Components of Creativity: Language Model-based Predictors for Clustering and Switching in Verbal Fluency

Sina Zarrieß, Simeon Junker, Judith Sieker, Özge Alaçam
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{sina.zarriess;j.sieker;simeon.junker;oezge.alacam}@uni-bielefeld.de

Abstract

Verbal fluency is an experimental paradigm used to examine human knowledge retrieval, cognitive performance and creative abilities. This work investigates the psychometric capacities of LMs in this task. We focus on *switching* and *clustering* patterns and seek evidence to substantiate them as two distinct and separable components of lexical retrieval processes in LMs. We prompt different transformer-based LMs with verbal fluency items and ask whether metrics derived from the language models' prediction probabilities or internal attention distributions offer reliable predictors of switching/clustering behaviors in verbal fluency. We find that token probabilities, but especially attention-based metrics have strong statistical power when separating between cases of switching and clustering, in line with prior research on human cognition.

1 Introduction

The processes underlying human creative abilities have been an important topic of research in several fields. Research in cognitive science suggests that semantic association and search are core aspects of creative thinking (Mednick, 1962; Gilhooly et al., 2007; Beaty and Silvia, 2012). Therefore, creative abilities in humans are commonly tested and measured using semantic search tasks such as verbal fluency, in which participants are asked to list lexical items for a given category in a short period of time (e.g., name as many animals as possible in 60 seconds) (Beaty et al., 2014a).

Human responses to such tasks exhibit a well-known search pattern, which has been termed "clustering and switching" or "exploitation and exploration" (Troyer et al., 1997). During clustering, humans generate sequences of words that belong to the same subcategory, exploiting the neighbourhood of previous items in the semantic space. As this subcategory becomes increasingly exhausted, they switch to other subcategories, shifting their attention to a different patch in their conceptual space (see Figure 1). Recent work suggests that clustering and switching are two fundamental components of semantic search related to creative abilities and has aimed to identify neurocognitive correlates of these processes (Ovando-Tellez et al., 2022).

In this paper, we investigate whether transformer language models (LMs) provide further evidence for the hypothesis that creative semantic search in verbal fluency involves two distinct, separable processes related to clustering and switching. The design of our experiments follows Ovando-Tellez et al. (2022), who tested correlations between the occurrence of clusters and switches in participants' responses to fluency tasks and metrics for participants' creativity, semantic network structure, and brain connectivity. In our study, we replace these metrics of human neuro-cognitive processes with a set of probability and attention-based measures computed with language models over human verbal fluency sequences. We test whether these measures provide predictors of clusters and switches in the human sequences, e.g., whether attention is distributed differently in the LM when retrieving a word within a cluster as compared to a switch.

Our motivation for studying clustering and switching in verbal fluency using LMs is twofold: First, we note that cognitive science has a long-standing interest in computational models that capture human behavior in verbal fluency and other creative search tasks. Existing models in this area typically implement graph-based semantic networks and explicit search algorithms on top of these networks (Hills et al., 2012; Zemla and Austerweil, 2017). We believe that LMs are an obvious alternative modeling approach worth exploring here since their implicit semantic representations and word prediction processes have been shown to excel in a variety of generative tasks. LM-based correlates of clustering and switching would demonstrate the

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 216–232
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

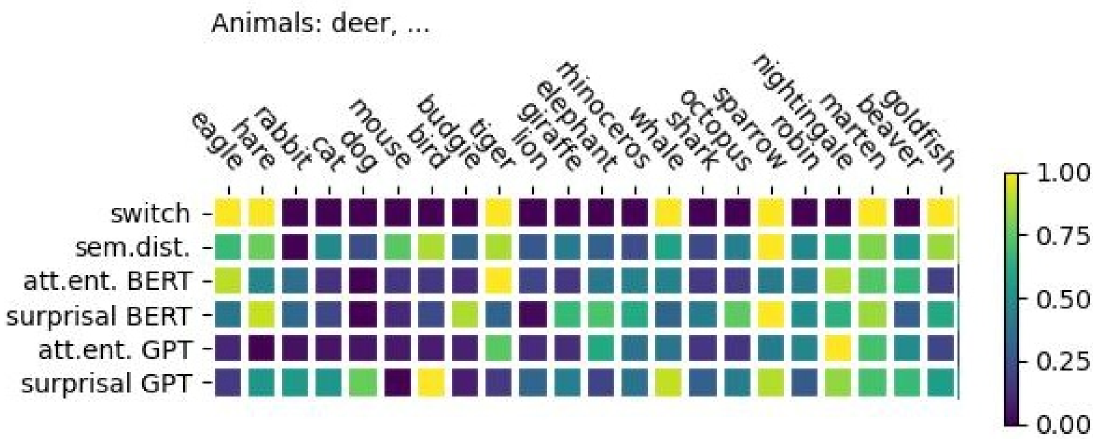


Figure 1: Translated verbal fluency response from BIEFU (Alacam et al., 2022) with annotations of clusters and switches (first row); semantic distances (cosine distances of ConceptNet embeddings) between consecutive items; LM predictors: attention entropy and surprisal predictors from BERT and GPT respectively (all scores are min-maxed normalized for visualization). "Animals: deer" is the LM prompt used to (re)-generate the sequence.

potential of LMs to complement the landscape of computational approaches in this field and, in particular, to provide an account of general language and word sequence processing mechanisms in verbal fluency that are hard to come by in small-scale net-work-based models (cf. Heineman et al., 2024). At the same time, research on LMs is increasingly interested in testing their elementary language processing abilities. Recent studies have tested the extent to which surprisal or attention-based scores computed with LM predict human reading times, providing a cognitively plausible account of processing difficulties in reading and language comprehension (see Oh and Schuler, 2022; Shain et al., 2024). The verbal fluency paradigm complements the landscape of existing probing tasks and analysis methods toward production-oriented tasks involving semantic search and creative abilities. In this study, we ask whether LM-based metrics separate between clustering and switching, as two central components of creative semantic search. Our results suggest that LMs provide novel and strong predictors for modeling human behavior in the verbal fluency task and that attention distribution in LMs has predictive power in accounting for clustering and switching.

2 Background

2.1 Verbal fluency

The verbal fluency task is a neuropsychological test of verbal functioning that is commonly used to measure cognitive performance in e.g. lexical knowledge and retrieval or executive control (Shao et al., 2014). We focus on categorical fluency, which involves repeated retrieval of lexical items for the same category. This gets more challenging when easily accessible words are exhausted and participants are required to transition from fast, associative processes to a more controlled semantic search (Demetriou and Holtzer, 2017). Verbal fluency data is often analyzed in terms of *clusters* and *switches* structuring the word sequence, i.e., word spans that fall into the same semantic subcategories or transitions between subcategories (Troyer et al., 1997; Kim et al., 2019). In Figure 1, for example, the sequence *rabbit, cat, dog, ...* corresponds to a cluster followed by switch from *budgie* to *tiger*. Words within a cluster are typically produced in a fast, associative manner. Switches, in turn, show longer retrieval times as they involve effortful search, executive control (i.e. inhibition of common or previous items), and efficient navigation of long-term semantic memory (Michalko et al., 2023).

The interaction of clustering and switching that typically appears in human verbal fluency responses plays an important role in creativity research (Silvia et al. 2013; Beaty et al. 2014b; Beaty and Kenett 2023, among others). Ovando-Tellez et al. (2022) show that clustering is related to *divergent thinking*, i.e., generating new and effective ideas, while switching is connected with *convergent thinking* or combining available information in creative ways, and both are characterized by distinct brain connectivity patterns. They argue that clustering involves associative abilities, while switching requires controlled memory retrieval processes, executive functions and memory.

2.2 Computational Models of Verbal Fluency

The computational modeling of verbal fluency data has received considerable attention in cognitive science research. Existing models typically implement the generation of verbal fluency responses as a search over a semantic network or graph (Hills et al., 2012; Abbott et al., 2015; Zemla and Austerweil, 2017; Avery and Jones, 2018), where clustering and switching emerges from the search strategy as in the foraging model by (Hills et al., 2012) or from the underlying structure of the network as in the model by (Abbott et al., 2015). To a similar end, other approaches make use of biologically inspired neural networks (Kajić et al., 2017) or, more recently, pre-trained transformer models (Nighojkar et al., 2022) and LLMs (Heineman et al., 2024; Wang et al., 2025). In general, these models are tested for their ability to predict or simulate human fluency sequences on a word level.

Other computational work on verbal fluency focused explicitly on automatically annotating clustering-switching patterns in sequences produced by humans. Some studies have explored the use of distributional semantic representations and word embeddings for scoring semantic fluency data (Linz et al., 2017; Paula et al., 2018; Kim et al., 2019; Alacam et al., 2022) or the ability of pre-trained LMs in predicting category switches (Heineman et al., 2024).

In contrast to these models, our study does not aim to explicitly reproduce or simulate the semantic search strategies observed in human verbal fluency responses with LMs. Instead, we focus on investigating their underlying word retrieval and prediction processes. Inspired by Ovando-Tellez et al. (2022), we ask whether we can identify distinct components of verbal fluency, i.e. clustering and switching, from processing-related behavioural measures computed with an LM.

2.3 Linguistic and Cognitive Probing of LMs

The analysis of linguistic and cognitive capabilities captured in LMs has become an important area of research (Belinkov and Glass, 2019; Baroni, 2022; Chang and Bergen, 2023; Binz and Schulz, 2023; Strachan et al., 2024). A common paradigm in LM probing is behavioral analysis, which treats the pretrained LM as a black box and uses carefully controlled test suites or experimental datasets from (psycho-)linguistics to compare model outputs against human productions or judgments. This paradigm is useful for testing whether LMs learn particular linguistic rules and generalizations, in particular in the domain of syntax (Warstadt et al., 2020), but provide very limited insights into how underlying processing mechanisms in LMs align to human language processing and cognition (cf. Baroni, 2022; Chang and Bergen, 2023).

Other work on probing LMs focuses on their ability to account for effects of processing difficulty, and mostly goes back to the idea of "surprisal" (Hale, 2001; Levy, 2008; Demberg and Keller, 2008; Smith and Levy, 2013). Surprisal is defined as the negative log probability of a word in context and has been demonstrated to provide a very robust predictor for human processing times (e.g., to reading times) when computed with language models of different sizes or perplexities (Goodkind and Bicknell, 2018; Shain et al., 2024). These findings lend support to expectation-based accounts of sentence processing in psycholinguistics, aligning word prediction processes in LMs with humans' anticipation of upcoming material in sentence reading. A few recent studies explored further predictors complementing surprisal. Thus, the attention mechanism of transformer LMs has been considered to approximate aspects of memory and attention in human cognition (Ryu and Lewis, 2021; De Varda and Marelli, 2024). Most importantly for our study, Oh and Schuler (2022) showed that attention distribution and distance metrics from internal layers of the LM yield very powerful predictors for self-paced reading times and gaze durations in naturalistic reading, drawing connections to memory-based accounts of sentence processing. As memory is an important aspect of semantic search in the verbal fluency task (Ovando-Tellez et al., 2022), our study will examine both surprisal (or, more generally, probability-based) predictors computed at the LM's output layer as well as attention-based predictors from the internal layers.

However, although LMs are now frequently used as computational testbeds for theories of language processing and cognition, the field is still debating which of the many existing LMs can provide the most robust and cognitively plausible predictors of human processing. Oh et al. (2022) tested surprisal estimates from GPT-2 models of different sizes and showed that the surprisal computed with smaller model sizes achieved a better fit with human reading times than larger model sizes. Similar observations have been made in (Kuribayashi et al., 2022; Oh and Schuler, 2023). Wilcox et al. (2023), on the other hand, trains LMs of small and medium size on a range of languages and finds that LM quality generally correlates with its psychometric predictive power. Therefore, in the following, we will rely on some less recent but widely used LMs such as BERT or GPT-2, but also include variants of more recent models available in different sizes.

3 Motivation and research questions

The main question of this work is whether current transformer LMs can account for effects of processing difficulty in a creative word retrieval task – verbal fluency – where clear differences in retrieval difficulty have been widely observed in terms of clustering-switching patterns (Troyer et al., 1997; Hills et al., 2012). In the following, we will detail the assumptions underlying this question.

Why could prediction and attention mechanisms implemented in LMs explain effects of processing difficulty in the verbal fluency task? At a basic level, verbal fluency involves repeated retrieval of lexical items, which aligns well with the autoregressive, left-to-right word prediction objective implemented in modern LMs. Research on verbal fluency in psychology and linguistics typically emphasizes that the verbal fluency task involves a whole range of different cognitive and verbal abilities, such as access to the mental lexicon, semantic knowledge, search strategies, language processing, executive control functions, long-term memory, and attention (Kim et al., 2019; Michalko et al., 2023; Ovando-Tellez et al., 2022). Importantly, many previous studies have found strong evidence for lexical access and language production processes being critical components in verbal fluency (Weckerly et al., 2001; Whiteside et al., 2016; Marko et al., 2023). Therefore, we believe that LMs with their complex underlying architecture for representing and modeling word sequences may offer additional benefits over traditional, relatively small-scale network models (Hills et al., 2012; Abbott et al., 2015) building on Markovian assumptions and being detached from general language processing accounts (Heineman et al., 2024).

Which LM-based predictors can be expected to account for processing effects of clustering and switching in verbal fluency? While previous modeling approaches typically rely on some form of semantic distance to account for clustering-switching patterns, this work proposes to use word prediction and attention-based measures computed from LMs as proxies of retrieval difficulty in verbal fluency. We expect these predictors to inherently account for sequence processing effects and to capture retrieval difficulties beyond semantic-taxonomic distances. As a motivating example, consider the first cluster of the sequence in Figure 1 corresponding to common “pets” (*rabbit, dog, cat,* ...). Here, attention entropy and surprisal scores computed with BERT predict that these words are easy to retrieve, matching the annotation as a cluster. However, the semantic distance predicts greater difficulty, potentially due to taxonomic distances between, e.g. *mouse* and *bird*. In simple terms, we assume that words corresponding to switches and higher retrieval difficulty in humans are modeled as less predictable and requiring higher attention-entropy in LMs. For instance, the word *dog* following *cat* should have a relatively low surprisal compared to the word *tiger* following *budgie*, as illustrated in Figure 1. Higher attention entropies, in turn, indicate that the model distributed attention weights more evenly across the preceding sequence which in Figure 1 is often the case for words corresponding to switches (*tiger, whale, sparrow, ...*).

4 Experimental Method

4.1 Data

We base our experiments on BIEFU (Alacam et al., 2022), a dataset of German verbal fluency responses, which covers a fairly high number of categories. The BIEFU data was collected from 100 participants and contains verbal fluency responses that enumerate words for 10 different semantic categories (e.g., animals, hobbies, body parts). An overview of the data is shown in Table 4 (App. A).

Soft and Hard Switches The BIEFU dataset includes manual annotations of lexical items with semantic subcategories. Based on these, we determine soft (fluid) and hard (static) switches, following Zemla and Austerweil (2019). A soft cluster switch occurs when the next word in a list does not share a sub-category label with the previous word, while a hard switch occurs whenever the next word does not share a sub-category label with any of the previous words since the start of the last cluster. Soft switches are the most commonly examined types of switches in the literature and we will focus on these in the following.

4.2 Prompting

To obtain prompts from human verbal fluency sequences, Nighojkar et al. (2022) replaced the last item in a partial verbal fluency sequence with a mask token, cf. (1).

(1) *[C]*s I know are w_{n-1-ct}, ..., w_{n-1}, and the [MASK] .

Here, w_{n-1-ct} (ct being the context size) is the initial and w_{n-1} the penultimate item in a sequence produced for category *C*. [MASK] always represents the last item. We adopt this scheme and iteratively mask out subsequent items in each human-produced sequence, i.e., shift the masked token from left to right by truncating them at the position of the masking token, cf. the prompts in Table 1. Baseline prompt (pr-0), which consists of a simple enumeration preceded by the category name, is added for comparison. Since LMs can be very sensitive to the specification of their prompts, we conducted further experiments with prompt design that addresses both auto-regressive and bidirectional prompt strategies with different wording variations, see Table 5 (App. B.2) for additional results on these.

Seq:	dog, cat, mouse, ...
pr-0_1	Animals: dog, [MASK]
pr-0_2	Animals: dog, cat, [MASK]
pr-1_1	Animals I know are dog, [MASK]
pr-1_2	Animals I know are dog, cat, [MASK]

Table 1: A (translated) sample of a human response and derived LM prompts for two subsequent steps in a verbal fluency sequence for autoregressive prompting.

4.3 Language Models

Since our investigation is one of the first to test the predictive power of LMs in distinguishing clustering and switching, we select basic transformer LMs that have also been widely used in the literature on cognitive probing – GPT-2 (Radford et al., 2019), BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020). Next to these, we also include open-source German or multilingual models that come in different size – Bloom (350m, 1b5, 1b7) (Scao et al., 2023) and XGLM (560M, 1b7) models (Lin et al., 2021). This model selection ensures a representative comparison across transformer architectures that employ different versions of the self-attention mechanism: BERT as a bidirectional encoder model, GPT-2, Bloom and XGLM as unidirectional autoregressive decoder models, and T5 as an encoder-decoder transformer.

4.4 Predictors of Switching and Clustering

We use generalized linear mixed-effect models to test the predictive power of probability-based and attention-based metrics derived from LMs to separate clustering and switching in verbal fluency data. In the following, we describe the predictors we include in this statistical analysis.

4.4.1 Psycholinguistic Predictors

We implement a strong baseline model that predicts clustering/switching based on fixed and random effects established in recent verbal fluency literature (Michalko et al., 2023). These predictors are *temporal order*, *task demand*, *Typicality*, *Inter-response similarity*. We add the participants and semantic categories as a crossed random effect to the initial model ($m0$).

Temporal order (TEMP). The normalized temporal order (TEMP) corresponds to the current position of the word in a sequence divided by the length of that sequence (range between 0 and 1). This predictor captures the fact that words are more difficult to produce the longer the sequences become.

Task demand (TD). This predictor reflects that certain verbal fluency categories are systematically easier to enumerate than others, due to their familiarity, frequency, and lexical specificity. For instance, categories like *animals* and *vegetables* are easier to enumerate since they are more frequent, while other categories like *fabrics* or *insects* are less easily accessible. Following Michalko et al. (2023), we manually group the verbal fluency categories into three so-called "task demand categories".

Typicality (TYP). Next, we add a fixed effect that captures the typicality of an item within a verbal fluency category (TYP). TYP is calculated as the logarithm of the absolute number of occurrences of a word among all items enumerated by all participants within that particular category. See App. A for further detail.

Inter-response similarity (IRS) We compute the semantic similarity of subsequent lexical items in a verbal fluency sequence. Here, we deviate slightly from Michalko et al. (2023) and use the cosine similarity between the items' word embeddings, computed with the ConceptNET Numberbatch word embedding. This semantic space is enriched with ConceptNet taxonomic relations (Speer et al., 2017), achieves the best performance in predicting clustering and switching patterns in BIEFU data Alacam et al. (2022).

Retrieval latency (RL) Our data records time stamps of every typed character in the verbal fluency sequence. We define retrieval latency as the time span as the offset between a preceding item and the onset of the next item. We calculate it by subtracting the offset of the first item from the onset of the second item.

4.4.2 Probability-based Predictors

Our first set of LM predictors is derived from word probabilities. We regard these as measures of retrieval difficulty or predictability in sequence generation, mirroring the notion of "expectation" in sequence understanding (Shain et al., 2024). We expect that clustering corresponds to less surprising items, whereas switching should show higher surprisal and lower probabilities. The handling of words composed of subwords across different LM architectures is detailed Appendix B.1. To test this hypothesis, we consider the following predictors:

Surprisal (Surp.). We transform word probabilities into surprisal scores, quantifying the information content it conveys in the context in which it appears. The surprisal of a word w is calculated as the negative log-likelihood of its probability. We expect a positive correlation with switching.

$$\text{Surprisal}(w_i) = -\log_2 p(w_i \mid w_{<i})$$

Rankings (Rank). This predictor derives determines the rank of the word w in the word probability distribution. We expect a positive correlation with switching. The rank parameter is highly dependent on the vocabulary size of the LM architecture. In our analysis, the rank scores are normalized, but see Appendix B.4 for more information.

$$\text{Rank}(w) = \text{arg min}_i \, \{p(w \mid \text{context}) : i = 1, 2, \ldots, N\}$$

Entropy (Ent.). As another account of retrieval difficulty in context, we include the entropy of the word probability distribution, quantifying the model's uncertainty in the given context, regardless of the probability or rank of the target item. We expect a positive correlation with switching.

$$\text{Entropy}(w_i) = -\sum_{w_i} p(w_i \mid w_{<i}) \log_2 p(w_i \mid w_{<i})$$

4.4.3 Attention-based Predictors

The second set of LM predictors derives from the model's internal attention distributions as measures of cognitive effort, related to monitoring and shifting working memory and attention (Ryu and Lewis, 2021; De Varda and Marelli, 2024). We expect that switching corresponds to higher cognitive effort, e.g., wider attention distributions across layers and heads, than clustering which we expect to show more localized attention patterns.

We extract the attention-based predictors considering different layers and attention heads in the transformer architecture (144 heads in total for the smaller LMs, 256 for the larger LMs). We first transform the embeddings of tokens or hidden states of a sequence to a triple of query (q), key (k), and value (v) embeddings. The heads then compute the attention weight between the query and key vectors for all pairs of tokens in the input prompt as soft-max-normalized dot products.

$$\alpha_{ij} = \frac{exp(q_i^T k_j)}{\sum_{l=1}^{n} exp(q_i^T k_l)}$$

The diffuseness of attention obtained from these attention maps α can be calculated in different ways. We follow Clark et al. (2019) and consider attention head entropy and distance between attention distribution for subsequent items in the sequence.

Average Attention-Heads Entropy (AHE). The attention entropy is calculated in a similar way to the probability-based entropy metric. The key distinction lies in its application to attention weight distributions instead of a softmax-adjusted probability distribution. Subsequently, the attention entropy is obtained by averaging across all heads for the respective iteration of the input prompt. High entropy is associated with bag-of-words context incorporation (Clark et al., 2019).

$$Entropy(\text{head}) = -\sum_{i=1}^{N} \alpha(i) \log_2 \alpha(i)$$

Here, $\alpha(i)$ represents the probability associated with the i-th element in the attention distribution.

Average JS-Divergence in attention heads (AH-JSD). To explore whether attention heads in the same layer can be grouped based on similar behavior, we compute the distances between all pairs of attention heads. This pairwise distance between the attention distribution of each pair of heads H_i and H_j is calculated using Jenson-Shannon Divergence following (Clark et al., 2019). Lower divergence indicates that all heads process the inputs in a similar way.

$$JSD = \sum_{token \in Prompt} JS(H_i(token), H_j(token))$$

5 Experiments

We now describe our experiments, testing the predictive power of LM predictors in distinguishing between clustering and switching in a creative semantic search task. All analyses were carried out in R version 2024.12.x (R Core Team, 2021). The models are compared using ANOVA and all numerical values are (z-)normalized using the *scale* function in R.

5.1 Baseline Models

We use mixed-effect logistic regression (glmer) and fit them on annotations of switching and clustering in human verbal fluency responses. The dependent variable is coded as a binomial variable (0: cluster, 1: switch), indicating clustering or switching between consecutive words in a sequence.

For the baseline model, we applied forward stepwise inclusion starting with $m0$ which has only crossed-random effects of participant and category. The order of the inclusion of the parameters in the baseline is from more basic (temporal order) to complex (retrieval latency). For model $m1$ to $m5$, we add the baseline predictors from Section 4.4.1 as follows:

$$\begin{aligned}
m0 &: switch + (\sim 1|part.) + (\sim 1|cat.)\\
m1 &: m0 + \text{TEMP}\\
m2 &: m1 + \text{TASKDEMAND}\\
m3 &: m2 + \text{TYP}\\
m4 &: m3 + \text{IRS}\\
m5 &: m4 + \text{RL}
\end{aligned}$$

The temporal order parameter did not improve the model fit ($\chi^2(1) = 1.31, p > .05$). Adding task demand (TEMP) has a significant effect ($\chi^2(2) = 6.64, p < .05$). The main effects of the typicality (TYP) and of the inter-response similarity parameter (IRS) were also found significant (($\chi^2(1) = 44.63, p < .0001$) and ($\chi^2(1) = 3384, p < .001$), respectively). For the hard switch, all parameters significantly contributed to model fit (see Appendix B.3 for the details). The results indicate that $m5$ is the strongest baseline for switch modeling.

This set of baseline models, commonly used in the verbal fluency literature, enables us to quantify and compare the individual contributions of a rich array of LM predictors that we propose.

5.2 Models with LM predictors

Next, we analyze the power of LM predictors in modeling clustering and switching. The following model list shows in which order the probability and attention-based variables from Sections 4.4.2 and 4.4.3 are included:

$$\begin{aligned}
lm_m6 &: (m3,\ m4, or\ m5) + Prob_{LMtype}\\
lm_m7 &: (m3,\ m4, or\ m5) + Rank_{LMtype}\\
lm_m8 &: (m3,\ m4, or\ m5) + Ent_{LMtype}\\
lm_m9 &: (m3,\ m4, or\ m5) + AHE_{LMtype}\\
lm_m10 &: (m3,\ m4, or\ m5) + AH - JSD_{LMtype}
\end{aligned}$$

Thus, adding LM predictors to $m3$ shows the contribution of probability and attention-based predictors to a model that includes the baseline predictors of temporal order, task demand, and typicality. Then, we test the predictive power of LM parameters to the $m4$ model, which includes a significant predictor for semantic similarity between consecutive words (IRS). Finally, we add them to the $m5$ model, which further includes retrieval latency (RL), a highly predictive variable for clustering and switching.

5.3 Results

Table 2 summarizes the contribution of each LM predictor for soft switch modeling when added to the defacto baseline model ($m3$). The results for $m3$ in Table 2 show clear evidence for the predictive power of LM predictors, in separating between clustering and switching processes. The attention-based metric AH-JSD, in particular, models these processes very robustly and independently from the underlying LM, i.e. it is highly significant for all LMs. This also holds for the AHE metric, which achieves slightly lower values across the board, though. The probability-based metrics are less consistent across LMs: T5, Bloom350, and XGLM yield a highly significant RANK variable while surprisal is less significant. However, SURPRISAL derived from BERT achieves substantial predictive power, comparably to AHE. Most probability-based predictors from GPT-2 are insignificant.

Analysis with Concept Similarities and Retrieval Latency. We further investigate the re-

Table 2: Soft Switch: the individual contributions of LM-predictors to the base model ($m3$) (Chi-Square). *** denotes significance (p) < 0.001. ** : p < 0.01 and * : p < 0.05

		BERT	T5	GPT-2	Bloom350	Bloom1b5	Bloom1b7	XGLM560	XGLM1b7
$m3$	**Prob**	37.44***	11.28***	2.20	0.65	0.62	2.64	4.68*	15.26***
	Rank	9.64**	51.25***	1.49	50.41***	0.74	2.41	67.79***	76.78***
	Surprisal	64.08***	12.89***	3.86*	46.99***	23.09***	2.78	30.25***	17.25***
	Entropy	2.91	0.83	3.54	33.02***	0.72	1.03	63.16***	3.21
	AHE	60.43 ***	33.66***	45.02***	32.31***	32.31***	31.97***	52.68***	52.68***
	AH-JSD	**106.26** ***	**63.64*****	**92.35** ***	**71.07*****	**68.34*****	**73.56*****	**85.11*****	**79.52*****

Table 3: Soft Switch: the individual contributions of LM-predictors on top of $m4$ and $m5$ models (Chi-Square). *** denotes p < 0.001. ** : p < 0.01 and * : p < 0.05

		BERT	T5	GPT-2	Bloom350	Bloom1b5	Bloom1b7	XGLM560	XGLM1b7
$m4$	**Prob**	1.56	22.85 ***	29.05 ***	**56.94** ***	50.15	0.005	9.77**	15.11***
	Rank	4.26 *	16.50 ***	8.35 **	35.96 ***	48.61***	8.92**	14.92***	29.39***
	Surprisal	10.76 **	0.89	1.34	53.55 ***	**74.28*****	1.22	0.02	4.19*
	Entropy	0.15	0.97	1.96	42.27***	0.79	0.01	**71.10*****	7.19**
	AHE	46.65 ***	21.03 ***	31.24 ***	20.27 ***	20.27***	15.79***	34.95***	34.95***
	AH-JSD	**71.41** ***	**29.28** ***	**58.64** ***	38.61 ***	35.16***	**34.88*****	43.88***	**39.10*****
$m5$	**Prob**	1.85	**24.05** ***	**30.05** ***	**56.13** ***	50.69***	0.001	8.81**	16.83***
	Rank	4.54 *	17.57 ***	6.80 **	33.12***	43.35***	**7.13****	15.82***	**28.39*****
	Surprisal	8.95 **	2.06	0.93	51.61 ***	**74.14*****	1.55	0.20	4.68*
	Entropy	0.49	1.59	2.05	35.39 ***	1.32	0.02	**69.39*****	6.12*
	AHE	14.99 ***	3.51 *	7.44 **	2.67	2.67	1.11	8.73**	8.73**
	AH-JSD	**24.93** ***	4.02 *	17.93 ***	7.28 **	5.71*	5.13*	8.64**	6.35*

lationship between LM parameters and semantic similarity (IRS) – one of the most frequently used NLP metrics in verbal fluency modeling – as well as retrieval latency (RL) as a strong behavioural measure of processing difficulty. Table 3 summarizes the contribution of each LM predictor for soft switch modeling when added to the $m4$, and $m5$ models, respectively. Looking at the results for $m4$, we find that a number of LM predictors remain highly significant, even on top of the strong similarity variable IRS. This holds in particular for the attention-based metrics, most notably for AH-JSD. This confirms our hypothesis that attention distributions in the internal layers of LMs capture aspects of processes in semantic search beyond static similarities in embedding space. However, we also see notable differences in how predictors from different LMs interact with IRS. Bloom350 and Bloom1b5's attention-based metrics seem to be more closely aligned with the IRS parameter (resulting in lower contributions) compared to their probability-based parameters. The probability-based predictors of BERT, however, are not significant anymore when combined with IRS.

The results for $m5$ closely align with those of $m4$, with the primary difference being a substantial decrease in the magnitude of contribution for attention-based models. As $m5$ includes the highly significant retrieval latency parameter from the human data, we take this as a promising finding suggesting that attention-based metrics derived from LMs show some alignment with humans internal retrieval processes. The inclusion of retrieval latency does not influence the contribution of probability-based metrics which supports the view that they capture complementary aspects of clustering and switching in our data.

LM Comparison. When comparing all three testing conditions, attention-based metrics are the most robust predictors across different LM architectures. Their predictive power only decreases when added after the retrieval latency parameter, which suggests that attention-based predictors are highly aligned with retrieval latency in humans. For the final $m5$ model, the probability-based metrics from small German Bloom models remain highly significant. Interestingly, we observe a similar effect here to other studies on surprisal (Oh and Schuler, 2023), i.e. their predictive power decreases with increasing model size. Similarly, we see some advantages of the smaller XGLM560 over the larger XGLM1b7. Finally, next to model size, we see great differences between predictors computed from different transformer architectures (BERT, GPT2, T5). For instance, AH-JSD from BERT remains significant in $m5$, while the same is not true for T5 or GPT-2. This suggests that attention patterns learned in different architectures capture different aspects of humans' cognitive processes, supporting further research into novel LM architectures (Charpentier and Samuel, 2024).

Finally, we complement the chi-square-based evaluation with the model ranking according to AIC scores (quantifying model fitness) in Appendix Figure B.5. The AIC-based analysis confirms the pattern described above. Among all variations for the base model ($m3$), AH-JSD metric derived from BERT had the highest model fit. However, for the enriched models incorporating semantic similarity ($m4$) and retrieval latency ($m5$), larger models – particularly BLOOM1b5 and XGLM560 – demonstrate superior performance.

5.4 Discussion

Our experiments on verbal fluency add to the existing evidence that language models show some degree of human-likeness in their internal processing mechanisms (cf. Kuribayashi et al., 2025). Thus, we find that well-known predictors derived from LMs' word predictions, i.e., surprisal and related measures, as well as predictors computed from LMs' attention distributions, have strong statistical power when separating between clustering and switching in human verbal fluency responses.

For research on creativity in human cognition, this result supports the assumption that different processes are at play in creative semantic search tasks (Ovando-Tellez et al., 2022). When LMs regenerate humans' verbal fluency responses, they show clearly distinct attention and prediction patterns that neatly align with annotations of clustering and switching in these sequences. Previous studies identified these patterns based on distances in word embedding spaces (Alacam et al., 2022). Our study complements this with further metrics computed, in particular, from the LMs' internal attention distribution. These attention-based LM predictors remained significant even when added to a baseline model that included a semantic distance-based variable (IRS). This suggests that attention distributions capture processing-related mechanisms in verbal fluency beyond semantic distances.

The fact that attention-based predictors are superior to probability-based metrics in our verbal fluency setting supports previous work proposing that attention patterns in transformer LMs could reflect processes or retrieval and memory search (Ryu and Lewis, 2021; De Varda and Marelli, 2024). The creative search processes involved in verbal fluency pose particularly strong demands on memory and executive processes of working memory and inhibition (Shao et al., 2014). This further underlines the plausibility of our findings and explains why surprisal predictors, which are prominent in studies on processing difficulty in natural reading, show less consistent patterns than attention-based metrics.

While recent work on cognitive probing of LMs has mostly focused on autoregressive GPT-style architectures, our results show that attention predictors from encoder models like BERT outperform GPT models. This is surprising since autoregressive word prediction and causal, left-to-right self-attention seem intuitively more aligned with incremental sequence generation in verbal fluency. A hypothesis to explore in future work is that the bidirectional self-attention in the BERT encoder could allow the model to obtain a richer semantic space and account for more complex attention and retrieval operations involved in a challenging semantic search task.

Finally, our study points to new directions for the cognitive probing of LMs. Whereas most previous work looked at modelingreading times, our study shows the fitness of LM predictors in accounting for generative and creative tasks. Future tasks to consider could be related to naming (Silberer et al., 2020), reference (Junker and Zarrieß, 2024) or association (Chen and Ding, 2023).

6 Conclusion

Our work contributes to understanding the processing mechanisms of LMs with the help of verbal fluency, an established experimental task from cognitive science research. We showed that LMs can distinguish two central components of creative sematic search, clustering and switching, via their metrics derived from their attention and probability distributions. Our study is one of the first to show that distributions of attention weights in the internal layers and attention heads of the transformer architecture correlate to a great extent with processing difficulty in a creative semantic search task.

Limitations

We have employed the vanilla versions of the selected language models and all the metrics derived from the models were not subjected to heavy transformations except the basic soft-max, negative log-likelihood, or pooling over layers and attention heads. Since the evidence from the analysis points towards the advantage of using attention-based metrics, further investigation on calculating different attention scores (Oh and Schuler, 2022) is a promising line of research.

The verbal fluency data were processed using off-the-shelf NLP text processing tools. Compound words are generally common in German, and the vocabulary used by participants also frequently contains compound words such as “Klavierspielen” (piano playing), “Krankenpfleger” (health nurse), “Fahrradfahren” (bike riding). Unfortunately, many of the compounds do not exist in the vocabulary of the static embedding models such as ConceptNet, whereas BERT and succeeding language models can deal with out-of-vocabulary tokens due to their sub-word tokenization method.

Ethical Statement

Our study utilizes a published and openly available dataset with annotations on verbal fluency, without annotator-related information. Additionally, we ensure that our use of the dataset aligns with its intended purpose.

Acknowledgments

We thank the anonymous reviewers for their helpful and detailed comments. This research has been partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project no. 512393437, project B02. The authors also acknowledge financial support by the project “SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems” (Grant ID NW21-059A), which is funded by the program “Netzwerke 2021” of the Ministry of Culture and Science of the State of North Rhine-Westphalia, Germany.

References

Joshua T Abbott, Joseph L Austerweil, and Thomas L Griffiths. 2015. Random walks on semantic networks can resemble optimal foraging. In *Neural Information Processing Systems Conference; A preliminary version of this work was presented at the aforementined conference.*, volume 122, page 558. American Psychological Association.

Özge Alacam, Simeon Schüz, Martin Wegrzyn, Johanna Kißler, and Sina Zarrieß. 2022. Exploring semantic spaces for detecting clustering and switching in verbal fluency. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 178–191, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Johnathan Avery and Michael N Jones. 2018. Comparing models of semantic fluency: Do humans forage optimally, or walk randomly? In *CogSci*.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic structures in natural language*, pages 1–16.

Roger E. Beaty and Yoed N. Kenett. 2023. Associative thinking at the core of creativity. *Trends in Cognitive Sciences*, 27(7):671–683.

Roger E Beaty and Paul J Silvia. 2012. Why do ideas get more creative across time? an executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of aesthetics, creativity, and the arts*, 6(4):309.

Roger E Beaty, Paul J Silvia, Emily C Nusbaum, Emanuel Jauk, and Mathias Benedek. 2014a. The roles of associative and executive processes in creative cognition. *Memory & cognition*, 42:1186–1197.

Roger E. Beaty, Paul J. Silvia, Emily C. Nusbaum, Emanuel Jauk, and Mathias Benedek. 2014b. The roles of associative and executive processes in creative cognition. *Memory amp; Cognition*, 42(7):1186–1197.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language Model Behavior: A Comprehensive Survey. *Computational Linguistics*, pages 1–55.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. BERT or GPT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Honghua Chen and Nai Ding. 2023. Probing the “creativity” of large language models: Can models produce divergent semantic association? In *Findings* *of the Association for Computational Linguistics: EMNLP 2023*, pages 12881–12888, Singapore. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Andrea De Varda and Marco Marelli. 2024. Locally biased transformers better align with human reading times. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 30–36, Bangkok, Thailand. Association for Computational Linguistics.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Eleni Demetriou and Roee Holtzer. 2017. Mild cognitive impairments moderate the effect of time on verbal fluency performance. *Journal of the International Neuropsychological Society*, 23(1):44–55.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth J Gilhooly, Evridiki Fioratou, Susan H Anthony, and Victor Wynn. 2007. Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, 98(4):611–625.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

David Heineman, Reba Koenen, and Sashank Varma. 2024. Towards a path dependent account of category fluency. In *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*.

Thomas T Hills, Michael N Jones, and Peter M Todd. 2012. Optimal foraging in semantic memory. *Psychological review*, 119(2):431.

Simeon Junker and Sina Zarrieß. 2024. Resilience through scene context in visual referring expression generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 344–357, Tokyo, Japan. Association for Computational Linguistics.

Ivana Kajić, Jan Gosmann, Brent Komer, Ryan W. Orr, Terrence C. Stewart, and Chris Eliasmith. 2017. A biologically constrained model of semantic memory search. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, London, UK. Cognitive Science Society.

Najoung Kim, Jung-Ho Kim, Maria K. Wolters, Sarah E. MacPherson, and Jong C. Park. 2019. Automatic scoring of semantic fluency. *Frontiers in Psychology*, 10.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *arXiv preprint arXiv:2502.01615*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Nicklas Linz, Johannes Tröger, Jan Alexandersson, and Alexandra Konig. 2017. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*, pages 1–7, Montpellier, France).

Martin Marko, Drahomír Michalko, Jozef Dragašek, Zuzana Vančová, Dominika Jarčušková, and Igor Riečanskỳ. 2023. Assessment of automatic and controlled retrieval using verbal fluency tasks. *Assessment*, 30(7):2198–2211.

Sarnoff Mednick. 1962. The associative basis of the creative process. *Psychological review*, 69(3):220.

Drahomír Michalko, Martin Marko, and Igor Riečanskỳ. 2023. Executive functioning moderates the decline of retrieval fluency in time. *Psychological Research*, 87(2):397–409.

Animesh Nighojkar, Anna Khlyzova, and John Licato. 2022. Cognitive modeling of semantic fluency using transformers. In *Cognitive Aspects of Knowledge Representation workshop at 31st International Joint Conference on Artificial Intelligence*.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5:777963.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Marcela Ovando-Tellez, Mathias Benedek, Yoed N Kenett, Thomas Hills, Sarah Bouanane, Matthieu Bernard, Joan Belo, Theophile Bieth, and Emmanuelle Volle. 2022. An investigation of the cognitive and neural correlates of semantic memory search related to creative ability. *Communications Biology*, 5(1):604.

Felipe Paula, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. Similarity measures for the detection of clinical conditions with verbal fluency tasks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 231–235, New Orleans, Louisiana. Association for Computational Linguistics.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Zeshu Shao, Esther Janse, Karina Visser, and Antje S. Meyer. 2014. What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, 5:1–10.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paul J. Silvia, Roger E. Beaty, and Emily C. Nusbaum. 2013. Verbal fluency and creativity: General and specific contributions of broad retrieval ability (gr) factors to divergent thinking. *Intelligence*, 41(5):328–340.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7):1285–1295.

Angela K Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *neuropsychology*, 11(1):138.

Ye Wang, Yaling Deng, Ge Wang, Tong Li, Hongjiang Xiao, and Yuan Zhang. 2025. The fluency-based semantic network of llms differs from humans. *Computers in Human Behavior: Artificial Humans*, 3:100103.

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

Jill Weckerly, Beverly Wulfeck, and Judy Reilly. 2001. Verbal fluency deficits in children with specific language impairment: Slow rapid naming or slow to name? *Child Neuropsychology*, 7(3):142–152.

Douglas M Whiteside, Tammy Kealey, Matthew Semla, Hien Luu, Linda Rice, Michael R Basso, and Brad Roper. 2016. Verbal fluency: language or executive function measure? *Applied Neuropsychology: Adult*, 23(1):29–34.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7511, Singapore. Association for Computational Linguistics.

Jeffrey C Zemla and Joseph L Austerweil. 2017. Modeling semantic fluency data as search on a semantic network. In *CogSci... Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference*, volume 2017, page 3646. NIH Public Access.

Jeffrey C Zemla and Joseph L Austerweil. 2019. Analyzing knowledge retrieval impairments associated with alzheimer's disease using network analyses. *Complexity*, 2019.

Appendix

A BIEFU data

Table 4 presents basic statistics for word counts and retrieval latencies for BIEFU verbal fluency sequences within each category and across categories (as *global*). This overview highlights some characteristic differences between the categories: participants enumerated almost 11.5 items on average. For the *animals* and *countries*, the number is high as 19.11 and 18.5 respectively, while it is around or below 10 items for *fabrics*, *insects*, and *vessels*. Correspondingly, retrieval latency for *countries*, *animals*, *groceries* and *body parts* are significantly lower than categories that are less easy to enumerate such as *fabrics or insects*.

Table 4 also includes typicality and IRS scores that we will use as predictors in our baseline model. The IRS is the cosine similarity between consecutive words calculated with ConceptNet Numberbatch embeddings (Speer et al., 2017). We observe that the categories *insects* and *fabrics* which elicited the smallest number of words (tokens and types) across participants show the lowest typicality values, i.e. participants retrieved relatively few and rather divergent sets of words. Interestingly, *hobbies* and *occupations* exhibit high typicality, i.e. show more overlap between participants, but also show the lowest IRS scores, i.e. they contain words that have more distant embedding in semantic space. The categories *clothes*, *body parts*, *insects*, and *vessels* exhibit the highest IRS scores. Based on the provided dataset, we further calculate the retrieval latencies between each consecutive items. The mean retrieval latencies shown in Table 4 further differentiate the overall picture. Here, the categories *countries* and *animals*, the most widely used category in verbal fluency, show the lowest mean retrieval latencies, together with high typicality and medium IRS.

Task demands For creating the task demand categories for BIEFU in a similar way as in Michalko et al. (2023), we have looked at the held-out sequences (from another 100 participants on the same categories, but without retrieval latency scores) and calculated the basic statistics similar to Table 4 except the retrieval latency score. Based on these scores, we categorized the BIEFU categories into three groups depending on the cognitive effort needed to enumerate them. The low-demand category consists of *animals*, *body parts* and *countries*. *Hobbies*, *occupations*, *groceries* and *clothes* belong to the moderate category. Finally, the high demand category includes *fabrics*, *vessels* and *insects*.

B Language Models

We utilize the verbal fluency data in German by (Alacam et al., 2022) and we employ various distinct language models for German : (i) a pretrained German BERT model[1] (ii) a pretrained German GPT-2 model[2] , and (iii) a pretrained T5 model[3] for German.

In this way, we aim to minimize any potential impact of the training data's nature on the overall performance of our models. We generally use the Hugging Face[4] framework for reproducibility.

Next to these common LMs, we evaluate two more recent autoregressive models on the dataset, investigating the effects of model size and the difference between monolingual and multilingual language models. Specifically, we employ (i) a monolingual BLOOM model that is trained from scratch on German data, comprising 350M parameters[5],

[1] https://huggingface.co/dbmdz/bert-base-german-cased.

[2] https://huggingface.co/dbmdz/german-gpt2.

[3] https://huggingface.co/t5-base.

[4] https://huggingface.co/.

[5] https://huggingface.co/malteos/bloom-350m-german.

Table 4: BIEFU: Basic statistics (Max, min, and average values of sequences, retrieval latency and sub-category counts per semantic category)

Categories	Token Count in a Sequence	Mean Retrieval Latency (in sec.)	Total Token (Type) Count	Subcat. Count	Typicality (mean)	IRS Similarity (mean)
animals	**Max: 34, Min: 8, Mean: 19.11**	1,96	1548 (202)	22	4.53	.39
body parts	**Max: 28 , Min: 8 , Mean: 17.02**	2.50	1571 (144)	8	3.98	.50
clothes	Max: 24, Min: 7, Mean: 16.5	2.31	1434	15	4.04	.52
countries	**Max: 33, Min: 6, Mean: 18.5**	1.81	1688 (140)	6	4.19	.42
fabrics	Max: 14, Min: 5, Mean: 7.9	**5.06**	633 (142)	15	3.94	.39
groceries	Max: 28, Min: 7, Mean: 16.6	2.32	1550 (276)	14	4.69	.42
hobbies	Max: 25, Min: 6, Mean: 14.49	2.63	1333 (302)	31	4.86	.32
insects	Max: 17, Min: 5, Mean: 9.47	**4.21**	843 (99)	14	3.67	.49
occupations	Max: 20, Min: 6, Mean: 12.23	2.89	1113 (296)	19	4.91	.35
vessels	Max: 17, Min: 5, Mean: 10.13	3.83	902 (166)	9	4.13	.46
Global	Max: 34, Min: 5, Mean: 11.51	3.05	19518 (2763)	153	4.13	.43

(ii) a multilingual BLOOM model adapted to the German language via the CLP-Transfer method with 1.5B parameters[6], and (iii) a multilingual BLOOMoom with 1.7B parameters[7]. Furthermore, we use (iv) a multilingual XGLM model with 564M parameters[8], comparable in size to the monolingual BLOOM model, and (v) a multilingual XGLM model with 1.7B parameters[9], equivalent in size to the biggest multilingual BLOOM model.

We omit models like Chat-GPT or GPT-4 from our analysis since these do not generally provide token probabilities or attention distributions through their respective APIs and, hence, do make it possible to compute the type of measures and predictors we need for our investigation.

B.1 Tokenization

We first tokenize the masked prompt with the word w masked out by a single mask token m) and pass it through the model. We then restrict the output logits of the model to the position of the masked token and pass them through a softmax function to obtain a probability distribution over the model's vocabulary for the position of m. In the resulting distribution, we select the probability of w, the entropy of the distribution as well as the rank of w in the model's vocabulary sorted by the probability. In addition to this, we also store the attention map over the whole sequence. The subword tokenization of BERT and T5 complicates this process, i.e. w is not always represented by a single token in the model's vocabulary, but may consist of multiple subword tokens (such as *[Kol, ##ib, ##ri]* for the word *Kolibri (hummingbird)*). In such cases, we iteratively replace m with each subword token for w and take the average of the log probabilities of all subwords as well as the lowest rank of any subword as representative of the whole item w. Such a method is considered useful for extracting a more meaningful score for the multiword expressions like *[Großer Panda (Big Panda), Rote Paprika (Red paprika)]*. For the autoregressive GPT-2, BLOOM and XGLM models, where utilizing a masked token isn't feasible, we truncate the prompt at the position of the masked item and then pass it through the models. The process of extracting probabilities, ranks, surprisal scores, and entropies with GPT-2, BLOOM and XGLM models mirrors that are utilized for BERT and T5 models. This also extends to the handling of the subword tokens, as the autoregressive models employ the same tokenization strategy.

B.2 Prompt Design

Since existing LMs can be very sensitive to the specification of their prompts, we also test several prompt variations for the calculation of probabilities and attention distributions for verbal fluency sequences. Depending on the type of LM, these prompts can be divided into (i) unidirectional prompts that only include left context for masked tokens and (ii) bidirectional prompts where masked tokens are presented in a left and right context. In the following, we describe the design of the verbal fluency prompts.

[6] https://huggingface.co/malteos/bloom-1b5-clp-german.

[7] https://huggingface.co/bigscience/bloom-1b7.

[8] https://huggingface.co/facebook/xglm-564M.

[9] https://huggingface.co/facebook/xglm-1.7B.

Table 5: A sample of a human response and derived LM prompts for two subsequent steps in a verbal fluency sequence (1st step/left, 2nd step/right column), as input for autoregressive prompting. For T5, we use identical prompts to BERT but replace [MASK] with the sentinel token.

Original Sequence	Hund (dog), Katze (cat), Maus (mouse)	
Target token	Katze in the 1st step	Maus in the 2nd step
Prompt-0	*(Animals: Dog, [MASK])* Tiere: *Hund, [MASK]**	Tiere: *Hund, Katze, [MASK]**
Prompt-1	*(Animals I know are dog and [MASK].)* Tiere, die ich kenne, sind *Hund und* [MASK]*	Tiere, die ich kenne, sind *Hund, Katze und* [MASK]*
Prompt-2	*(Examples of animals are dog, [MASK])* Beispiele für Tiere sind *Hund und [MASK]**.	Beispiele für Tiere sind *Hund, Katze, und [MASK]**.
Prompt-3	*(The first animals that come to my mind are dog, [MASK], mouse.)* Die ersten Tiere, die mir einfallen, sind *Hund und [MASK]**.	Die ersten Tiere, die mir einfallen, sind *Hund, Katze und [MASK]**.
Prompt-4	*(Animals one can know are dog and [MASK].)* Tiere, die man kennt, sind *Hund und* [MASK]*	Tiere, die ich kenne, sind *Hund, Katze und* [MASK]*
Prompt-5	*(When I think of animals, I think of dog and [MASK].)* Wenn ich an Tiere denke, dann denke ich an *Hund und* [MASK]*	Wenn ich an Tiere denke, dann denke ich an *Hund, Katze und* [MASK]*

B.3 Hard Switches

Table 6 summarizes the results for the hard switch modeling when the LM metrics are added to $m3$, $m4$ and $m5$ models.

Unlike soft-switch modeling, the contribution of various metrics in this specific case of switches varies significantly, without exhibiting a consistent pattern across all conditions. A closer examination reveals that among the probability-based metrics, RANK and SURPRISAL are the most influential, often performing on par with AH-JSD or even surpassing it in modeling hard-switch cases. It is important to note that a hard switch occurs when a previously unmentioned subcategory appears in the enumeration. This necessitates metrics that are sensitive to a broader contextual lookback.

Overall, for detecting hard-switches, probability-based metrics demonstrate greater predictive power in decoder-only models, whereas models with encoders benefit substantially from AH-JSD. Further details on these results are provided in Appendix B.3.

Psycholinguistic parameters. In the hard switch condition, adding the retrieval order parameter (TEMP) improves model fit ($\chi^2(1) = 11.58, p < .001$). The task demand also significantly improves the model ($\chi^2(2) = 6.97.87, p < .0001$). The main effects of typicality (TYP) ($\chi^2(1) = 19.76, p < .001$) and the inter-response similarity parameter (IRS) also significantly contributed to explaining the data ($\chi^2(1) = 2990.75, p < .0001$) as well as the retrival latency.

m3 + LM predictors. It is obvious that A closer look reveals that among the probability-based metrics, Rank and Surprisal are the most prominent ones except the GPT-2, Bloom1b5 and Bloom1b7 models. Furthermore, all attention-based metrics contribute significantly to the model fit to a differing extent. Despite not having the highest contribution, almost all metrics derived from XGLM adds explanatory power.

m4 + LM predictors . When we look at the effect of LM metrics for the model with IRS, it is also difficult to see one distinct pattern. Again, Rank and Surprisal parameters are generally more informative than probability or entropy metrics. Bloom1b7 seems to have no contribution on top of basic psycholinguistic parameters. $Entropy$ only contributes to the fitness for Bloom350m.

m5 + LM predictors. In addition to the de facto psycholinguistic parameters, we investigate the effect of a less common parameter in verbal fluency analysis – the retrieval latency – as an indicator of lexical computation in explaining switching /clustering behavior. Then we also examine the alignment between retrieval latency with the LM predictors. To do that, we add the retrieval latency to the $m4$ model. In the both hard and soft switch conditions, we find that the retrieval latency RL further improves the model fitness significantly: ($\chi^2(1) = 344.88, p < .001$) and ($\chi^2(1) = 265.17, p < .001$) respectively.

As summarized in Table 6, the Bloom350 model continues to exhibit a significant effect for its probability-based metrics, followed by Bloom 1b5. Attention-based metrics continue to contribute to the model fitness only for the BERT model, on top of retrieval latency.

Table 6: Hard Switch: the individual contribution of LM-predictors on top of $m3$, $m4$ and $m5$ models (Chi-Square)

		BERT	T5	GPT-2	Bloom350	Bloom1b5	Bloom1b7	XGLM560	XGLM1b7
$m3$	**Prob**	49.67 ***	9.95 **	1.72	0.32	0.37	2.06	19.69 ***	27.02***
	Rank	12.75 **	**57.89***	1.67	44.32 **	0.52	7.65**	**94.86***	**66.05***
	Surprisal	**107.08***	0.06	9.86**	**66.82** **	31.30***	2.07	76.61 **	27.25***
	Entropy	5.24*	0.61	2.05	24.61***	0.89	0.33	21.87**	2.32
	AHE	37.12**	24.97***	18.17**	16.40**	16.40**	16.36**	24.64***	24.64***
	AH-JSD	73.34***	54.89***	**43.45***	40.03***	**37.96** **	**45.91** **	53.31***	48.43***
$m4$	**Prob**	0.10	23.33 ***	18.62 ***	49.31 ***	41.76***	0.01	0.87	**28.64***
	Rank	7.14 **	11.76 ***	**21.49** ***	32.58 ***	26.23***	**19.05** **	**32.53***	23.52***
	Surprisal	39.25 ***	9.73 **	0.52	**78.20***	**89.56***	0.8	14.11***	10.71**
	Entropy	1.50	0.04	2.37	30.98 ***	1.09	0.14	22.05***	4.87*
	AHE	24.40 ***	13.69 ***	8.49 **	7.40 **	7.40**	5.14*	11.65**	11.65***
	AH-JSD	**43.37** ***	**28.45** ***	16.12 ***	16.22 ***	14.09**	16.58**	21.68***	18.13***
$m5$	**Prob**	0.04	**24.80** ***	20.04 ***	48.93 ***	43.01***	0.003	0.48	31.65***
	Rank	7.25 **	9.77 **	**23.08** ***	29.23 ***	21.22***	**15.87***	**34.57***	22.44***
	Surprisal	**35.21** ***	8.62 **	1.55	**75.08** ***	**88.46***	1.15	17.65***	11.8**
	Entropy	2.67	0.03	3.57	24.31 ***	1.77	0.07	20.69***	**38.9***
	AHE	2.28 ***	0.15	0.06	0.37	0.37	1.26	0.00	0.001
	AH-JSD	6.41 *	1.67	0.00	0.01	0.17	0.04	0.11	0.02

B.4 Effect of Vocabulary size on the Rank Parameter

Among the explored metrics, the rank score is highly dependent on the vocabulary size of the language model. Figure 2 plots the predictive power (Chi2) of the RANK parameter when added to the to $m3$, $m4$ and $m5$ models. On the left, the graph shows models with smaller vocabulary sizes (BERT, T5); in the center, models with (relatively) moderate vocabulary sizes (GPT-2, Bloom350m, Bloom1b5); and on the right side, multilingual models with substantially larger vocabulary sizes (Bloom1b7 and XGLM models). This graphs reveals that when the rank is added to $m3$ model, no clear pattern is observed with respect to vocabulary size. On the other hand, when the rank score is added to the m4 and m5 models, there is a slight upward, suggesting a possible relation between rank score and the vocabulary size. However, this trend is still not consistent across models with similar vocabulary sizes.

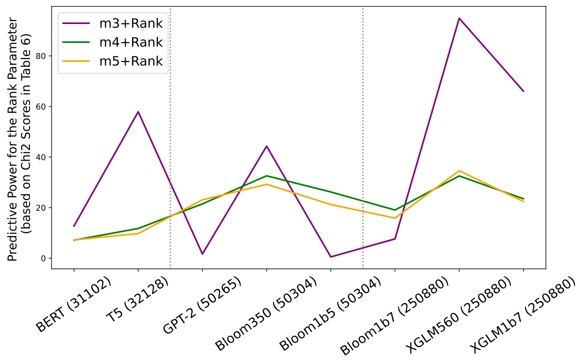
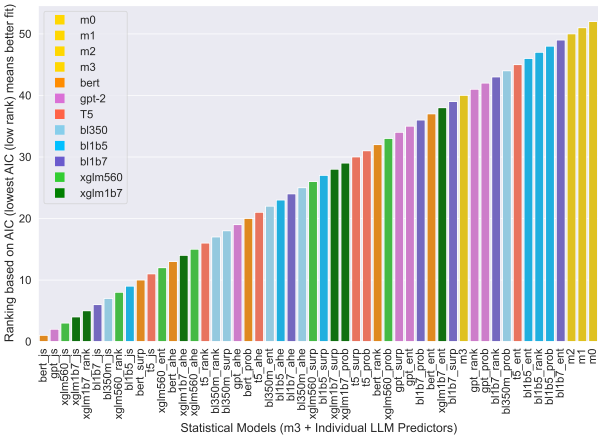


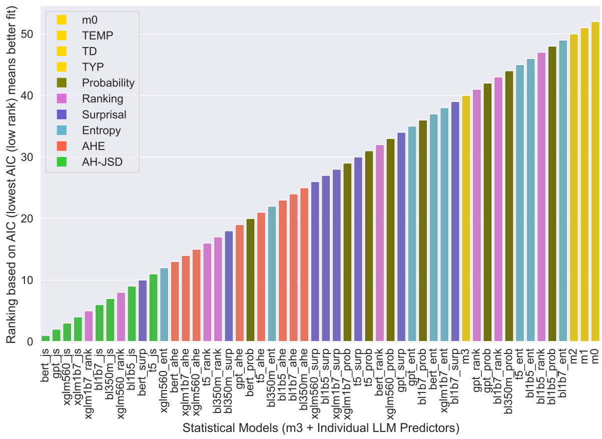
Figure 2: LMs architectures ordered w.r.t their vocabulary size. Y-axes denotes the Chi2 Scores for the rank parameter added to $m3$, $m4$ and $m5$ models.

B.5 AIC Based Ranking

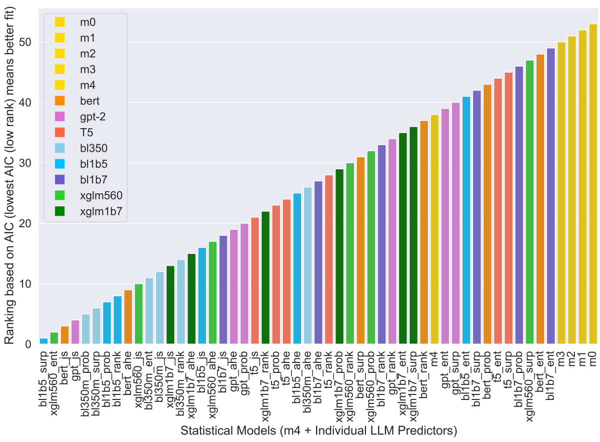
Complementary results for the Section 5.2. While the sub-figures positioned next to each other show the same data, they highlight the different aspects: for example, Figure B.5 (a) is color-coded with respect to the LM type, and Figure B.5 (b) for the effect of the metric. The lowest AIC corresponds to the lowest rank (1st rank/best model).



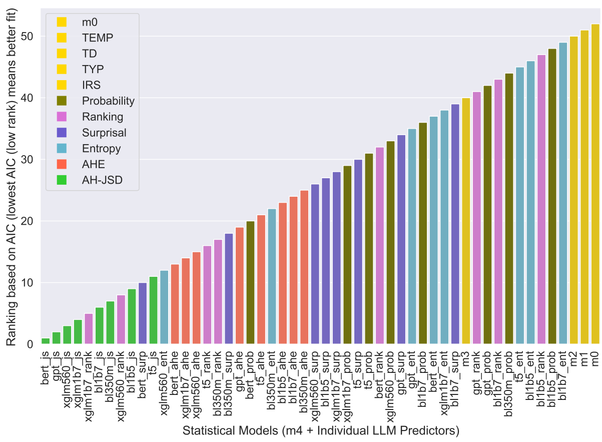
(a) $m3$: Model-based color coded



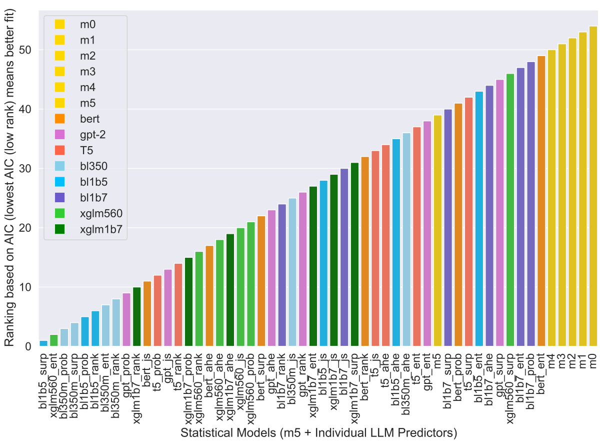
(b) $m4$: Metric-based color coded



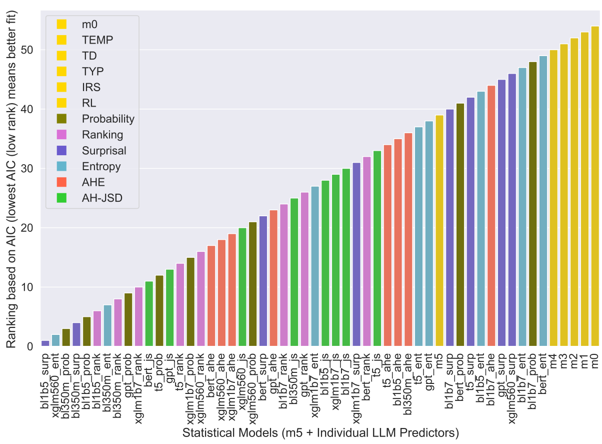
(c) $m4$: Model-based color coded



(d) $m4$: Metric-based color coded



(e) $m5$: Model-based color coded



(f) $m5$: Metric-based color coded

Figure 3: Individual Models' fitness (based on AIC scores