Probing the Category of Verbal Aspect in Transformer Language Models

Anisia Katinskaia,[*♢] **Roman Yangarber**[♢]
[*] Department of Computer Science
[♢] Department of Digital Humanities
University of Helsinki, Finland
first.last@helsinki.fi

Abstract

We investigate how pretrained language models (PLM) encode the grammatical category of verbal aspect in Russian. Encoding of aspect in transformer LMs has not been studied previously in any language. A particular challenge is posed by *"alternative contexts"*: where either the perfective or the imperfective aspect is suitable grammatically and semantically. We perform probing using BERT and RoBERTa on alternative and non-alternative contexts. First, we assess the models' performance on aspect prediction, via behavioral probing. Next, we examine the models' performance when their contextual representations are substituted with counterfactual representations, via causal probing. These counterfactuals alter the value of the "boundedness" feature—a semantic feature, which characterizes the action in the context. Experiments show that BERT and RoBERTa do encode aspect—mostly in their final layers. The counterfactual interventions affect perfective and imperfective in opposite ways, which is consistent with grammar: perfective is positively affected by adding the meaning of boundedness, and vice versa. The practical implications of our probing results are that fine-tuning only the last layers of BERT on predicting aspect is faster and more effective than fine-tuning the whole model. The model has high predictive uncertainty about aspect in alternative contexts, which tend to lack explicit hints about the boundedness of the described action.

1 Introduction

This paper focuses on the grammatical category of *verbal aspect*. It is a category that involves both morphology and semantics of the verb, and expresses how an action denoted by the verb extends over time. Linguistic theory of aspect is intricate: different languages make different aspectual distinctions, e.g., languages can have distinct perfective/imperfective/progressive categories of aspect, while some make no distinction at all. Aspect is also one of the most complex categories in many languages: even advanced experts, who are non-native speakers, continue to make errors in the choice of aspect (Forsyth, 1970; Bar-Shalom and Zaretsky, 2008). We focus on the Slavic aspectual system, in particular in Russian, which displays significant differences in the semantics of the perfective/imperfective opposition to other languages (Dahl, 1985).

How aspect is encoded in pretrained language models (PLMs) has not been previously studied for any language, although other grammatical properties—number agreement, predicate-argument syntactic relations, etc.—have been studied. It is challenging to identify what linguistic phenomena in the context affect the choice of aspect. A special challenge concerning aspect is posed by *"alternative contexts"*, where more than one aspect form is acceptable grammatically and semantically.

We investigate the following research questions: RQ1. Do BERT and RoBERTa encode the category of aspect, and if they do—how? RQ2. How does the encoding of aspect in these models correspond to linguistic theory of aspect? RQ3. Is encoding of aspect in alternative contexts different from non-alternative contexts?

We perform two kinds of probing: behavioral and causal. In behavioral probing, we inspect layers one by one, observing how the model predicts which aspect form best suits the context. If the model fails, we infer that it does not encode the target linguistic property (here—aspect). We introduce two types of behavioral probing via filling a mask: iterative masking and aspect inference. In both methods, the model's preference for aspect is reflected in the probabilities it assigns to verb forms in the masked position. For causal probing, we *intervene* in the model's representations at each layer: we manipulate the semantics of the action described by the target verb and its context—whether the action is *bounded* or *unbounded*. If the intervention is relevant for predicting the target property, the model's

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3347–3366

performance on the task will be affected.

Our findings can be summarised as follows: (1) All probing methods indicate that BERT and RoBERTa do encode aspect, predominantly in their final layers. (2) Interventions in sentence semantics cause effects consistent with theory of aspect: imperfective verbs typically describe *unbounded* actions, while perfective verbs describe *bounded* actions. (3) Fine-tuning only the final layers of BERT for aspect prediction results in improved performance, confirming our first finding. (4) Both pretrained and fine-tuned models exhibit *high uncertainty* regarding aspect preference in *alternative* contexts, where multiple aspect forms are valid. (5) *Alternative* contexts are more sensitive to causal intervention in the semantics of boundedness. (6) Such contexts often lack explicit hints about the action's boundedness, which makes both humans and PLMs uncertain about the choice of aspect.

2 Related Work

Several studies focus on the internal representation of linguistic information inside PLMs. *Correlation probing* methods are based on *parametric probes*, i.e., linear or non-linear classifiers trained on model representations to predict specific linguistic properties (Adi et al., 2017; Conneau et al., 2018; Tenney et al., 2019; Hewitt and Manning, 2019; Dalvi et al., 2019; Maudslay et al., 2020; Weissweiler et al., 2022; Conia and Navigli, 2022; Arps et al., 2022). Some have questioned the efficacy of probing classifiers, and whether the original model, which was used as an encoder, actually uses the information discovered by probes (Hewitt and Liang, 2019; Tamkin et al., 2020; Ravichander et al., 2021). In response to this criticism, a number of methodologies are proposed (Hewitt et al., 2021; Pimentel et al., 2020; Voita and Titov, 2020; Immer et al., 2022; Wang et al., 2023). Belinkov (2022) gives an extensive review of probing classifiers as an approach, their advantages and shortcomings.

Non-parametric correlation probing, or *behavioral probing*, tests the behavior of PLMs without additional classifiers. To isolate the target linguistic property, a PLM is evaluated by using a set of carefully designed examples (Linzen et al., 2016; Gulordava et al., 2018; Ribeiro et al., 2020; Warstadt and Bowman, 2020; Newman et al., 2021; Wu et al., 2020; Li et al., 2023; Amini et al., 2023; Kim et al., 2023). Ravfogel et al. (2019) propose a methodology for creating synthetic examples, which differ by various linguistic properties. While most work focuses on English, Mueller et al. (2020) introduce the CLAMS dataset for syntactic evaluation of models for five languages, including Russian. Hlavnova and Ruder (2023) propose Multilingual Morphological Checklist (M2C), a framework for behavioral probing of typological features in 12 languages, e.g., motion verbs in Russian.

Causal probing relies on controlled interventions into the LM's internal components (or into the input), and studying consequent changes in the model's behavior (Giulianelli et al., 2018; Vig et al., 2020; Elazar et al., 2021; Kaushik et al., 2020; Geiger et al., 2021; Voita et al., 2021; Finlayson et al., 2021; Lasri et al., 2022b; Rozanova et al., 2023; Yamakoshi et al., 2023; Li et al., 2023). *Amnesic* probing (Elazar et al., 2021) builds on the intuition that removing a property from the representation will weaken the model's ability to solve a task, if the property is important for the task. The approach is based on an algorithm—Iterative Null-space Projection (INLP)—for removing linear information from representations (Ravfogel et al., 2020). Ravfogel et al. (2021) apply INLP to generate counterfactual representations and use these to test how changing particular linguistic features affects the model's behavior. Despite some criticism of INLP (Kumar et al., 2022), we use it to investigate the behavior of LMs on aspect prediction.

3 Background on Aspect

The category of aspect in Russian characterizes the action described by a verb in terms of its progress—continuous vs. punctual, completed vs. uncompleted, etc.—or from the observer's perspective—retrospective vs. synchronous. The meaning of aspect opposition has long been a subject of debate. In this paper, we adhere to the theory that *boundedness*—reaching a limit—is the factor determining the aspect form (Vinogradov, 1947; Dahl, 1985). We assume that every verb has two aspect forms—*perfective* and *imperfective*—though in some rare cases the two forms may coincide, e.g., "обещать" (*to promise—perf. or imperf.*).

Unlike most grammatical categories, aspect has no unique marker in the verb form and is tightly connected with the verb's lexical meaning. Aspect can be expressed by the root, e.g., "**говор**ить" (imp.) vs. "**сказ**ать" (perf.), *to say*; by the suffix, e.g., "толк**а**ть" (imp.) vs. "толк**ну**ть" (perf.), *to push*; by the prefix, e.g., "делать" (imp.) vs. "**с**делать"

(perf.), *to do/make*.[1] Examples (1) and (2) show the aspect pair of verbs "дуть / дунуть" (*to blow*):

(1) На побережье всегда дул (imp.) ветер.
Wind always blew on the coast.

(2) И вдруг резко дунул (perf.) ветер.
And suddenly the wind blew sharply.

In these contexts, only one aspect form is acceptable. We call such contexts **non-alternative**. However, in some (narrow) contexts it may not be possible to decide which aspect fits best, since both may fit, albeit with slight differences in meaning. We call such contexts **alternative**. For example, in the sentence below both perfective and imperfective are acceptable:

Я уже позвонил (perf.) в клинику и вызвал врача.
Я уже __звонил (imp.) в клинику и вызвал врача.
I already rang the clinic and called the doctor.

For any particular instance, we use the term *expected* for the original verb form found in the text vs. the opposite form, which we call *complementary*. We perform experiments by probing aspect of the expected vs. the complementary form; we also investigate model behavior in the non-alternative vs. alternative contexts.

4 Experiments

For probing experiments, we use the Russian BERT-base, BERT-large (Devlin et al., 2019), and RoBERTa-large (Liu et al., 2020).[2] We mostly focus on experiments with BERT-large, since other models showed similar performance.

4.1 Data

There are no pre-existing datasets for probing verbal aspect, so we perform our analysis using the following data. For *alternative* contexts, we collected short paragraphs from the ReLCo corpus (Katinskaia et al., 2022); the contexts contain exercises offered to learners of Russian, where they inserted verb forms that *differ* from the expected answers only by the aspect feature. The learners used the Revita language teaching and learning system (Katinskaia et al., 2018, 2017). These forms were manually annotated as acceptable by several native speakers. For non-alternative contexts, we created our own dataset by randomly selecting sentences from the Omnia corpus (Shavrina and Benko, 2019). In each context,[3] we pick one verb (hereafter, the **target**). We generate the target verb's complementary aspect form using a morphological generator (Korobov, 2015); further details in Appendix A.

We tried to ensure that the target verbs are lexically varied. The collected contexts with hidden target verbs and the generated aspect pairs were manually annotated by two native speakers. The annotation task was to assess whether the given verb form fits the context grammatically and semantically. We collected 750 non-alternative contexts—with 375 examples for each aspect—featuring 542 distinct target verb aspect pairs. We expanded the set of alternative contexts to 496 instances in total, with 238 perfective and 258 imperfective verbs. The agreement between the annotators was 84.5%, conflicts were resolved through discussion.

We release the annotated data and the first Russian Aspect Bank with over 2K unique aspect pairs with this paper.[4] The Aspect Bank was manually created in collaboration with experts in Russian linguistics and language pedagogy.

4.2 Behavioral Probing

First, we probe BERT and RoBERTa as Masked Language Models (MLM), in the alternative vs. non-alternative contexts. We evaluate the model's ability to predict aspect in the context by measuring its preference for particular grammatical forms. Typically in this task, the model is prompted to fill in the MASK given the context. The model is deemed successful if it assigns a higher probability to the correct form (Marvin and Linzen, 2018; Lasri et al., 2022a; Amini et al., 2023).

Since Russian is a morphologically rich language, with heavy inflection, its words are often split into segments during tokenization. This is especially relevant for verbal forms, which have multiple inflectional and derivational affixes. Considering this challenge, and the fact that aspect can be marked in the prefix, the stem, or the suffix of a verb, we performed and compared two types of behavioral probing: iterative masking and aspect inference.

Iterative masking entails several *iterations* of filling the mask.[5] First, we pre-segment the target verb V in the input sequence $\mathbf{X}$ into a list of n sub-word tokens $V = [V_1 ... V_n]$, where $n \geq 1$.

[1] The prefix may affect the meaning, but lexical vs. grammatical changes are often very difficult to disentangle; therefore we consider such verb pairs to be aspect pairs.

[2] huggingface.co/ai-forever

[3] An instance is a long sentence or several shorter sentences.

[4] github.com/RevitaAI/AspectProbing

[5] For full detail, please see the algorithm in Appendix 1.

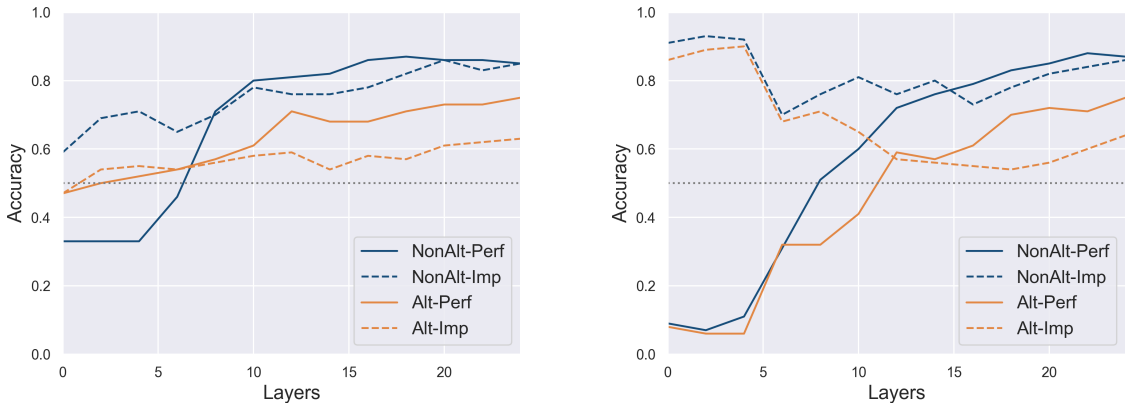

Figure 1: Performance of BERT-large on iterative masking (left) and aspect inference (right) for target verbs. *Perf* and *Imp* denote perfective and imperfective aspect in non-alternative (*NonAlt*) and alternative (*Alt*) contexts. Black dotted lines indicate random guessing between perfective and imperfective.

Then we feed the input sequence $\mathbf{X}$ to the model n times. On the first iteration, we replace all n target tokens of V with one [MASK] token to get $P(V_1|\mathbf{X} \setminus V)$. On each i-th iteration, $i > 1$, we feed $\mathbf{X}$ as input, with the tokens up to i, $V_1...V_{i-1}$ unmasked, and replace all remaining tokens $V_i...V_n$ with one [MASK]. We accumulate the probabilities $P(V_i|X \setminus V,\ V_1...V_{i-1})$ that the model assigns to each target token V_i. After the final n-th iteration, we calculate target verb's probability as the *average* of the accumulated conditional probabilities: $P(V) := \frac{1}{n}\sum_{i=1}^{n} P(V_i)$. We perform iterative masking for each instance twice: for the perfective and imperfective forms—and compare the two target probabilities.

We evaluate the model's performance by dropping one layer at a time. For BERT-large, Figure 1 (left) shows consistently higher performance for both aspect forms in non-alternative contexts on layers deeper than layer 15, with peak performance (85-88% accuracy) achieved in the final 8 layers of the model. BERT-base yields analogous results, although performance is slightly lower in the final layers, and its ability to predict both aspects steadily improves after layer 6 (Figure 7, Appendix B).

Performance on alternative contexts is significantly lower—since both aspect forms fit the context, the LMs show less preference for either aspect. Although we expect accuracy to be $\approx 50\%$ in alternative contexts, BERT picks the expected form more often. This may indicate the tendency of LMs to be more conservative when judging grammaticality (Prange and Wong, 2023). However, the probabilities assigned to the expected and complementary forms in alternative contexts are much closer together than in non-alternative contexts, particularly after layer 15 (see Figure 8 in Appendix B).

For RoBERTa-large, iterative masking shows significantly lower performance across all layers, the ability to differentiate between aspects is observed only after layer 18 in non-alternative contexts, see Figure 11 in Appendix B.

Aspect inference is a method based on verbs in the model's dictionary, which are *not* segmented into sub-words—call these *complete* verb forms. We feed the input sequence $\mathbf{X}$ to the model only once, replacing the target verb with a [MASK] token. We gather the top-k most probable tokens for the [MASK] position, and for each token we check whether it is also a complete verb form, with a known aspect. Then, we calculate aspect *preference*: e.g., preference for perfective aspect is given by:

$$P(\textit{perf}) = \sum_{i=1}^{k} 1_{\{\exists \text{ aspect = perf}\}} \cdot P(x_i)$$

$P(x_i)$ is the probability assigned by the model to a complete verb x_i. The parameter k is set to 10% (12K tokens) of the model's vocabulary. If most forms are perfective and have higher probabilities, we conclude that the model *systematically* prefers perfective in the target position.

As Figure 1 (right) shows, performance of BERT-large improves steadily for both aspects after layer 15 (after 8 for BERT-base). In the last 6 layers, aspect inference shows a similar performance to iterative masking (82-88% accuracy in non-alternative contexts). Our observations suggest that the capability to differentiate aspects develops after layers 12–14 for BERT-large (6–8 for BERT-base).

Setting k to 1% of the vocabulary size gives a similar performance, except for the first 2 layers

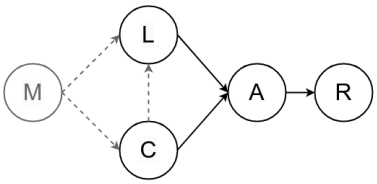

Figure 2: Causal model of dependencies between intended meaning (M) of instance, lemma of target verb (L), context (C), choice of aspect (A), and contextual representation (R) of target verb.

(Figure 10 in Appendix B). From layer 0 to layer 12 for BERT-large (0–8 for BERT-base), the model seems to favor one aspect over the other. To investigate this tendency, and whether predictions in early layers are conclusive, we inspected how many words out of the top-k for the masked position are complete verbs, for k =1.2K and k =12K. See further details in Appendix B.

For RoBERTa-large, the pattern of performance on aspect inference is similar to BERT-base, although in the early and middle layers the model seems to favor perfective first and then imperfective, unlike BERT. On layers 15–20, RoBERTa starts to differentiate between aspects. The last 5 layers perform similarly to the last 5 layers of BERT-base and BERT-large. As was noted by Belinkov (2022), the results of probing depend on the probed model, its original task, and the pre-training dataset. A much smaller vocabulary (50K for RoBERTa vs. 120K for BERT) and a different pre-training dataset can cause differences between the probed models, which requires further studies.

Aspect inference is similar to syntactic evaluations by Newman et al. (2021). These evaluations address the model's *systematicity* by conjugating a large set of verbs[6] and checking the model's *likely behavior*, by computing the probability that the models place on the correct form given the context. To get a higher score, the model must conjugate more verbs correctly, instead of only preferring some well-conjugated form. The authors show that neural models prefer to correctly conjugate verbs they deem likely in the target position.

4.3 Causal Probing

In the following experiments on causal probing, we continue with *aspect inference* to estimate the model's behavior, and with BERT-large, due to its superior performance in the final layers. We use all layers for causal probing. Although aspect inference is limited to verbs that are *complete* (unsegmented), this method gives a reliable assessment, since the percentage of complete verbs among the top-k predictions is high. Further, aspect prediction does not depend on the lemma of the verb, on its original form in the instance and its aspect pair, or on segmentation. It is also significantly faster than iterative masking.

We use a causal model of relations between the choice of aspect A and the intended meaning M conveyed in the context: M affects the choice of lemma L for the target verb and the choice of the surrounding context C (Figure 2). Since aspect is a grammatical category, we do not draw a direct connection between M and A. We focus on interventions into L and C: we will (1) remove the effect of lemmas by masking them and (2) alter the semantics of the context C by replacing the model's original representation with a counterfactual one. To generate counterfactual representations, we use the *AlterRep* method (Ravfogel et al., 2021). It is designed to study how the model uses a particular linguistic feature—by altering the representation of the studied feature, and investigating whether the resulting changes in the model's behavior agree with linguistic theory.

Boundedness: Aspect differs from previously studied syntactic phenomena—such as number agreement between subject and verb, etc.—for which it is easy to identify linguistic features that are directly involved in the phenomenon and can be used for causal probing. To probe aspect, we leverage the semantics of the context: in particular, how the meaning of *bounded vs. unbounded action* affects the choice of aspect. In Example (1), e.g., the imperfective verb form and the adverb "всегда" (*always*) conveyed that action is unbounded.

We identify cue words in the context—"Resultative", "Inception" words, etc.—which give cues regarding the boundedness of the action described by the verb and determine the choice of aspect.[7] The action is bounded if the target verb:

(1) has a "Resultative" adverbial modifier or argument, e.g.: "Внезапно она все поняла." (*Suddenly, she understood everything.*)

(2) has a "Duration" argument, e.g., "Она пробежала круг за 5 минут." (*She ran the lap in 5 minutes.*)

[6]The authors consider only *unsegmented* verb forms present in the models' vocabulary.

[7]The complete list of cue words is given in Appendix C.

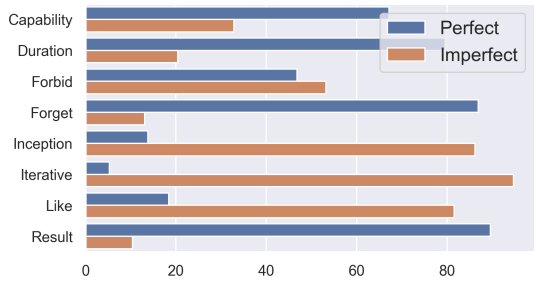

Figure 3: Percentage of sentences with detected cue words where target verb is perfective or imperfective.

(3) is a complement of a "Capability" verb, e.g.: ("Она смогла его понять.") (*She was able to understand him.*)

(4) is a complement of a "Forget" verb, e.g.: "Она забыла зарядить телефон." (*She forgot to charge the phone)*

The instance is unbounded if the target verb:

(1) is a complement of an "Inception" verb, e.g.: "Он начал петь." (*He began to sing*).

(2) has an "Iterative" adverbial modifier, e.g., "Гулять в лесу каждый вечер." (*To walk in the forest every evening.*)

(3) is a complement of a "Like" verb, e.g.: "Она любила читать." (*She liked to read.)*

INLP: Following Ravfogel et al. (2021), we denote by T a set of words in context, H the set of contextual representations of T, $\vec{h_t} \in \mathbb{R}^d$ the representation of word t. Let F be a linguistic feature encoded in H—here: *boundedness*. INLP defines the "feature sub-space" R of the original representation space where F is encoded. R is spanned by m learned directions—weight vectors of m linear classifiers trained to predict F given H, where all m are mutually orthogonal. Counterfactual representations $\vec{h_t^+}$ and $\vec{h_t^-}$ encode that the word t has positive or negative values of F, regardless of the true value of F encoded in the original $\vec{h_t}$. Counterfactuals are generated by pushing $\vec{h_t}$ further away in the opposite directions from the separating planes learned by m classifiers, see Appendix D.

We use *boundedness* as the feature F encoded in the representations H. F has two values: '+' if the action in the context is bounded, or '−' if unbounded. We train m SVM classifiers to define a sub-space of boundedness R and use it to manipulate the value of F in the target verb by generating counterfactual context representations $\vec{h_t^+}$ and $\vec{h_t^-}$.

Data: To train INLP classifiers, we need a dataset with pairs of contexts where the described action is bounded or unbounded. We collect instances automatically from the Omnia corpus and make sure that they do not appear in the test data. In a sentence parsed with a dependency parser (Burtsev et al., 2018), we pick the verb as a target only if it participates in syntactic relations with one or more cue words indicating boundedness.

We collect 8160 instances for each value of F. The choice of the types of relations was guided by grammatical rules, materials for language teaching (Kagan et al., 2014; Volkova and Phillips, 2015), and statistics derived from the SynTagRus corpus (Droganova et al., 2018), see Figure 3. In some constructions, both perfective and imperfective verb forms can be found.[8]

Training INLP: For every collected instance, we replace the target verb with a [MASK] token—to remove the influence from its lemma—and feed INLP classifiers with *contextual vectors of the cue words* from different BERT layers. Since the cue may be segmented into multiple sub-word tokens, we average the representation from the vectors of all cue segments. See training details in Appendix D.

Effect of Interventions: To assess the impact of counterfactuals, we measure the accuracy of aspect prediction using the aspect inference method. As in Ravfogel et al. (2021), for each sentence, we mask the target verb, start the forward pass, perform interventions on the verb representation at the specific layer, and continue the forward pass. Then, we retrieve the top-k tokens for the masked position and compute the model's preference for aspect. Figure 4 shows the results on the data used for behavioral probing, for non-alternative (top plots) vs. alternative contexts (bottom). The left plots display the results using negative counterfactuals—shifting representations toward unbounded action, and the right plots—positive counterfactuals, shifting toward bounded action. The X-axis indicates the layer at which the intervention is performed.

The most significant changes in the accuracy of predicting aspect in the masked position are seen in

[8] E.g., we can find examples of perfective verbs that depend on "Inception" verbs, but they are very infrequent in the corpus.

More complex is the situation with certain *ambiguous* cue words, e.g., нельзя (*impossible/forbidden*), which can appear equally with either aspect: its complement can be imperfective—"Здесь нельзя курить", (*Smoking is prohibited here)*, or perfective—"Нельзя закурить при сильном ветре", (*Impossible to smoke in strong wind).*

We exclude instances of such constructions from the data.

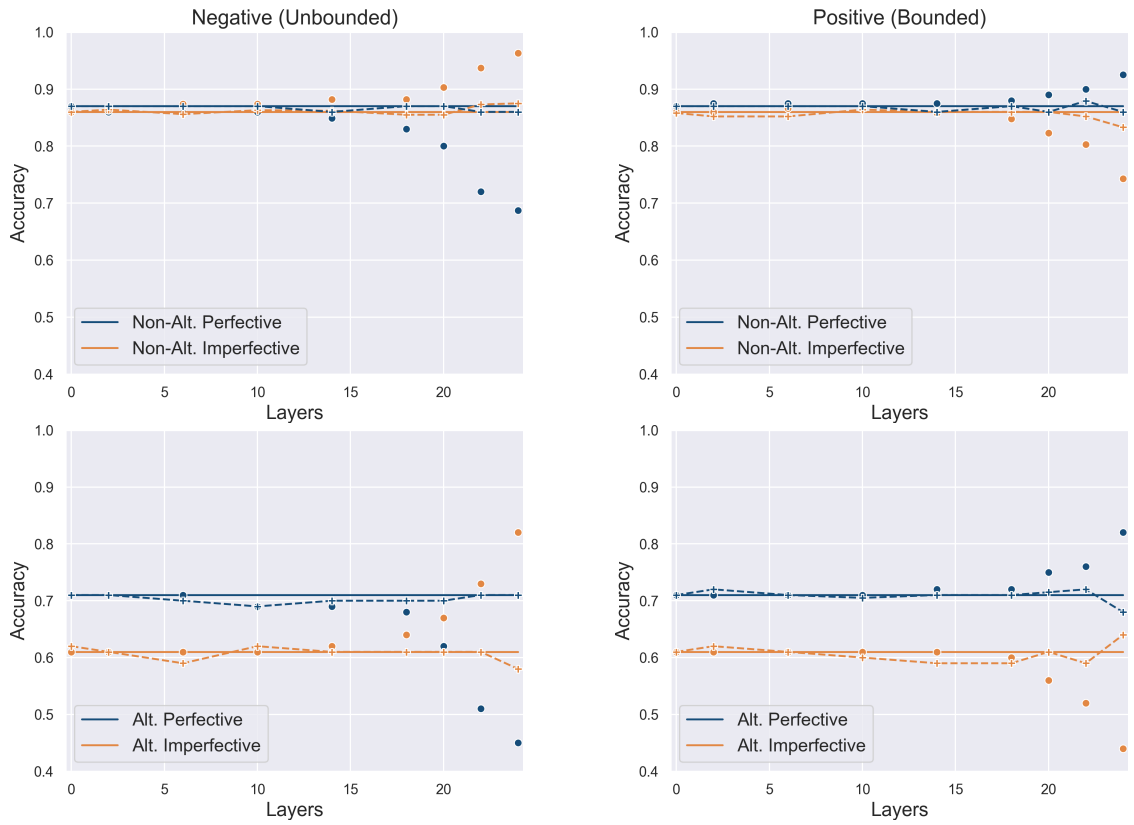

Figure 4: Accuracy of predicting correct (expected) aspect, using *aspect inference* method after intervention on BERT-large representations. Top plots—non-alternative contexts; bottom plots—alternative contexts. Left plots—negative intervention: toward unbounded action. Right plots—positive intervention: toward bounded action. **Flat** lines show performance before intervention; **dots**—after intervention. **Dashed** lines—after random interventions.

the model's latter layers (post layer 20)—compare the flat lines, indicating performance before interventions vs. dots, indicating an intervention. This trend is observed for both aspects, using negative and positive interventions in the alternative and non-alternative contexts. It agrees with the findings of behavioral probing, where the peak performance for both aspects was evident mostly in the final layers. The results align with our hypothesis and grammatical theory: shifting representations toward the "unbounded" sub-space improves the predictions of imperfective aspect and significantly increases the error in predictions of perfective aspect; moving representations in the opposite direction of the "bounded" sub-space has the opposite effect—the accuracy of perfective rises, while the accuracy of imperfective deteriorates.

Negative interventions influence imperfective in both alternative and non-alternative contexts: the maximum accuracy shift is $+21\%$ and $+10.3\%$, respectively, in layer 24. Similarly for positive interventions: the maximum accuracy shift for imperfective is -17% in alternative and -11.7% in non-alternative contexts. The impact of negative interventions on perfective is higher for alternative (-26%) and non-alternative contexts (-18.3%), as compared to the effect of positive interventions: in alternative contexts, perfective accuracy increases by 11%, and for non-alternative—by 5.5%. The plots show that interventions have stronger impacts in alternative contexts. Further, negative intervention has a stronger effect in both types of contexts.[9] This could be caused by the data used to train the INLP classifiers—cue words indicating unbounded action appear with imperfective verbs more consistently, see Figure 3.

We apply causal probing to RoBERTa-large and observe a similar pattern: only layers 18-24 are affected by interventions (Figure 14 in Appendix E). The influence of intervention is the same as for BERT. However, the difference between accuracy shift in alternative vs. non-alternative contexts is not as striking as for BERT.

Selectivity: To ensure the *selectivity* of the probe,

[9]We observe a similar patterns for BERT-base, see Figure 12 in Appendix E.

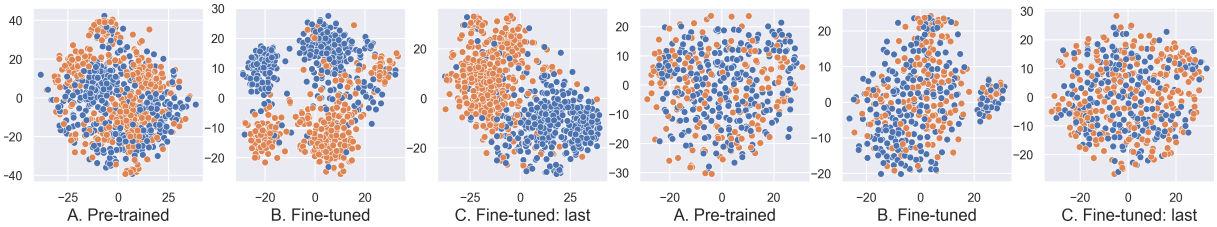

Figure 5: t-SNE visualization of representations from BERT-large layer 24 for masked target verbs in *non-alternative* (**left** 3 plots) and *alternative* (**right** 3 plots) contexts using A. pretrained models, B. fine-tuned models, and C. models with fine-tuned last layers. Orange indicates imperfective, and blue—perfective.

Model	non-alternative		alternative	
	$F_{0.5}^{\text{perf}}$	$F_{0.5}^{\text{imp}}$	$F_{0.5}^{\text{perf}}$	$F_{0.5}^{\text{imp}}$
A. Pretrained	36.3	49.2	54.0	51.1
B. Fine-tuned	85.9	84.0	67.5	57.0
C. Fine-tuned last 5 layers	**88.5**	**88.0**	69.1	64.2

Table 1: Performance in terms of $F_{0.5}$ for aspect prediction in non-alternative and alternative contexts.

we verify that random changes in representations do not impact aspect prediction in the same manner. Random counterfactuals were generated using 20 random sub-spaces. Dashed lines in Figure 4 show that changes in accuracy are smaller and do not follow the pattern observed with altering boundedness. Additionally, we ensure that the interventions targeting context semantics do not affect the predictions of other grammatical categories in the same way as they affect aspect. We perform the same experiment, but measure the accuracy of predicting the grammatical number of the masked target verbs. We choose number because it has no relation to aspect and frequently appears in verb forms. The results indicate no significant change in the prediction of number on any layer (Figure 13, Appendix F).

Probing with Iterative Masking: We checked whether causal probing shows similar results with different methods of evaluating the model's performance. Using iterative masking instead of aspect inference confirms the above observations. The main difference is the absolute value of the accuracy shift: it is in the range 2%–18% for the last layers of BERT-large.

4.4 Fine-tuning for Aspect Prediction

To utilize the information found through probing, we fine-tune BERT-large for the aspect prediction task. We formulate the task as a 2-way classification, where the model predicts whether the masked verb is perfective or imperfective. We use the SynTagRus corpus to create training and validation data.[10] Inspired by the probing results, we fine-tune layers 20-24 of the BERT encoder and the last classification layer, keeping all other layers frozen.

Table 1 shows the classification performance in three experiments: A. prior to fine-tuning; B. after fine-tuning all layers; and C. after fine-tuning the final 5 layers. Rows B–C show performance averaged across 5 fine-tuning runs. Freezing layers up to layer 20 speeds up fine-tuning and increases performance for aspect prediction, especially for imperfective aspect. Fine-tuning can yield performance comparable to the performance of BERT-large as a MLM at its final layers in non-alternative contexts. In alternative contexts, results are lower. Details on data, training, and evaluation with other layers are in Appendix G.

To visualize the changes in the model's representations, we use t-SNE (van der Maaten and Hinton, 2008) to project the masked verb representations onto the 2-D plane, Figure 5. Notably, fine-tuning the final layers results in more refined clustering of representations based on aspect. The lack of structure in the verb representations within alternative contexts aligns with our observations from the two behavioral probing methods—consistently lower preference for either aspect form.

4.5 Error Analysis

Uncertainty: The aspect inference method does not allow us to directly calculate the *uncertainty* of aspect prediction for a given instance. Therefore, we use Monte Carlo dropout (Gal and Ghahramani, 2016) to estimate the confidence of the fine-tuned model with frozen layers. For every input, we repeatedly sample 20 predictions with dropout activated, and calculate the variance; see plot (b) in Figure 6. The model has much higher predictive uncertainty for alternative contexts: BERT cannot

[10]The model is trained only on verb forms that have aspect tags in their morphological analysis.

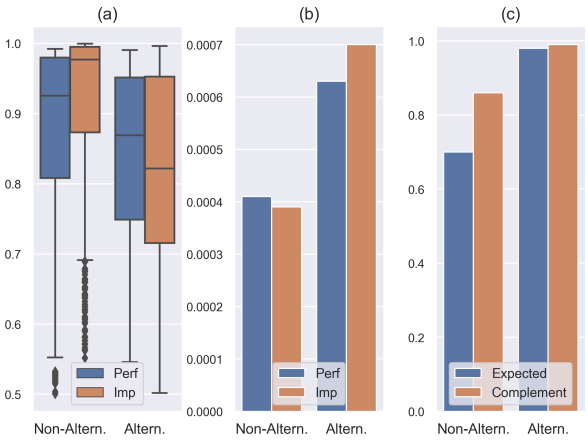

Figure 6: (a) Scores assigned to imperfective and perfective classes. (b) Variance of scores assigned to imperfective and perfective classes. (c) Percentage of contexts lacking cue words when the model predicts: *Expected* aspect vs. *Complementary* aspect.

make a preference for a particular aspect; see the orange bars indicating high variance.

We perform an automatic and a manual analysis of both types of contexts, to examine the possible reasons why BERT struggles with the aspect, either as a MLM or fine-tuned. We collect all contexts where BERT as MLM prefers the expected (blue bars in plot (c), Figure 6) vs. complementary aspect (orange). Preference is calculated using the aspect inference method. We calculated predictive uncertainty for the preferred aspect in each of these contexts. Then, we manually inspected the contexts with the highest variance. We observed the main difference—almost none of the alternative contexts contain cue words that could inform the preference for one aspect over the other.

Absence of Cues: Appendix C lists many cue words that indicate bounded vs. unbounded action. Although this list is not exhaustive, it provides a rough estimate of the difference between the contexts, and can aid in checking the manual analysis. We used this list to automatically inspect how many of the contexts do not contain cue words, and to check our manual evaluation. In Plot (c) the two left bars show that in non-alternative contexts, when BERT as a MLM predicts the complementary aspect (rather than the expected), most such contexts have no cues (orange bar, more than 85%). Of the contexts where the model predicted the aspect correctly, 70% (left blue bar) also have no cues, which indicates that other types of contextual evidence must be present in the context. This requires further study.

Almost 100% of alternative contexts have no cue words at all (plot (c), right two bars), which might explain why counterfactuals have more impact in alternative contexts—positive or negative interventions introduce the missing "hints" into the representations.

5 Conclusions and Future Work

We investigate the encoding of the grammatical category of verbal aspect for Russian in PLMs—particularly, BERT and RoBERTa—via behavioral and causal probing. Encoding of aspect has not been studied to date for any language or model. All types of probing show that these models do encode aspect and learn to distinguish between aspect forms primarily in their final layers. Using this finding, we fine-tune BERT for aspect prediction, which leads to more effective and faster tuning.

In line with linguistic theory, information about the *boundedness* of the action is encoded in the model's context representation and affects the choice of aspect: shifting representations towards the "bounded" space positively affects prediction of perfective forms (and negatively—of imperfective), and vice versa. Prediction of aspect is not affected by random interventions. We checked that the causal probe is selective and does not affect irrelevant categories, e.g., number.

A particular challenge is caused by contexts where more than one aspect form can fit grammatically and semantically, which we call *alternative*. We investigated whether encodings of aspect differ in these contexts from non-alternative ones. We find that BERT is consistently uncertain about aspect forms in alternative contexts. Causal interventions also have a stronger effect in such contexts. Our error analysis shows that these contexts do not have enough cues to help the model (or a human) decide which aspect to use.

In future work, we plan to explore additional languages; investigate how transformers encode relations between verbs and the cue words; and inspect the connections between aspect and tense of verbs, and context words expressing time, by intervening in the attention weights. We also plan to investigate aspect prediction in contexts lacking cue words, where information affecting the choice of aspect is presented in the neighboring sentences and requires reasoning. The practical goal is to deploy the aspect prediction in the production language teaching/learning system, to help learners master this advanced and complex feature of Russian.

Acknowledgements

This research was supported in part by Business-Finland Project "Revita" (Grant 42560/31/2020), and by a grant from the Helsinki Institute for Information Technology (HIIT).

Limitations

This work has a number of limitations to consider:

(A) The experimental design of the paper was limited to a single language. Aspectual systems vary significantly across languages. Therefore, adding a new language requires linguistic expertise and a new experimental setting. For Russian, we performed causal intervention in the context's meaning of boundedness and compared perfective vs. imperfective verb forms. Many languages do not have the opposition of these two forms as in Russian, and the meaning of boundedness may not be as significant for the choice of aspect in context. The closest aspect system to Russian among Slavic languages is Polish. Probing it would require a substantial investment of resources, which our team lacks.

(B) We experimented only with masked language models available for Russian since they have access to the full context, which is more relevant for the aspect prediction task.

(C) Due to resource constraints, we could not engage more people in data collection and annotation. While we recognize that our dataset is relatively small, we believe it is crucial to share the data we have. We hope it draws the research community's attention to the complex problem of aspect probing.

(D) We acknowledge that there is no consensus regarding several important questions among linguists studying the category of aspect in Slavic languages: the meaning of aspect opposition or whether aspect pairs represent forms of the same verb or different verbs. There are well-founded different opinions on each of these questions. We shape particular views for clarity of our experiments.

(E) We also recognize that our list of cue words, which indicate the boundedness of actions, is not exhaustive. We also ignore for now other contextual evidence indicating whether an action was completed, and whether its result is observable at the moment. Identifying this information is more complex and, we believe, requires reasoning. We plan to extend our work to investigate various types of contexts and larger PLMs.

(F) Due to the page limit, we did not include the effects of the removal of the linguistic feature of boundedness in the current experiment which could be an interesting extension of the experiment in the future versions of the paper.

(G) Our current experiments do not include an investigation of attention weights which we plan to do in future work.

Ethics Statement

We used publicly available data, code, and models for the described experiments.

Annotated data that we release together with this paper will be freely available for the research community to be used for extending probing experiments. Data does not have any personal information, does not identify individual people, and does not include offensive content. Annotators are volunteers who have previous experience in annotation and are aware of how the annotated data is going to be used. We also do not see any potential risk that might be caused by our work.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks. In *International Conference on Learning Representations*.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic Causal Probing for Morpho-Syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for Constituency Structure in Neural Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6738–6757, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eva G Bar-Shalom and Elena Zaretsky. 2008. Selective attrition in Russian-English bilingual children: Preservation of grammatical aspect. *International journal of bilingualism*, 12(4):281–302.

Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid

Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. In *Proceedings of ACL 2018, System Demonstrations*, pages 122–127, Melbourne, Australia. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2022. Probing for Predicate Argument Structures in Pretrained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Östen Dahl. 1985. *Tense and aspect systems*. Oxford.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

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.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian UD treebanks. In *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, volume 155, pages 53–66.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

James Forsyth. 1970. *A grammar of aspect: Usage and meaning in the Russian verb*. Cambridge University Press.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Atticus Geiger, Hanson Lu, Thomas F Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ester Hlavnova and Sebastian Ruder. 2023. Empowering cross-lingual behavioral testing of NLP models with typological features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7181–7198, Toronto, Canada. Association for Computational Linguistics.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. 2022. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1839–1851, Dublin, Ireland. Association for Computational Linguistics.

Olga E Kagan, Kudyma Anna, and Frank J Miller. 2014. *Russian: from intermediate to advanced*. Routledge.

Anisia Katinskaia, Maria Lebedeva, Jue Hou, and Roman Yangarber. 2022. Semi-automatically annotated learner corpus for Russian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 832–839, Marseille, France. European Language Resources Association.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. Revita: a system for language learning and supporting endangered languages. In *6th Workshop on NLP for CALL and 2nd Workshop on NLP for Research on Language Acquisition, at NoDaLiDa*, Gothenburg, Sweden.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Najoung Kim, Jatin Khilnani, Alex Warstadt, and Abdelrahim Qaddoumi. 2023. Reconstruction probing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8240–8255, Toronto, Canada. Association for Computational Linguistics.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, editors, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. Probing classifiers are unreliable for concept removal and detection. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022a. Does BERT really agree ? Fine-grained analysis of lexical dependence on a syntactic task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022b. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. 2023. Assessing the Capacity of Transformer to Abstract Syntactic Representations: A Contrastive Analysis Based on Long-distance Agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A Tale of a Probe and a Parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic

structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Jakob Prange and Man Ho Ivy Wong. 2023. Reanalyzing L2 preposition learning with Bayesian mixed effects and a pretrained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12722–12736, Toronto, Canada. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. 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 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Julia Rozanova, Marco Valentino, Lucas Cordeiro, and André Freitas. 2023. Interventional probing in high dimensions: An NLI case study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2489–2500, Dubrovnik, Croatia. Association for Computational Linguistics.

Tatiana Olegovna Shavrina and Vladimír Benko. 2019. Omnia russica: even larger Russian corpus. In *Corpus Linguistics-2019*, pages 94–102.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1393–1401, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, abs/2004.12265.

Victor Vinogradov. 1947. *Russkij yazyk. Grammaticheskoe uchenie o slove. [The Russian language. A grammatic theory of the word].* Uchpedgiz.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Natalija Volkova and Del Phillips. 2015. *Let's improve our Russian! Advanced Grammar Topics for English Speaking Students.* Zlatoust.

Zi Wang, Alexander Ku, Jason Baldridge, Thomas L Griffiths, and Been Kim. 2023. Gaussian Process Probes (GPP) for Uncertainty-Aware Probing. *arXiv preprint arXiv:2305.18213*.

Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *ArXiv*, abs/2007.06761.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? Probing pretrained

language models for the English comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Takateru Yamakoshi, James McClelland, Adele Goldberg, and Robert Hawkins. 2023. Causal interventions expose implicit situation models for commonsense language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13265–13293, Toronto, Canada. Association for Computational Linguistics.

Appendices

A Generating Aspect Test Data

For each target verb, we generated an aspect pair that differs only by the category of aspect. For this purpose, we used a list of over 2K verb lemmas with their aspect pairs which was manually created in collaboration with several linguists and Russian teaching experts. We generated an aspect pair for each target verb form automatically using a morphological generator which takes as input the verb lemma and a list of grammatical tags. For example, for a perfective verb form "получила" (*received*) in the past tense, singular number, and feminine gender, we generate an imperfective form "получала", which has the same tense, number, and gender.

> История получила шикорое освещение в газетах.
> *The story received extensive coverage in the newspapers.*

However, the paradigms of imperfective and perfective verb forms are not symmetrical. Perfective forms do not have present tense forms in the indicative mood, so we skip generation of an aspectual pair if the target verb is in imperfective present indicative form. There are no present tense forms for passive participles and transgressive forms, thus, in the context of this paper, we ignore participles and transgressives as targets.

There is also a difference between future tense forms for perfective and imperfective: imperfective verbs have analytic forms, e.g., compare "прочитает" (*will read*, perf.) and "будет читать" (*will read*, imp.). We generate aspect pairs for future tense taking this difference into account.

B Behavioral Probing

Iterative Masking Algorithm 1 demonstrates the process of iterative masking described in subsection 4.2.

Figure 8 shows boxplots with removed outliers displaying differences between probabilities assigned by BERT-large to two aspect forms (expected and complementary) in alternative contexts vs. differences between probabilities assigned to two aspect forms in non-alternative contexts. The probability of each form is calculated using iterative masking. Probability difference is calculated by subtracting the probability of the expected form from the probability of the complementary form: $P_{\{\text{exp.}\}} - P_{\{\text{compl.}\}}$.

Aspect Inference We inspected how many words out of the top-k filled by BERT-large in the [MASK] position are complete verbs, see plots for k =1.2K and k =12K in Figure 9. For the first 6 layers, the number of complete verb forms is low and most of them are imperfective, for any masked position. The model starts to predict perfective forms only after layer 4.

Considering that early layers incorporate less context information (Rogers et al., 2021), a higher preference for imperfective can be caused by frequency differences between aspect forms in the BERT's training data. Since the data used for pretraining is not available to us, we compared form frequencies in the SynTagRus corpus (Droganova et al., 2018). Imperfective is indeed more frequent (55% vs. 44%) in SynTagRus. However, these statistics characterize only one dataset. The frequency of aspect forms can depend on the genre of texts in the corpus. For example, legal texts usually have present tense more frequently than past or future. As a result, imperfective forms dominate legal texts because present tense forms in the indicative mood do not exist for perfective in Russian.

BERT-base Figure 7 shows the performance of BERT-base using iterative masking (left plot) and aspect inference (right plot) methods.

C Cue Words for Aspect

This section includes lists of lemmas of cue words that were used for collecting and annotating training data automatically for INLP classifiers. We excluded sentences where the target verb is negative since the negation particle "не" (*not*) in some

Algorithm 1 Iterative Masking

Input: Sequence of tokens $\boldsymbol{X}$ with target verb V; pre-segmented target verb $V = [V_1, \ldots, V_n]$ where $n \geq 1$.
for $i = 1$ to n **do**
  if $i = 1$ **then**
    Replace all V with [MASK] and feed $\boldsymbol{X}$ to BERT.
    Calculate $P(x^1|X \setminus V)$
  else
    Keep target segments $V_1, \ldots, V_{i-1}$ unmasked in $\boldsymbol{X}$.
    Replace $V_i ... V_n$ with one [MASK] and feed $\boldsymbol{X}$ to BERT
    Calculate $P(V_i|X \setminus V, V_1, \ldots, V_{i-1})$
  end if
end for
Get averaged probability of the target:
$P(V) \coloneqq \frac{1}{n} \sum_{i=1}^{n} P(V_i)$
Execute iterative masking twice for $\boldsymbol{X}$: with perfective and imperfective target verb forms
Compare $P(V_{\text{perf}})$ and $P(V_{\text{imp}})$

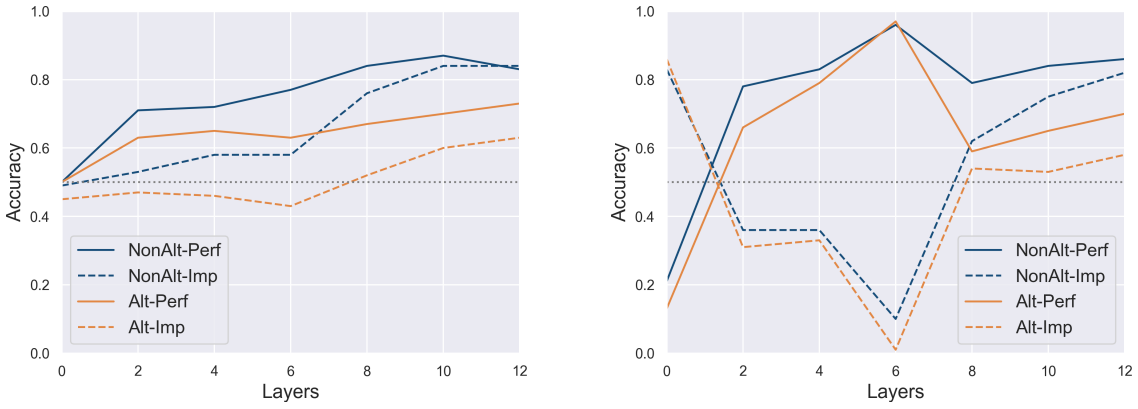

Figure 7: Performance of BERT-base on iterative masking (left) and aspect inference (right) for target verbs. *Perf* and *Imp* denote perfective and imperfective aspect in non-alternative (*NonAlt*) and alternative (*Alt*) contexts.

contexts causes a change of aspect from perfective to imperfective, e.g., in the imperative mood.

Lemmas in brackets denote that any word in the first list appears with any word from the second list, e.g., “каждый день” (*every day*) or “каждый год” (*every year*). There is also a possibility for these words to be interrupted by their own dependent words, e.g., “каждый новый год” (*every new year*).

“Forbid”: [запрещенный, дозволено, должен, надо, невозможно, нельзя, можно, нужно, обязан, опасный, рекомендуется, стоит]

“Iterative”: [бесконечно, бесперерывно, вечно, вновь, временами, всегда, часто, долго, изредка, непрерывно, как правило, постоянно, обычно, опять, регулярно, редко, систематически, снова], [[все, всякий, каждый, много, несколько, пара] + [век, весна, вечер, вторник, воскресенье, год, день, десятилетие, зима, лето, месяц, миг, минута, неделя, ночь, осень, раз, сезон, секунда, среда, суббота, сутки, период, полдня, полночи, понедельник, пятница, четверг, утро, час]], [[по] + [понедельник, вторник, среда, четверг, пятница, суббота, воскресенье, утро, вечер]]

“Duration”: [[за] + [век, весна, вечер, вторник, воскресенье, год, день, десятилетие, зима, лето, месяц, миг, минута, неделя, ночь, осень, раз, сезон, секунда, среда, суббота, сутки, период, полдня, полночи, понедельник, пятница, четверг, утро, час]]

“Inception”: [браться, бросать, бросить, давать, взяться, заканчивать, закончить, кончить, надоедать, надоесть, начать, начинать, оканчивать, окончить, отвыкать, отвыкнуть, передумать, передумывать, переставать, перестать, приниматься, приняться, продолжать, продолжить, раздумать, раздумывать,

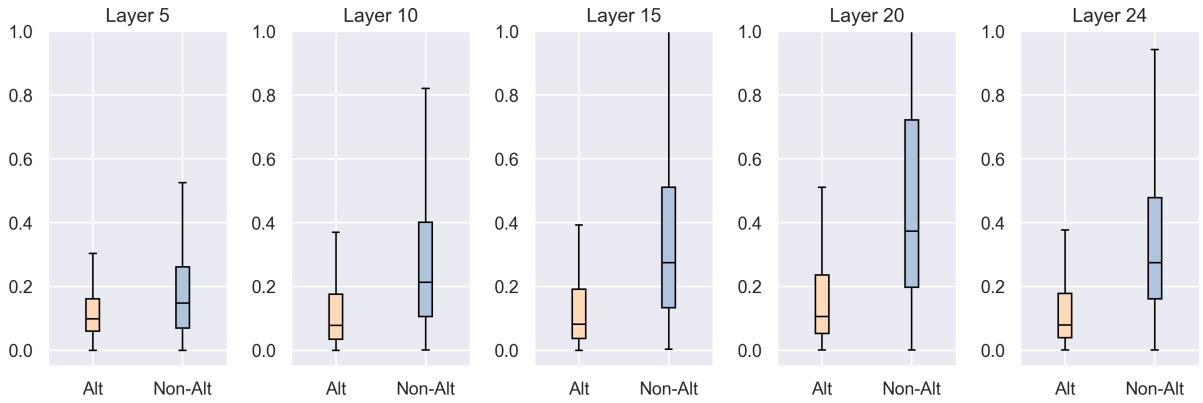

Figure 8: Difference between probabilities assigned by BERT-large to two aspect forms in alternative contexts (*Alt*) vs. differences assigned to aspect forms in non-alternative contexts (*Non-Alt*).

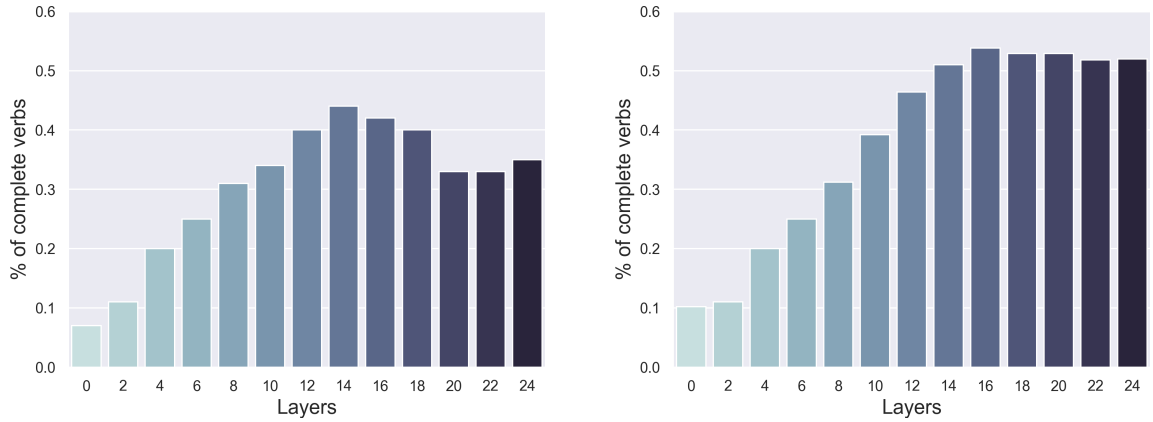

Figure 9: Percent of tokens that are complete valid verb forms among top-k tokens for the masked position, k = 12000 (10% of vocabulary size of BERT-large) on the left and k = 1200 (1% of vocabulary size of BERT-large) on the right.

разучиваться, разучиться, расхотеться, становиться, стать, уставать, устать]

"Like": [запрещать, запрещаться, избегать, любить, научить, научиться, нравиться, отговатьвать, привыкнуть, привыкать, следовать, уметь, учиться]

"Forget": [договариваться, договориться, забывать, забыть, обещать, согласиться, соглашаться, удасться, успевать, успеть]

"Capability": [мочь, смочь, способный]

"Result": [вдруг, внезапно, наконец, уже]

[[в] + [итог, конец, результат, финал]]

D Training INLP

We use SVM with stochastic gradient descent learning[11] as an INLP classifier and set the number of classifiers $m = 20$ and $\alpha = 4$ for BERT-large. Ravfogel et al. (2021) demonstrate that using different parameter values m and α does not substantially affect observed results. We use $m = 10$ for BERT-base. We increased m for BERT-large since its hidden representations are twice as big. Other parameters of the INLP classifier are: adaptive learning rate, early stopping set to True, and eta=0.1.

[11] sklearn.linear_model.SGDClassifier

E Effect of Counterfactuals on Aspect

Figure 7 shows the effect of counterfactual interventions on predicting aspect of the target verb using BERT-base. The effects of positive and negative interventions in both alternative and non-alternative contexts are similar to those observed using BERT-large. Interventions into the boundedness of action have a bigger impact on predicting aspect in alternative contexts. Positive and negative interventions affect aspect prediction in the last layers of BERT-base as well, predominantly after layer 8.

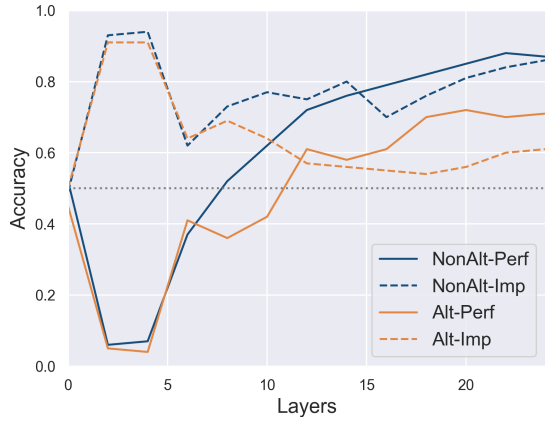

Figure 10: Performance of BERT-large on aspect inference for the target verbs with k = 1200. Perf and Imp denote perfective and imperfective aspect in non-alternative (NonAlt) and alternative (Alt) contexts

	non-alternative		alternative	
Model	$F_{0.5}^{perf}$	$F_{0.5}^{imp}$	$F_{0.5}^{perf}$	$F_{0.5}^{imp}$
Pretrained	36.3	49.2	54.0	51.1
Fine-tuned	85.9	84.0	67.5	57.0
Fine-tuned (up to last 5)	85.0	83.9	67.0	56.0
Fine-tuned (last 6)	87.0	88.0	69.0	64.0
Fine-tuned (last 5)	**88.5**	**88.0**	**69.1**	**64.2**
Fine-tuned (last 2)	87.0	87.0	67.0	60.0
Fine-tuned (last 1)	86.0	87.0	67.0	60.0

Table 2: Performance in terms of $F_{0.5}$ for imperfective and perfective aspects in non-alternative (non-alt.) and alternative (alt.) contexts.

F Affect of Boundedness on Category of Number

Figure 13 shows the effect of counterfactual interventions on predicting the number of the target verb. Interventions affect the meaning of the boundedness of the described action: whether the action is bounded or unbounded. Plots demonstrate that predicting the category of number is not affected by altering boundedness of the action, unlike aspect.

G Fine-tuning BERT for Aspect Prediction

Training data for fine-tuning BERT was generated using the SynTagRus corpus. For every sentence, we picked all verbs, labeled them with their aspect (`Perf` or `Imp` tag in the morphological analysis), and replaced them with a [MASK] token; all other words were labeled with `None`. Masking was used because the task is not to predict an aspect of a given verb form, but to predict which aspect fits in the given context. Also, during inference, we do not know which form should fit the context. We generated 60K training sentences and 7.5K validation sentences, where each sentence includes two masked verbs on average.

Parameters of training: learning rate = 5e-5, epochs = 3, batch size = 256, max input length = 512. The model was fine-tuned using 2 GPUs NVIDIA A100.

Testing was performed using the same data that we used for all probing tasks. The fine-tuned model is successful if the predicted label is the same as the expected aspect of the target verb. Table 1 and Table 2 report results averaged across 5 runs for each model configuration.

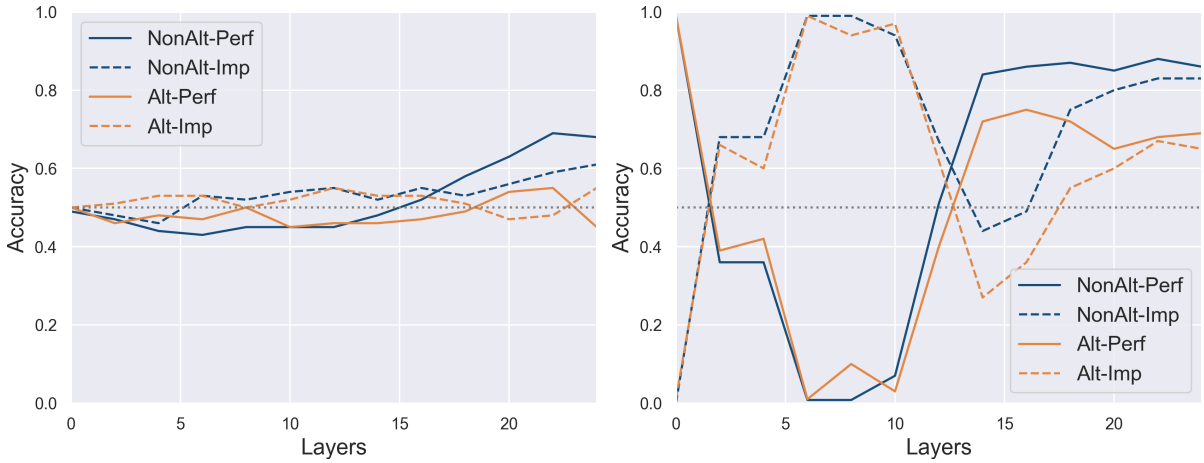

Figure 11: Performance of RoBERTa-large on iterative masking (left) and aspect inference (right) for target verbs. *Perf* and *Imp* denote perfective and imperfective aspect in non-alternative (*NonAlt*) and alternative (*Alt*) contexts.

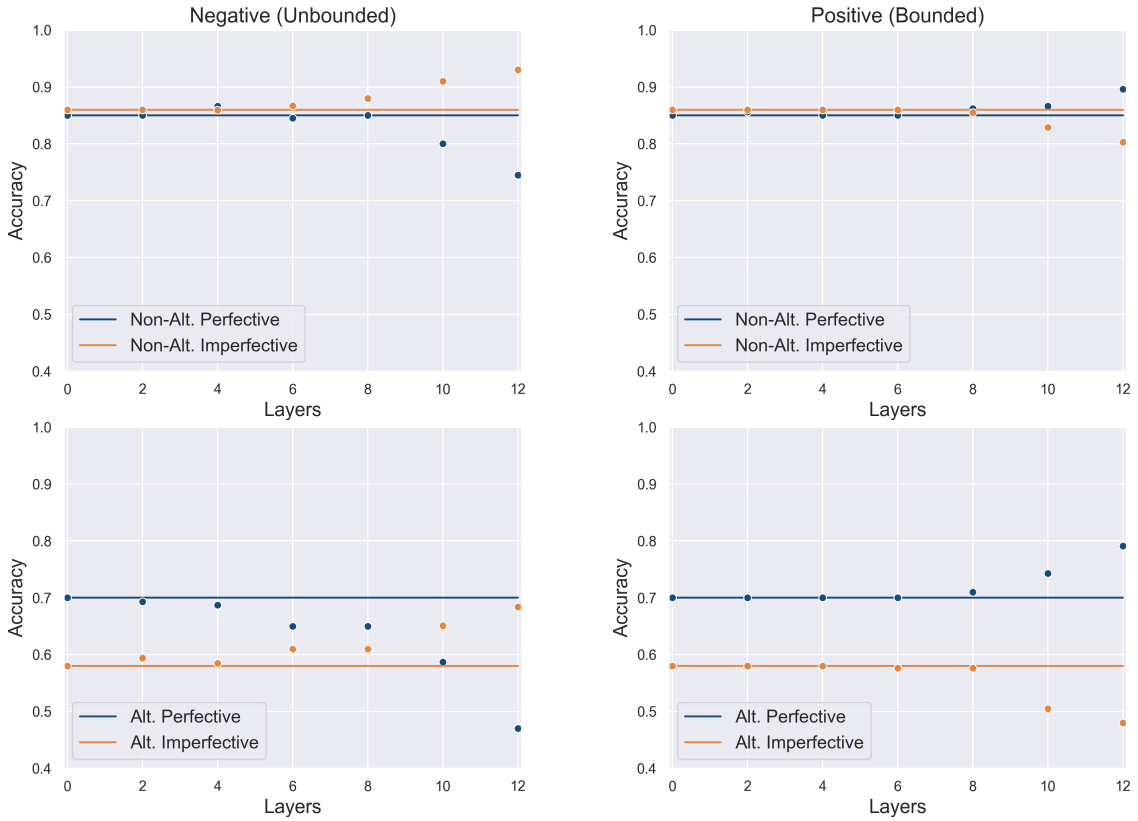

Figure 12: Change in accuracy of predicting correct (expected) aspect using aspect inference method after interventions on BERT-base representations. Top plots show results in non-alternative contexts; bottom plots—in alternative contexts. Left plots show the results of negative interventions: moving toward the meaning of unbounded action. Right plots—results of positive interventions: moving toward the meaning of bounded action. **Flat** lines indicate performance before interventions. **Dots**—after interventions. **Dashed**—after random interventions.

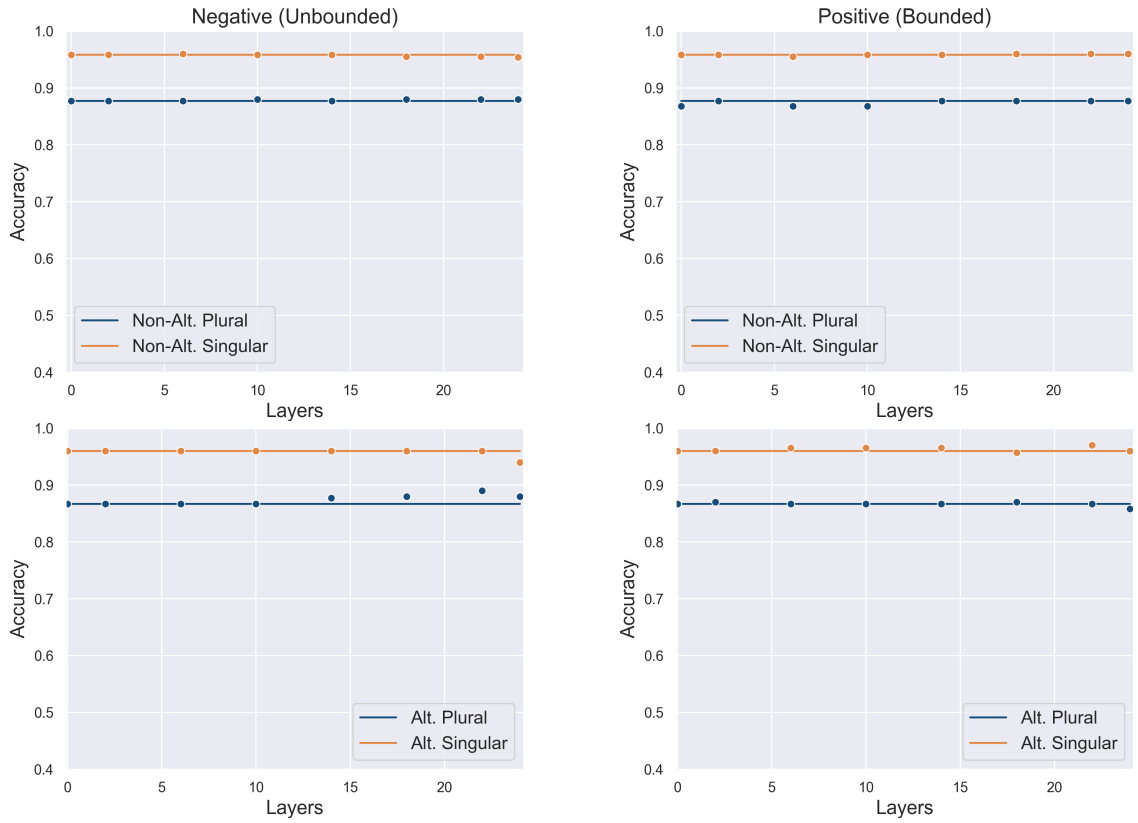

Figure 13: Change in accuracy of predicting correct number using the number inference after interventions on BERT-large representations. Top plots show results in non-alternative contexts; bottom plots—in alternative contexts. Left plots show the results of negative interventions: moving toward the meaning of unbounded action. Right plots—results of positive interventions: moving toward the meaning of bounded action. **Flat** lines indicate performance before interventions. **Dots**—after interventions.

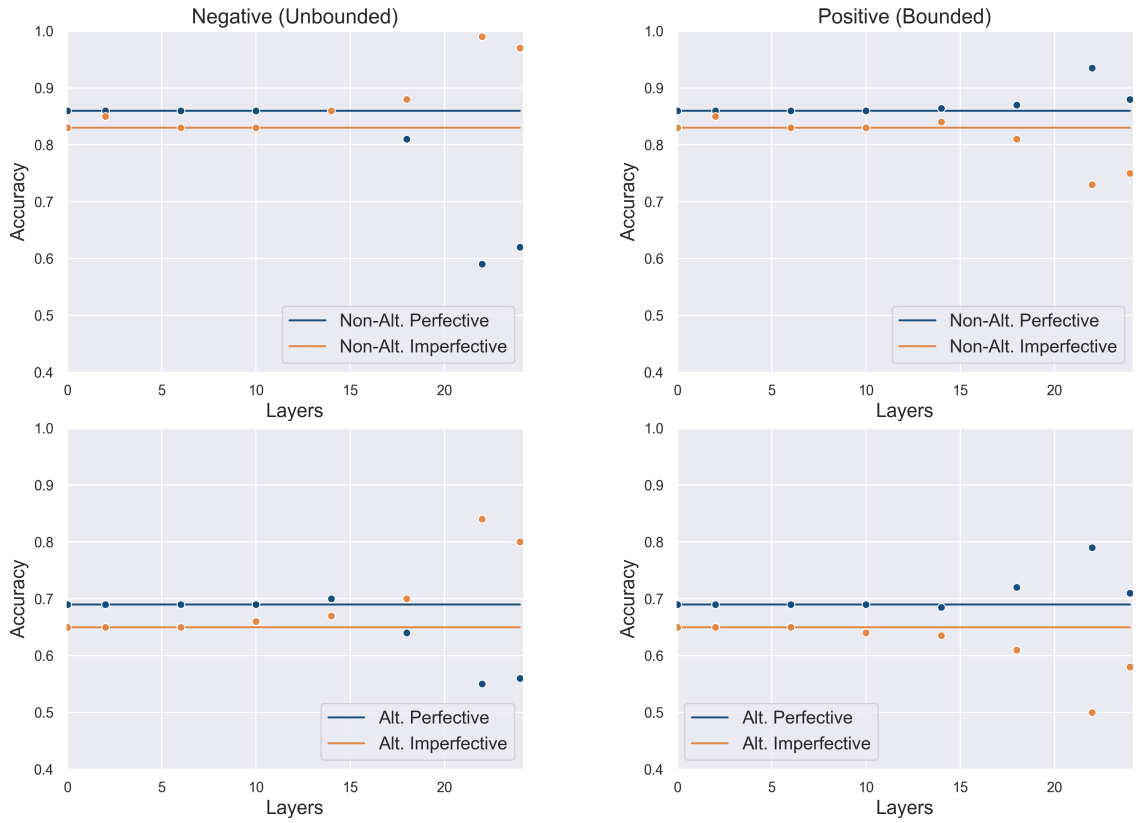

Figure 14: Change in accuracy of predicting correct (expected) aspect using the aspect inference after interventions on RoBERTa-large representations. Top plots show results in non-alternative contexts; bottom plots—in alternative contexts. Left plots show the results of negative interventions: moving toward the meaning of unbounded action. Right plots—results of positive interventions: moving toward the meaning of bounded action. **Flat** lines indicate performance before interventions. **Dots**—after interventions.