adj) which is absurd, unimaginable, unacceptable, patently foolish*]	pazzesco [*crazy*]	IT5-small	'inimmaginabile' [*unimaginable*], 'assurdo' [*absurd*], 'indecifrabile' [*indecipherable*]
			IT5-base	'irrazionale' [*irrational*], 'inconcepibile' [*inconceivable*], 'ipocrita' [*hypocritical*]
			MT5-base	'inimmaginabile' [*unimaginable*], 'illusorio' [*illusory*], 'assurdo' [*absurd*]
			IT5-large	'**pazzesco**' [*crazy*], 'strampalato' [*bizarre*], 'improponibile' [*unthinkable*]
Neo.	(loc nom) ##neologismo## ##forestierismo## Settore economico focalizzato sui prodotti agricoli con Indicazione Geografica, contribuendo significativamente al valore del comparto agroalimentare nazionale. [*(noun phrase) ##neologism## ##foreignerism## Economic sector focused on agricultural products with Geographical Indication, significantly contributing to the value of the national agri-food sector.*]	dop economy	IT5-small	'agrifood', 'food-survivaid', 'food-relations'
			IT5-base	'food&wine', 'agroalimentare di qualità' [*high-quality agri-food*], 'horeca'
			MT5-base	'food economy', 'green economy', 'agrobusiness'
			IT5-large	'**dop economy**', 'wholesale agriculture', 'agroalimentare di qualità'
	(sost) ##neologismo## Insegnante specializzato per individui non vedenti o ipovedenti [etimologia: dal greco antico τυφλός e διδάσκω] [*(noun) ##neologism## Specialized teacher for blind or visually impaired individuals [etymology: from ancient Greek τυφλός and διδάσκω)]*	tiflodidatta [*tifloteacher*]	IT5-small	'tiflonaio', 'pedodidaco', 'timidida'
			IT5-base	'tiflodidasco', 'tuflodidattico', 'totodidattico'
			MT5-base	'tifoso' [*fan*], 'tettodidattico', 'trofoide'
			IT5-large	'tiflodidattico' [*tifloeducational*], 'tiflologo' [*tiflologist*], 'tuflologo'
	(sost) ##neologismo## la capacità di esprimersi e di interagire con un tono e in uno stile conversevole, vivido ed efficace [*(noun) ##neologism## the ability to express oneself and interact in a vivid, conversational, and effective tone and style*]	conversevolezza [*conversationality*]	IT5-small	'pop-up', 'comunicativa' [*communicative*], 'teleconducibilità'
			IT5-base	'conversalità' [*conversality*], 'eloquenza' [*eloquence*], 'conversalità'
			MT5-base	'dialettica' [*dialectic*], 'dialogismo' [*dialogism*], 'conversazione' [*conversation*]
			IT5-large	'conversalità' [*conversality*], 'bidimensionalità' [*bidimensionality*], 'iperbole' [*hyperbole*]
	(sost) ##neologismo## Termine usato per descrivere un colpo di stato narcotrafficante, spesso associato a organizzazioni criminali che prendono il controllo di un governo o di un territorio. [*(noun) ##neologism## Term used to describe a drug-trafficking coup, often associated with criminal organizations taking control of a government or territory.*]	narco-golpe	IT5-small	'narcotrafficante' [*drug-trafficker*], 'narcotrafficante', 'antimafia'
			IT5-base	'narco- colpo' [*narco-coup*], 'narco colpo', 'super- colpo di stato' [super- coup]
			MT5-base	'narcotraffico' [*drug trafficking*], 'narcoterrorismo' [*narcoterrorism*], '**narco-golpe**'
			IT5-large	'narcocrazia' [*narcocracy*], 'desaparecido', 'colpo di stato' [*coup*]

DM	source	target	models	predictions
Dict.	##astronomia## ##meteorologia## celeste (agg) [*##astronomy## ##meteorology## celestial (adj.)*]	di o relativo al cielo [*related to the sky*]	IT5-small	che riguarda il sole [*related to the sun*]
			IT5-base	che riguarda il cielo [*related to the sky*]
			MT5-base	che riguarda il sole [*related to the sun*]
			IT5-large	che riguarda il cielo [*related to the sky*]
	classico (sost) [ESEMPIO: la Divina Commedia è un classico della letteratura] [*classic (noun) [EXAMPLE: The Divine Comedy is a classic of literature*]]	qualsiasi opera di riferimento o di particolare importanza [*any work of reference or particular importance*]	IT5-small	genere letterario di grande formato, soprattutto nella letteratura e nel mondo contemporaneo [*a large format literary genre, especially in literature and contemporary world*]
			IT5-base	un libro che ha una lunga tradizione, o che è stato oggetto di studio per molti anni [*a book that has a long tradition, or has been studied for many years*]
			MT5-base	persona che è considerata un grande classico [*a person considered a great classic*]
			IT5-large	libro o opera di letteratura, letteraria, artistica e scientifica che appartiene alla tradizione [*a book or work of literature, arts, or science that belongs to tradition*]
	acciaccare (verb) (senso figurato) [ESEMPIO: la malattia ha un po' acciaccato il nonno, ma sono certo che si riprenderà] [*to weaken, figuratively [EXAMPLE: the illness weakened grandpa, but I'm sure he'll recover*]]	causare grande spossatezza o debolezza fisica, detto in particolare di malattie, sforzi etc. [*to cause great fatigue or physical weakness, especially due to illness, effort, etc.*]	IT5-small	colpire con forza [*to hit hard*]
			IT5-base	sconquassarsi, infermandosi [*to be shaken, becoming ill*]
			MT5-base	rendere difficoltoso, insofferente [*to make difficult, irritable*]
			IT5-large	provocare un affaticamento, una debolezza [*to cause fatigue, weakness*]
Neo.	##neologismo## deplastificazione (sost) [ESEMPIO: L'azienda ha iniziato un percorso di de-plastificazione su prodotti e imballi] [*##neologism## deplasticization (noun) [EXAMPLE: The company has begun a deplasticization process for products and packaging]*]	Rimozione di sostanze plastiche da un materiale o da un prodotto [*Removal of plastic substances from a material or product*]	IT5-small	rimozione di plastificazione [*removal of plasticization*]
			IT5-base	rimozione dei residui di polimerizzazione da una materia [*removal of polymerization residues from a material*]
			MT5-base	Procedura che consiste nell'eliminazione di un tessuto plastico da una parte dell'imballo, in particolare dal legno [*Procedure that eliminates plastic tissue from part of the packaging, particularly from wood*]
			IT5-large	riduzione della quantità di plastica presente in un oggetto [*reduction of the amount of plastic in an object*]
	##neologismo## ecofiction (sost) [ESEMPIO: La prima ecofiction italiana è napoletana. 'La famiglia Lindi', in onda dalla fine di giugno su Canale 9, si ispira alla crisi dei rifiuti.] [*##neologism## ecofiction (noun) [EXAMPLE: The first Italian ecofiction is Neapolitan. 'The Lindi Family', airing at the end of June on Channel 9, is inspired by the waste crisis.]*]	Un genere letterario che include opere narrative focalizzate sulla natura e sull'ecosistema come elementi centrali della trama, del punto di vista o del messaggio [*A literary genre that includes narrative works focused on nature and the ecosystem as central elements of the plot, perspective, or message*]	IT5-small	arte e tecnica di coniugare una cultura eco-sostenibile con la natura [*art and technique of combining an eco-sustainable culture with nature*]
			IT5-base	romanzo o fiction che tratta tematiche ambientali [*novel or fiction dealing with environmental themes*]
			MT5-base	(Iron.) Storia incentrata sul rispetto dei processi ambientali [*(Iron.) Story focused on respect for environmental processes*]
			IT5-large	film o serie televisiva che si ispira alla tutela dei valori ecologici [*film or TV series inspired by the protection of ecological values*]
	##neologismo## plusvalente (agg) [ESEMPIO: Pallotta vende prima tutti i giocatori plusvalenti e poi la Roma stessa] [*##neologism## plus-value (adj), [EXAMPLE: Pallotta first sells all the plus-value players, and then Roma itself]*]	che ha un valore maggiore rispetto a quello di mercato [*that has a greater value compared to the market value*]	IT5-small	che è capace di avere plusvalore [*that is capable of having surplus value*]
			IT5-base	(iron.) di più valore, che ha un notevole pregio [*(iron.) of more value, with notable merit*]
			MT5-base	(Iron.) Che è favorito per qualcosa di importante [*(Iron.) Favored for something important*]
			IT5-large	che vale di più, che possiede un valore maggiore [*that is worth more, that has a greater value*]
Nonce	##neologismo## ecoedificio (sost) [ESEMPIO: L'ecoedificio ha ricevuto numerosi premi per il suo design ecologico.] [*##neologism## eco-building (noun), [EXAMPLE: The eco-building has received numerous awards for its ecological design.]*]	Edificio costruito interamente con materiali riciclati e sostenibili, progettato per un impatto ambientale minimo [*Building entirely constructed from recycled and sustainable materials, designed for minimal environmental impact*]	IT5-small	edificio che fa uso di fonti rinnovabili, ad esempio per edifici pubblici [*building that uses renewable energy sources, e.g., for public buildings*]
			IT5-base	edificio che rispetta i criteri di efficienza energetica [*building that meets energy efficiency criteria*]
			MT5-base	Edificio che è ambientalmente sostenuto [*Building that is environmentally supported*]
			IT5-large	edificio costruito secondo i principi della sostenibilità ambientale [*building constructed according to sustainability principles*]
	##neologismo## fototrasmettere (verb) [ESEMPIO: L'azienda ha sviluppato un nuovo sistema per fototrasmettere dati a velocità incredibili.] [*##neologism## photo-transmit (verb) [EXAMPLE: The company developed a new system to photo-transmit data at incredible speeds.]*]	Trasmettere dati o informazioni utilizzando impulsi di luce, spesso impiegato in tecnologie avanzate di comunicazione [*To transmit data or information using light pulses, often used in advanced communication technologies*]	IT5-small	mandare dati in forma di video e viceversa [*send data in video form and vice versa*]
			IT5-base	trasmettere immagini e dati con mezzi di comunicazione elettronica [*transmit images and data via electronic means of communication*]
			MT5-base	Trasmettere dati mediante la trasmissione di foto [*To transmit data through the transmission of photos*]
			IT5-large	trasmettere dati, informazioni e immagini con la luce [*to transmit data, information, and images using light*]
	##neologismo## solarmobile (sost) [ESEMPIO: Con il solarmobile, ha attraversato l'intero continente senza mai fermarsi per fare rifornimento.] [*##neologism## solar-car (noun) [EXAMPLE: With the solar-car, he crossed the entire continent without ever stopping to refuel.]*]	Automobile alimentata esclusivamente da pannelli solari integrati, progettata per viaggi a lungo raggio senza emissioni [*Automobile powered exclusively by integrated solar panels, designed for long-range travel with no emissions*]	IT5-small	nave da guerra dotata di due passeggeri [*warship equipped with two passengers*]
			IT5-base	telefono cellulare alimentato da energia solare [*cell phone powered by solar energy*]
			MT5-base	Automobile destinato al volo dell'energia solare [*Automobile intended for solar energy flight*]
			IT5-large	mezzo di trasporto pubblico a propulsione elettrica, che sfrutta l'energia solare per produrre carburante [*electric-powered public transport vehicle, harnessing solar energy to produce fuel*]

EM	source	target	models	predictions
Dict.	determinare (verb) [DEFINIZIONE: essere causa di uno specifico effetto o risultato] [*to determine (verb) [DEFINITION: to cause a specific effect or result]*]	l'eccessiva fiducia in sé stesso ha determinato il suo insuccesso [*excessive self-confidence determined his failure*]	IT5-small IT5-base MT5-base IT5-large	determinare la morte di un individuo [*to determine the death of an individual*] la mancanza di lavoro determina una perdita [*lack of work determines a loss*] l'aumento del prezzo di un prodotto non determina la quantità di latte che si può ricavare da una bottiglia [*the increase in a product's price does not determine the amount of milk that can be obtained from a bottle*] il trauma ha determinato la morte del paziente [*the trauma determined the death of the patient*]
	schiavo (agg) [DEFINIZIONE: (per estensione) che non riesce ad avere contegno] [*slave (adj) [DEFINITION: (by extension) someone who cannot control themselves]*]	tu sei schiavo di te stesso [*you are a slave to yourself*]	IT5-small IT5-base MT5-base IT5-large	è schiavo di sé stesso [*he is a slave to himself*] è schiavo di quello che dice [*he is a slave to what he says*] è schiavo di amicizia [*he is a slave to friendship*] è schiavo della droga [*he is a slave to drugs*]
	zompettare (verb) [DEFINIZIONE: ##familiare## muoversi compiendo dei piccoli salti] [*to hop (verb) [DEFINITION: ##familiar## to move by making small jumps]*]	alcuni conigli zompettavano sull'erba del prato [*some rabbits were hopping on the grass in the meadow*]	IT5-small IT5-base MT5-base IT5-large	mi zompettavo in un arco [*I was hopping in an arch*] il cane zompetta per terra [*the dog hops on the ground*] si è zompettato per un po' [*he hopped around for a bit*] il bambino zompettava per casa [*the child was hopping around the house*]
Neo.	iper-estate (sost) [DEFINIZIONE: ##neologismo## Termine usato per riferirsi a un periodo estivo di caldo estremo e prolungato...] [*hyper-summer (noun) [DEFINITION: ##neologism## a term used to refer to an extreme and prolonged hot summer, often associated with heatwaves and drought]*]	In particolare, i fattori che possono contribuire all'iper-estate includono: Effetto serra, Cambiamenti climatici e urbanizzazione [*In particular, the factors contributing to the hyper-summer include: Greenhouse effect, climate change, and urbanization*]	IT5-small IT5-base MT5-base IT5-large	e' la guerra che si disputa ogni due giorni e che si conclude con una fase di iper-estate, al confine tra il nord e l'estonia. [*it is the war that takes place every two days and concludes with a hyper-summer phase, at the border between the north and Estonia*] il rischio è che la vera iper-estate porti con sé anche il surriscaldamento globale. [*the risk is that the real hyper-summer will bring with it global warming*] La maggior parte dei giorni dell'estate ha scoperto che è molto caldo per quanto riguarda l'estate. [*Most of the summer days have found it very hot in terms of summer.*] in questo anno di iper-estate il termometro segna un incremento dell'1,7% rispetto al 1990, e la temperatura si sta rapidamente innalzando. [*in this year of hyper-summer, the thermometer shows a 1.7% increase compared to 1990, and the temperature is rising rapidly*]
	iperturismo (sost) [DEFINIZIONE: ##neologismo## Sovraffollamento turistico, concentrato in alcuni periodi dell'anno...] [*over-tourism (noun) [DEFINITION: ##neologism## high levels of tourism, concentrated in certain periods of the year]*]	Vienna dice no all'iperturismo. [*Vienna says no to over-tourism*]	IT5-small IT5-base MT5-base IT5-large	l'iperturismo è un fenomeno che inonda il turismo, iperturismo è in grado di far diventare iperturistico e scomodo a pochi istanti dai propri monumenti e monumenti. [*over-tourism is a phenomenon that floods tourism, making it overwhelming and uncomfortable near monuments.*] è evidente che il cosiddetto iperturismo, il turismo di massa, è una vera e propria moda. [*it is evident that the so-called over-tourism, mass tourism, is a real trend.*] Lo stesso che iperturismo (da ieri al 22 gennaio) è l'insieme dei turisti che di solito non hanno a che fare con l'Italia come una città. [*The same as over-tourism (from yesterday to January 22) is the group of tourists who usually have nothing to do with Italy as a city.*] «una delle cause principali dell'iperturismo è il sovraffollamento delle città, in particolare del centro storico, e il conseguente deterioramento delle opere architettoniche. [*one of the main causes of over-tourism is the overcrowding of cities, particularly in the historic center, and the consequent deterioration of architectural works.*]
	brevimirante (agg) [DEFINIZIONE: ##neologismo## Attributo di chi mostra una visione ristretta e manca della capacità di anticipare saggiamente gli eventi futuri...] [*short-sighted (adj) [DEFINITION: ##neologism## attribute of someone with a narrow vision and lack of ability to wisely anticipate future events]*]	In un panorama imprenditoriale brevimirante, attento solo e soltanto alla curva del profitto... [*In a short-sighted entrepreneurial landscape, focused only on the profit curve....*]	IT5-small IT5-base MT5-base IT5-large	se l'interstizio tra due stati è il sintomo di un «divido» di cui l'italia si è liberata per il «veicolo» della «pleopamir [*if the gap between two states is the symptom of a "divide" from which Italy has freed itself thanks to the "vehicle" of "pleopamir"*] la sua è una strategia brevimirante. [*his is a short-sighted strategy.*] Nel 2005 ci fu un'inchiesta sulla scelta del governo della destra che la Francia, oltre a trasformare in un escenario, prolungato, è stata un'ipotesi brevimirante di una partita culturale a lungo [*In 2005, there was an inquiry into the right-wing government's choice that France, besides transforming into a long scenario, was a short-sighted hypothesis in a long cultural game*] il governo, e soprattutto il pd, cercano di essere meno brevimiranti possibile. [*the government, especially the PD, are trying to be as little short-sighted as possible.*]
Nonce	geovela (sost) [DEFINIZIONE: ##neologismo## ##geologia## ##tecnologia## Vela navigabile che raccoglie dati geologici mentre si sposta su laghi o mari, utilizzata in esplorazioni scientifiche.] [*ecosail (noun) [DEFINITION: ##neologism## ##geology## ##technology##sail that collects geological data while moving across lakes or seas, used in scientific explorations]*]	La geovela ha fornito nuove informazioni sulla composizione del fondale marino. [*The ecosail provided new information about the composition of the seabed.*]	IT5-small IT5-base MT5-base IT5-large	«in effetti, il valore della geovela è determinato dal fatto che le geovela per l'afa, la maremma e l'atmosfera sono al centro della sua esistenza. [*in fact, the value of the ecosail is determined by the fact that ecosail for the heat, the Maremma, and the atmosphere are at the center of its existence.*] la geovela, le geovele hanno fatto il giro del mondo. [*the ecosail, the ecosails have traveled around the world.*] Il termine «geovela» può essere utilizzato per indicare una nuova ipotesi per la ricerca a livello del Mediterraneo e della ricerca di terra, delle scienze naturali, del patrimonio umano o di una regione geografica [*The term "ecosail" can be used to indicate a new hypothesis for research in the Mediterranean region and land research, natural sciences, human heritage, or a geographic region.*] il brevetto per le geovela è stato rilasciato in svezia nel 2016. [*the patent for the ecosail was issued in Sweden in 2016.*]
	quantosensibile (agg) [DEFINIZIONE: ##neologismo## ##fisica## Sensibile ai fenomeni quantistici, spesso utilizzato per descrivere dispositivi o materiali che reagiscono a variazioni quantistiche.] [*quantum-sensible (adj) [DEFINITION: ##neologism## ##physics## sensitive to quantum phenomena, often used to describe devices or materials that react to quantum variations]*]	Il rilevatore quantosensibile ha misurato anche le più piccole fluttuazioni di energia. [*The quantum-sensitive detector measured even the smallest energy fluctuations.*]	IT5-small IT5-base MT5-base IT5-large	per il professor antonio mazzini si tratta della prima «gara d'avorio» (cioè una «carte di acqua») di elica emilio rocchi, la prima, ma ancora più importante: «ti [*for Professor Antonio Mazzini, this is the first "ivory race" (i.e., a "water card") of Emilio Rocchi's propeller, the first, but even more important: "ti..."*] il nanoscopio nano-chimico è tantosensibile da essere considerato, a conti fatti, un ipotetico e ipotetico dispositivo analogico. [*the nano-chemical nanoscope is so sensitive that, in fact, it is considered a hypothetical and analog device.*] In Italia la soluzione è quella di un sistema quantosensibile, che ha a che fare a favore della distribuzione delle acque (circa 300 miliardi di chilometri), che rappresenta invece l'insieme delle onde di trasporto [*In Italy, the solution is a quantum-sensitive system, dealing with water distribution (about 300 billion kilometers), representing the set of transport waves.*] il nostro cervello è un corpo quantosensibile. [*our brain is a quantum-sensitive body.*]
	suonovento (sost) [DEFINIZIONE: ##neologismo## Un suono che sembra generato dal vento, portando con sé messaggi o emozioni non udibili altrimenti.] [*sound-wind (noun) [DEFINIZIONE: ##neologism## a sound that seems generated by the wind, carrying messages or emotions otherwise unheard.]*]	Il suonovento che attraversava la foresta narrava storie dimenticate da tempo. [*The sound-wind that passed through the forest narrated long-forgotten stories.*]	IT5-small IT5-base MT5-base IT5-large	è in programma il prossimo weekend, e le «fiamme» al «gridovento» al policlinico san paolo e al sant'agostino di santa lucia al sacro cuore. [*It is scheduled for next weekend, and the "flames" at the "gridovento" at San Paolo Policlinic and Sant'Agostino at the Sacred Heart.*] il suonovento, l'eco di elton john, in questi giorni è più intenso. [*the sound-wind, the echo of Elton John, has been more intense these days.*] Un brano di «suono vento» suscita molto commozione. [*A piece of "sound-wind" evokes a lot of emotion.*] in questi anni il suonovento ha fatto il boom, ma questa volta si è spostato dal pop all'emotivo: un mix di musica jazz, etnica e popolare che ha fatto impazzire il pubblico. [*in recent years, the sound-wind has boomed, but this time it has shifted from pop to emotional: a mix of jazz, ethnic, and popular music that has wowed the audience.*]

Definitions	Model	Predicted Word	Adhesion	Novelty
Un'acqua speciale che conserva i ricordi delle persone che vi si immergono, trasmettendoli a chi la tocca. [*A special water that preserves the memories of people who immerse themselves in it, transmitting them to anyone who touches it.*]	IT5-small	acquadimmergente [*immerging water*]	3.0	4.4
	IT5-base	acquarifuga [*memory-repelling water*]	1.8	5.0
	MT5-base	extraacquea [*extra-water*]	2.0	3.2
	IT5-large	acquamemoria [*memory-water*]	5.0	3.6
	GPT-4o	acquamemoria [*memory-water*]	4.2	3.6
##architettura## ##energia## Abitazione futuristica che sfrutta l'energia eolica per la propria autosufficienza energetica, progettata per ambienti ventosi. [*##architecture## ##energy## Futuristic house that harnesses wind energy for self-sufficiency, designed for windy environments.*]	IT5-small	eolita [*eolite*]	4.4	3.8
	IT5-base	windowcase	2.4	3.2
	MT5-base	ecoabitazione [*eco-habitation*]	3.8	3.2
	IT5-large	eohouse	3.4	4.2
	GPT-4o	eolicasa [*windhouse*]	4.2	4.2
(iron.) Formaggio talmente buono che sembra venire dallo spazio. [*Cheese so good it seems to come from space.*]	IT5-small	verzucchero	1.2	3.6
	IT5-base	supercereale [*super-cereal*]	2.0	2.6
	MT5-base	superformaggio [*super-cheese*]	3.0	1.6
	IT5-large	astrocacio [*astro-cheese*]	3.2	3.6
	GPT-4o	cosmaggio [*cosmo-cheese*]	3.2	3.4
Struttura volante utilizzata per raccogliere e analizzare i flussi d'aria a diverse altitudini, impiegata in studi climatici. [*Flying structure used to collect and analyze air currents at different altitudes, employed in climate studies.*]	IT5-small	autobottega [*auto-shop*]	1.4	2.6
	IT5-base	pneumoscopio [*pneumo-scope*]	2.4	2.6
	MT5-base	aereometro [*aerometer*]	4.0	2.8
	IT5-large	turbosnodo [*turbo-joint*]	2.4	3.8
	GPT-4o	aeroflux [*aero-flux*]	3.8	4.0
##design## Che cambia colore in risposta al tocco, utilizzato in arte interattiva e design tecnologico. [*##design## Changes color in response to touch, used in interactive art and technological design.*]	IT5-small	superlucido [*super-glossy*]	2.2	2.6
	IT5-base	ipercolore [*hyper-color*]	3.0	4.2
	MT5-base	trasformacolore [*color-transformer*]	3.8	2.0
	IT5-large	biosensibile [*bio-sensitive*]	3.0	2.8
	GPT-4o	cromotattile [*chromo-tactile*]	4.0	3.6
##biologia## Vernice biologica che cambia colore in risposta a stimoli ambientali, utilizzata per segnalazione visiva e decorazione interattiva. [*##biology## Biological paint that changes color in response to environmental stimuli, used for visual signaling and interactive decoration.*]	IT5-small	biopolvere [*bio-powder*]	2.4	3.4
	IT5-base	biopittura [*bio-paint*]	4.0	3.2
	MT5-base	biovernice [*bio-varnish*]	4.0	2.6
	IT5-large	biopittura [*bio-paint*]	4.0	3.2
	GPT-4o	biocromatica [*biochromatic*]	4.0	3.0
##astronomia## Una notte particolarmente luminosa, dove le stelle e la luna sembrano quasi illuminare a giorno. [*##astronomy## A particularly bright night, where the stars and the moon seem to almost light up the day.*]	IT5-small	supernotte [*super-night*]	2.6	2.2
	IT5-base	lunabeauty [*luna-beauty*]	2.2	4.8
	MT5-base	ultraluminescenza [*ultra-luminescence*]	3.0	2.0
	IT5-large	neonavigazione [*neo-navigation*]	1.4	3.4
	GPT-4o	nottechiara [*clear-night*]	4.8	2.4
##musica## Modificare o adattare il tono di un suono, utilizzato in contesti musicali o tecnologici per regolare le qualità sonore. [*##music## Modify or adapt the tone of a sound, used in musical or technological contexts to regulate sound qualities.*]	IT5-small	soundinserire [*sound-insert*]	2.0	4.2
	IT5-base	tonotizzare [*tonotize*]	3.0	3.8
	MT5-base	audioregolare [*audio-regulate*]	4.4	1.6
	IT5-large	audiomodificare [*audio-modify*]	4.4	2.6
	GPT-4o	sonomodulare [*sono-modulate*]	3.6	3.6
##tecnologia## Trasmettere dati o informazioni utilizzando impulsi di luce, spesso impiegato in tecnologie avanzate di comunicazione. [*##technology## Transmit data or information using light pulses, often employed in advanced communication technologies.*]	IT5-small	fototrasmettere [*photo-transmit*]	4.6	3.0
	IT5-base	fototrasmettere	4.2	2.4
	MT5-base	fototrasmettere	4.2	2.4
	IT5-large	fototrasmettere	4.4	3.0
	GPT-4o	fototrasmettere	4.0	2.4
##politica## Quando il potere politico di uno stato è influenzato da paesi esterni. [*##politics## When the political power of a state is influenced by external countries.*]	IT5-small	extracraziale [*extra-cratial*]	3.0	3.6
	IT5-base	intraneopolitismo [*intra-neopolitism*]	2.0	4.0
	MT5-base	esteropolitica [*foreign-politics*]	3.2	2.2
	IT5-large	mediorientalismo [*middle-orientalism*]	2.0	2.0
	GPT-4o	eteropolitica [*hetero-politics*]	2.0	3.4
##botanica## Piante geneticamente modificate in grado di codificare e immagazzinare dati digitali nelle loro strutture cellulari. [*##botany## Genetically modified plants capable of encoding and storing digital data in their cellular structures.*]	IT5-small	biodipendente [*bio-dependent*]	2.2	3.4
	IT5-base	nanobloccanti [*nano-blocking*]	2.8	4.0
	MT5-base	micropiante [*micro-plants*]	3.0	2.4
	IT5-large	digitoterapie [*digital-therapies*]	1.8	3.6
	GPT-4o	fitodigitale [*phyto-digital*]	4.4	3.6
##moda## Quasi trasparente, ma con una lucentezza opaca che lascia intravedere appena le forme sottostanti. [*##fashon## Almost transparent, but with a matte sheen that barely reveals the underlying shapes.*]	IT5-small	paraluminescente [*para-luminescent*]	3.8	3.2
	IT5-base	semichiaro [*semi-clear*]	3.0	1.8
	MT5-base	ultratrasparente [*ultra-transparent*]	1.8	2.6
	IT5-large	aeroluminescente [*aero-luminescent*]	2.0	3.0
	GPT-4o	luminebbioso [*lumi-misty*]	3.2	4.2
##musica## Dispositivo musicale che riproduce suoni e melodie basate sul ritmo biologico di chi lo utilizza, adattandosi all'umore del momento. [*##music## Musical device that plays sounds and melodies based on the user's biological rhythm, adapting to their mood.*]	IT5-small	biomemoria [*bio-memory*]	2.2	2.8
	IT5-base	etomusica	2.0	5.0
	MT5-base	biomelodico [*bio-melodic*]	3.8	2.8
	IT5-large	biotape	3.0	4.0
	GPT-4o	bioarmonico [*bio-harmonic*]	3.0	3.6
##psicologia## L'irrefrenabile desiderio di viaggiare. [*##psychology## The uncontrollable desire to travel.*]	IT5-small	viaggiapia	3.0	4.0
	IT5-base	micelinismo	2.0	3.8
	MT5-base	viaggiofobia [*travel-phobia*]	1.0	3.0
	IT5-large	travelmania [*travel-mania*]	3.6	2.0
	GPT-4o	viaggiomania [*travel-mania*]	4.0	2.4

Table 12: Additional generated nonce words with novelty and adhesion scores for various models.