SyntaxGym for French: Resource, Annotation, and Evaluation of French and Multilingual LLMs

Tatiana Bladier, Henri-José Deulofeu, Alexis Nasr
Aix-Marseille Université, CNRS, LIS, UMR 7020
58, BD Charles Livon, 13284 Marseille, France
tatiana.bladier@lis-lab.fr, jose.deulofeu@univ-amu.fr, alexis.nasr@lis-lab.fr

Abstract

Despite recent advances in large language models (LLMs), their syntactic competence remains insufficiently characterized, especially for languages other than English. While benchmarks such as BLiMP and SyntaxGym have enabled systematic syntactic evaluation in English and Spanish, no comparable resource exists for French. To address this gap, we present SyntaxGymFR, a manually curated evaluation suite for evaluating the syntactic abilities of French and multilingual LLMs. SyntaxGymFR consists of manually validated minimal sentence pairs targeting key syntactic phenomena in French. We describe the annotation methodology, the selection of linguistic constructions, and the validation procedures used to ensure the coverage of syntactic phenomena. Furthermore, we report experimental results obtained with several French and multilingual LLMs, analyzing their sensitivity to grammatical contrasts and cross-linguistic transfer effects. Our results provide new insights into the syntactic generalization capabilities of French LLMs and establish SyntaxGymFR as a benchmark for future research on language-specific evaluation of syntactic competence.

Keywords: Syntactic evaluation, French, Large Language Models, SyntaxGym, corpus annotation, benchmarking

1. Introduction

Performance of Large Language Models (LLMs) on syntactic benchmarks has shown substantial differences in their grammatical competence (Marvin and Linzen, 2018; Hu et al., 2020; Wilcox et al., 2018; Sinha et al., 2022; Leong and Linzen, 2023; Arora et al., 2024). In particular, syntactic generalization, i.e. the ability to handle hierarchical dependencies, agreement, or structure-sensitive syntactic phenomena, is a core aspect of linguistic knowledge, and a key indicator of whether models capture more than surface-level token statistics. Evaluating the grammar competence is therefore essential both for understanding model behavior and for ensuring their reliability in real world applications where syntactic precision might potentially affect meaning and the quality of downstream applications.

While a number of targeted evaluation benchmarks have been developed for English, such as CoLA (Warstadt et al., 2019), BLiMP (Warstadt et al., 2020), and SyntaxGym (Gauthier et al., 2020), there is only a limited amount of syntactic evaluation resources for French. Existing French datasets mostly target acceptability judgments like QFrCola (Beauchemin and Khoury, 2025) or cover limited subject-verb agreement phenomena such as QFrBliMP (Beauchemin et al., 2025) and the French portion of MultiBliMP (Jumelet et al., 2025). These resources also do not provide the kind of psycholinguistically motivated experiments that enable systematic probing of syntactic competence on a broader range of linguistic phenomena. This lack of targeted syntactic benchmark corpora containing a broad variety of linguistic phenomena beyond simple agreement hinders our understanding whether existing French-aware LLMs have learned the syntactic regularities of the French grammar, such as long distance dependencies, licensing, or subordination, beyond their exposure to surface token distributions during training.

In this paper, we address this gap by introducing SyntaxGymFR[1], a new benchmark for targeted syntactic evaluation of French language models. SyntaxGymFR extends the experimental paradigm of the original SyntaxGym framework to the French language, offering a collection of controlled test suites that target specific syntactic phenomena. Following the terminology of the original framework, SyntaxGymFR is organized into five evaluation circuits (i.e. agreement, gross syntactic state, licensing, long-distance dependencies, and center embedding) each comprising multiple test suites. Each test suite contains minimal sentence pairs with a clearly defined region of interest, for which the model's surprisal is computed. This design enables a fine-grained analysis of how models encode syntactic expectations. SyntaxGymFR is designed for French, ensuring that each construction reflects authentic syntax and usage. Our contributions are the following:

[1] https://github.com/TaniaBladier/SyntaxGymFR

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2277–2287
11–16 May 2026.

Condition	intro	subj	rel_pron	np_obj	**verb1**	**verb2**	continuation
gramm.	Le	soldat	que	le messager	suit	court	.
ungramm.	Le	soldat	que	le messager	court	suit	.

Surpisal evaluation: (**gramm.verb1** + **gramm.verb2**) < (**ungramm.verb1** + **ungramm.verb2**)

Table 1: Example from the Center Embedding circuit in SyntaxGymFR. The two critical regions, *verb1* and *verb2*, marked in green, indicate the embedded verbs whose order determines sentence plausibility. The evaluation formula shows that the model's surprisal for the sum of these regions is expected to be lower for grammatical than for ungrammatical sentences. We sum the surprisal of both verbs to capture the full surprisal cost of the construction, following the standard evaluation format established in SyntaxGym as proposed by Gauthier et al. (2020).

- *Resource:* We introduce SyntaxGymFR, the first targeted psycholinguistically motivated syntactic evaluation benchmark for French, comprising a diverse set of phenomena including agreement, long distance dependencies, control, subordination, and polarity. Each sentence is annotated with POS tags, lemmas, morphological features, and syntactic labels in the CoNLL format following the Universal Dependencies (Nivre et al., 2016) guidelines.
- *Evaluation:* We present a comparative evaluation of several French and multilingual LLMs, ranging from autoregressive models to masked architectures. These experiments underline differences in how models recognize syntactic dependencies in French and demonstrate the value of SyntaxGymFR as a diagnostic tool for model comparison in a transparent and reproducible way.

2. Related Work

Evaluating the syntactic competence of large language models has become an important question in recent years, leading to a range of targeted evaluation frameworks. SyntaxGymFR continues this line of research but goes beyond previous work by providing a French-native, linguistically controlled, and reproducible test suite specifically designed to probe syntactic knowledge in French language models. In contrast to earlier benchmarks, most of which were developed for English or adapted automatically to other languages, SyntaxGymFR is hand-crafted by linguists to capture constructions genuinely characteristic of French syntax.

The SyntaxGym platform (Gauthier et al., 2020) introduced the paradigm of targeted syntactic evaluation for English. It formalizes psycholinguistic test suites, allowing to assess whether models show human-like expectations in controlled syntactic contexts. However, SyntaxGym is English-centric: its templates and linguistic contrasts are grounded in English syntax and do not generalize naturally to other languages. SyntaxGymFR builds directly on this framework but extends it to French grammar, implementing several new templates that reflect French-specific phenomena such as clitic placement, gender and number agreement, subjunctive mood alternations, and relativization patterns, none of which are covered in the original SyntaxGym.

Similar to SyntaxGym, the BLiMP (Warstadt et al., 2020) benchmark evaluate syntactic and morphological competence using sub-datasets of automatically generated minimal pairs. It has become a standard for probing grammatical knowledge in English LMs. Yet, its methodology, automatic generation from English grammatical rules, makes it unsuitable for French, whose inflectional morphology and word order constraints differ substantially. While BLiMP tests whether a model prefers grammatical over ungrammatical sentences, it does not model the incremental processing structure central to SyntaxGym and SyntaxGymFR. The latter, by contrast, evaluates surprisal differences over regions of interest within sentences, thus connecting syntactic evaluation with psycholinguistic modeling.

To our best knowledge, there is no resource comparable to SyntaxGym in French. However, there exists an extension of the SyntaxGym resource to Spanish, called SyntaxGymES (Pérez-Mayos et al., 2021). SyntaxGymES demonstrated that the methodology can be transferred to another Romance language, but its coverage remains only to Spanish. SyntaxGymFR is based on both English and Spanish resources, and contains adapted test suites to be informative for French syntax. Recent work has also introduced acceptability and grammaticality benchmarks for French, which differ in focus and methodology from SyntaxGymFR.

QFrCoLA (Beauchemin and Khoury, 2025) presents a large acceptability corpus for Québec French, modeled after the English CoLA (Warstadt et al., 2019). It evaluates whether models can classify sentences as acceptable or unacceptable, covering a broad range of grammatical and ungrammatical constructions. However, QFrCoLA is not designed to isolate specific syntactic phenomena: its evaluation is global, not targeted. By

contrast, SyntaxGymFR focuses on explicitly controlled syntactic manipulations, each test isolates a single constructional contrast, allowing for direct linguistic interpretation of model behavior.

Finally, the French portion of MultiBLiMP (Jumelet et al., 2025) and the Quebec-French QFrBliMP (Beauchemin et al., 2025) both extend the BLiMP framework to French, using automatic generation based on Universal Dependencies and UniMorph. While this large-scale approach ensures broad multilingual coverage, it often produces unnatural or unidiomatic examples and lacks manual validation by native speakers. As a result, the French subset provides a coarse evaluation of grammaticality but not a linguistically precise or psycholinguistically motivated one. SyntaxGymFR differs by offering manually validated test suites for French that reflect psycholinguistic experiments.

3. SyntaxGymFR

This section describes a selection of SyntaxGymFR test suites that we developed for French. They were constructed both from adapting the original English SyntaxGym (Wilcox et al., 2019) and some Spanish SyntaxGymES (Pérez-Mayos et al., 2021) test suites. Each circuit in SyntaxGymFR consists of a series of test suites that vary in a controlled way across CONDITIONS defined by the experimental design. For each test suite in each circuit, PREDICTIONS specify how surprisal values should differ across conditions at specific sentence regions. For example, Table 1 shows an item from the Center Embedding circuit in SyntaxGymFR, contrasting a grammatical and an ungrammatical sentence that differ only in the order of the two embedded verbs (represented in columns *verb1* and *verb2*).

The surprisal-based evaluation tests whether the model assigns lower surprisal to the grammatical configuration. In some cases, additional modifiers increase the distance between two co-varying elements, making the task more demanding. The test suites are grouped into circuits that correspond to related syntactic phenomena. Table 2 shows the comparison of SyntaxGym-resources for English (EN), Spanish (ES), and French (FR).

Each subsection below describes one such circuit. *Notation:* an asterisk (*) marks an ungrammatical sentence, and a question mark (?) marks a sentence that is more difficult to process.

3.1. Annotation

Our annotation process combined automated assistance with extensive expert revision to ensure linguistic precision. Grammatical and ungrammatical conditions were initially generated through targeted ChatGPT prompts[2] designed to produce SyntaxGym-style items differing only in the relevant syntactic property (e.g., agreement, extraction, LDD, etc.). However, these automatically generated materials served only as a starting point: all items were substantially rewritten, restructured, and carefully adjusted by trained linguists. In many cases, lexical choices, inflectional forms, or word order, were manually modified to guarantee minimal contrasts and produce correct sentences. Crucially, SyntaxGym items must follow a highly constrained structural scheme to ensure the correct evaluation of the target region. Each item was therefore manually inspected and adapted so that (i) the grammatical and ungrammatical conditions differed exclusively in the intended syntactic manipulation, (ii) the critical region was properly isolated, and (iii) non-targeted regions were controlled in accordance with the SyntaxGym evaluation protocol.

The vocabulary of the SyntaxGymFR items was restricted to 19th-century literary French, based on a corpus comprising five major authors of the period[3]. All items were manually validated by trained linguists to verify grammaticality judgments, lexical consistency, and the intended syntactic contrast. In addition, we enriched the final dataset with dependency parses produced by the state-of-the-art UD_FRENCH-SEQUOIA-FLAUBERT model of the

	Agr	GP	CE	GSS	Lcs	LDD	Total
EN	3	6	2	4	10	8	33
ES	7	2	2	3	4	3	21
FR	7	–	2	4	4	6	23

Table 2: Number of test suites in the different circuits across the English (EN), Spanish (ES), and French (FR) SyntaxGym-based corpora. Abbreviations: Agr = Agreement, GP = Garden-Path, CE = Center Embedding, GSS = Gross Syntactic State, Lcs = Licensing, LDD = Long-Distance Dependencies.

[2]Given that all items were extensively revised and validated through discussion among the annotators, we did not compute inter-annotator agreement. Potential disagreements were resolved during the annotation process itself. For transparency and reproducibility, we provide the prompt templates used for initial item generation in Appendix A.

[3]This variety of French was selected because it provides a rich vocabulary and a broad spectrum of inflectional forms. The data used in the experiments presented in this paper is released as a frozen version 1.0 on our GitHub repository https://github.com/TaniaBladier/SyntaxGymFR. Similarly to the original SyntaxGym, the French SyntaxGymFR is intended to grow as a living resource, and we encourage linguists to contribute by designing new test suites, which we will review and publish.

French-based parser HOPS[4] (Grobol and Crabbé), providing an additional syntactic layer for quality control and subsequent analytical tasks.

3.2. SyntaxGymFR Phenomena

3.2.1. Center Embedding

A *center-embedded* clause is a clause that is inserted within a main clause, creating nested dependencies that are challenging for models. This circuit contains two test suites:

- *Center Embedding*
- *Center Embedding with Modifier*

The test suites, adapted from English, introduce an object relative clause immediately after the subject of the main clause. In such a situation, the verbs of the main and the relative clauses appear one after another at the end of the sentence. The verb of the relative clause must be transitive since the relative clause is an object clause while the verb of the main clause must be intransitive. Switching the two verbs create an ungrammatical situation and we expect the model to give a lower perplexity to the grammatical case.

An example from the test suite *Center Embedding with Modifier* is shown below.

(1) Le soldat **que** le messager que mon frère connaît **suit** court.
'The soldier **whom** *the messenger whom my brother know* **follows** *runs'*
* Le soldat **que** le messager que mon frère connaît **court** suit.
'The soldier **whom** *the messenger whom my brother know* **runs** *follows'*

3.2.2. Long-Distance Dependencies (LDDs)

Long-distance dependencies arise when two syntactically related constituents are separated by intervening material. This circuit includes the following six test suites:

- *Filler-Gap Hierarchy*
- *Filler-Gap Object*
- *Filler-Gap Subject*
- *Filler-Gap 2 Sentence embeddings*
- *Cleft Constructions*
- *Cleft Constructions with Embedded Modifier*

[4]All dependency annotations generated by HOPS were systematically reviewed and manually corrected by trained linguists. For reference, the UD_FRENCH-SEQUOIA-FLAUBERT model reports an UAS of 99.31 and a LAS of 94.78 on the gold UD French-Sequoia test set.

An example from the test suite *filler–gap 2 sentence embeddings* is shown below.

(2) Je sais **ce** que notre intendante a affirmé que ton cousin **avait dissimulé** dans le grenier .
'I know **what** *our housekeeper said that your cousin* **has concealed** *in the attic. '*
* Je sais que notre intendante a affirmé que ton cousin avait dissimulé dans le grenier.
'I know our housekeeper said that your cousin has concealed in the attic. '

In the first sentence, the object of the verb *avait dissimulé* is separated from it governor by a declarative clause *notre intendante a affirmé que*, causing a long distance dependency. In the second sentence, the object of the transitive verb is not present, causing the ungrammaticality of the sentence. This pair allows to check whether the model captures the long distance object dependency or not.

3.2.3. Garden-Path Effects

Garden-path effects occur when a sentence is initially parsed in a plausible but incorrect way and must later be reanalyzed. The English version of SyntaxGym includes examples of such sentences. But garden path sentences are often unnatural in French and we decided to ignore this circuit.

3.2.4. Licensing

Licencing is a lexico-syntactic phenomenon where the presence of a given word in a sentence constrains the nature of another word of the sentence. This constraint could be lexical or morphological. This circuit includes four test suites:

- *Licensing Subjunctive with Wish Verbs*
- *Licensing Negative Polarity Items*.
- *Licensing Subjunctive Mood, Negation and Belief Verbs*.
- *Licensing Subjunctive Mood, Negation and Belief Verbs with an Embedded Modifier*.

In the following example, from the suite *Licensing Subjunctive with Wish Verbs*, the verb *veux* (*want*) licences the subjunctive mood of the verb of its object subordinate clause. When the verb of the subordinate clause *partes* (*leave.*SBJ) is in indicative mood, the sentence is ungrammatical.

(3) Je **veux** que tu **partes** demain.
I **want** *you to* **leave.SBJ** *tomorrow*
* Je **veux** que tu **pars** demain.
I **want** *you to* **leave.IND** *tomorrow*

3.2.5. Agreement

Agreement is a morpho-syntactic relation where the features of one element constrain the form of another. The agreement circuit had to be modified compared to English, since agreement phenomena in French differ substantially. This circuit contains the seven following test suites:

- *Agreement Subject–Verb in Number*
- *Agreement Subject–Verb in Person*
- *Agreement Subject–Participle in Number*
- *Agreement Subject–Participle in Gender*
- *Agreement Subject–Verb in Number with Embedded Subject Relative Clause*
- *Agreement Subject–Verb in Number with Embedded Object Relative Clause*
- *Agreement Adjective-Noun*.

In the following example, taken from the suite *Agreement Subject–Verb in Number with Embedded Subject Relative Clause*, the verb of the main clause *parla* (*talked*) agrees in number with the singular subject *la femme* (*the woman*). The object of the relative clause *les gendarmes* (*the constables*) is plural and precedes directly the verb of the main clause, acting as a distractor.

(4) **La femme** qui a aidé les gendarmes **parla**.
The women *who helped the constables* **talked.**SING.
* **La femme** qui a aidé les gendarmes **parlèrent**.
The women *who helped the constables* **talked.**PLUR.

3.2.6. Gross Syntactic State (GSS)

This circuit concerns complex sentences formed by an adjunct subordinate clause and a main clause. In such structures, the subordinate clause is introduced by a subordinating conjunction. The absence of this element causes ungrammaticality. Its presence when the main clause is absent, causes also ungrammaticality. This circuit is made of the following four test suites:

- *Simple Subordination*
- *Subordination with Object Relative Clause*
- *Subordination with Subject Relative Clause*
- *Subordination with a Prepositional Phrase*

The last three test suites are modified versions of the basic subordination test, in which a relative clause serves as the object or subject modifier of the subject of the main clause.

The following example is taken from the suite *Simple Subordination*. In the first sentence, the subordinating conjunction *comme* (*as*) is present and the main clause is absent. In the second one, the main clause is present and the sentence is grammatical.

(5) * **Comme** le médecin étudiait le manuscrit.
' **As** the doctor was studying the manuscript.'*
Comme le médecin étudiait le manuscrit, le chirurgien entra dans la pièce.
*'**As** the doctor was studying the manuscript, the surgeon entered the room.'*

In contrast to the previous example (5), example (6) presents two grammatical sentences. However, the version where the two clauses are connected by a conjunction is expected to yield lower surprisal than the juxtaposed version (marked with a ?), reflecting the model's sensitivity to explicit clause linking.

(6) ? Le médecin étudiait le manuscrit, le chirurgien entra dans la pièce.
'? The doctor was studying the manuscript, the surgeon entered the room.'
Comme le médecin étudiait le manuscrit, le chirurgien entra dans la pièce.
*'**As** the doctor was studying the manuscript, the surgeon entered the room.'*

4. Experiments

4.1. Experimental Setup

We evaluate the syntactic abilities of a range of French and multilingual language models using surprisal-based accuracy on grammatical/ungrammatical minimal pairs. Surprisal values are computed for critical regions in a sentence, allowing us to measure the sensitivity of models to syntactic violations. For each critical region in question we compute the sum of the surprisals of every subword in this regions, as is standard practice in previous works using SyntaxGym. Accuracy is assessed on minimal sentence pairs within the SyntaxGymFR items that differ only in grammaticality, providing a straightforward evaluation of the syntactic knowledge of models on specific syntactic phenomena.

We selected the language model to test based on two dimensions that we considered important. The first one is the language of the training data, we selected models that were either trained exclusively on French data and models that were multilingual, in order to evaluate whether models that were trained on French data perform better than multilingual models. The second dimension

is directionality of the model, we wanted to compare autoregressive and masked language model. The idea is to test whether the access to the right context when predicting a word (as it is the case in masked language model) yields better results. These two dimensions define four families of models: **French autoregressive**, M**ultilingual autoregressive**, **French masked** and **Multilingual masked**: (see Table 3 for the results):

- **French-native autoregressive LMs:** gpt2-base-french (ClassCat, 2023), gpt2-finetuned-oscar-fr (Yong, 2023), fr-boris (Müller and Laurent, 2022), gpt-fr-cased-small (Simoulin and Crabbé, 2021), Claire-7B-0.1 (Louradour et al., 2024), Pagnol-XL (Launay et al., 2022).
- **French masked LMs:** flaubert_large_cased (Le et al., 2020), camembert-base (Martin et al., 2020), distilcamembert-base (Delestre and Amar, 2022).
- **Multilingual autoregressive LMs:** Mistral-7B-v0.1 (Jiang et al., 2023), xglm-4.5B (Lin et al., 2021), Llama-3.1-8B (Meta, 2024), CroissantLLMBase (Faysse et al., 2024), bloom-560m (BigScience, 2022).
- **Multilingual masked LMs:** mdeberta-v3-base (He et al., 2021a,b), xlm-roberta-base (Conneau et al., 2019), swissbert (Vamvas et al., 2023).

Our evaluation focuses on region-level analyses, in particular the critical region, i.e., the token or span where grammatical contrasts occur. Metrics are reported both at the critical-region level and as a mean across all regions, allowing for a finer-grained assessment of model behavior. We aggregate the results per phenomenon to obtain overall performance scores for each syntactic construction. For autoregressive models, probabilities of the critical regions are naturally computed in a left-to-right generative fashion. For bidirectional masked models, we sequentially feed tokens with masking to compute probabilities given the preceding context, following Pérez-Mayos et al. (2021) and Wang and Cho (2019).

4.2. Results and Discussion

The results in Table 3 show clear differences between model families in their syntactic competence as evaluated on the five circuits of SyntaxGymFR. Among the French autoregressive models, Pagnol-xl achieved the highest overall accuracy (86.10%), consistently outperforming other models across nearly all circuits, particularly in agreement (76.25%), center embedding (91.07%), and gross syntactic state (95.83%). The Claire-7B-0.1 model reached a mean accuracy of 58.41%, with comparatively lower performance on gross syntactic state (38.02%). Interestingly, some of the multilingual models like bloom-560m, Llama-3.1-8B or Mistral-7B-v0.1, reached a mean accuracy of around 80%, demonstrating that multilingual models can attain strong syntactic sensitivity when trained with balanced cross-lingual data, though they remain below Pagnol-XL.

Masked language models displayed a more heterogeneous profile. Among French-native masked models, camembert-base reached the highest mean score (64.23%), driven by strong performance on agreement and licensing circuits, though accuracy dropped on other circuits. The multilingual xlm-roberta-base model achieved the best overall performance within its category (64.2%), with particularly high scores on agreement and gross syntactic state phenomena, outperforming several monolingual baselines. However, multilingual masked models overall fall behind French masked models (57.29% vs. 62.17%), indicating that while cross-lingual pretraining may improve robustness on certain syntactic phenomena, it does not consistently enhance syntactic precision for French. In sum, these results show both the benefits of language-specific pretraining on the one hand and the persistent variability of syntactic sensitivity across architectures and training paradigms on the other hand. It is important to note, however, that it is difficult to draw strong conclusions about potential advantages of causal versus masked architectures, or of monolingual versus multilingual pretraining, because each model differs in key parameters that strongly affect performance: the number of model parameters, the size and composition of the training dataset, and the computational budget used for training, effects collectively described by the *scaling laws* in Kaplan et al. (2020). These differences make it challenging to attribute syntactic performance to architecture or linguistic specialization alone. It is also interesting to study whether the different models and family of models struggle on the same circuits. In order to perform such an analysis, we have reproduced the structure of Table 3 in Table 4 and have replaced the accuracies figures by a rank between 1 and 5. Rank 1 in the cell corresponding to model M and circuit C indicates that circuit C got the best performances for model M over the five different circuit. For example, Licensing (Lcs) got the best performances for GPT2-base-french, Center Embedding (CE) the second best and so on.

Some family of models, such as French autoregressive show very coherent results: almost all the models of this family rank the five circuits in the same order. It is interesting to note that, on average, for all models except French causal ones, the hardest circuit is long distance dependencies. The fact that long distance dependencies are hard to process does not come as a surprise since the

Models	Agr	CE	GSS	LDD	Lcs	Avg. (per model)
Causal Models (French)						
gpt2-base-french	54.52	75.00	68.23	65.11	**82.29**	70.01
gpt2-finetuned-oscar-fr	38.07	58.93	40.63	80.73	44.79	45.60
fr-boris	74.52	62.50	36.46	60.42	65.63	59.77
gpt-fr-cased-small	65.72	55.36	46.88	**88.02**	57.29	56.31
Claire-7B-0.1	71.98	64.29	38.02	56.77	59.38	58.41
Pagnol-xl	**76.25**	**91.07**	**95.83**	80.73	81.25	**86.10**
Mean	63.51	67.86	54.34	71.96	65.10	62.70
Causal Models (Multilingual)						
xglm-4.5B	68.57	60.72	79.69	56.77	73.66	70.66
bloom-560m	**76.35**	69.65	**99.48**	69.79	75.00	80.12
Llama-3.1-8B	73.47	**87.95**	93.23	**84.38**	68.75	80.85
Mistral-7B-v0.1	72.53	76.79	96.36	67.71	**78.13**	**80.95**
Mean	72.73	73.77	92.19	69.66	73.88	78.14
Masked Models (French)						
camembert-base	78.29	50.00	57.81	52.61	**70.83**	**64.23**
flaubert_large_cased	77.80	**62.50**	48.44	55.21	63.54	58.07
distilcamembert-base	**78.47**	51.79	**64.07**	**66.15**	**62.50**	64.21
Mean	71.52	54.76	56.77	57.99	65.62	62.17
Masked Models (Multilingual)						
xlm-roberta-base	**67.38**	44.65	**82.29**	40.11	**62.50**	**64.20**
swissbert	65.34	48.22	55.73	**51.56**	53.13	55.60
mdeberta-v3-base	49.72	**55.36**	43.75	51.04	59.38	52.05
Mean	60.82	49.41	60.59	47.57	**58.33**	57.29

Table 3: Performance metrics for different models. The table reports the mean accuracy across all test suites within each circuit. The upper half presents results for autoregressive models (French monolingual in orange and multilingual in blue), while the lower half presents masked models (French monolingual in grey and multilingual in green). Note that evaluation procedures differ slightly: for autoregressive models, probabilities of the critical regions are computed in a left-to-right generative manner, whereas for bidirectional masked models, they are obtained by sequentially masking tokens and conditioning on the preceding context. Abbreviations: Agr = Agreement, CE = Center Embedding, GSS = Gross Syntactic State, Lcs = Licensing, LDD = Long-Distance Dependencies.

two words between which the dependency operates are separated by a lot of linguistic material in the string, which might contain strong distractors. Other phenomena, such as agreement between a noun and an adjective that modifies it concern words that are directly following each other in the string and the models, on average, perform well on this test suite (60.82% on average across all models). It is tempting to interpret all results on the light of this criterion: the linear distance between the words on which a linguistic phenomenon applies. From this point of view, the complexity of the syntactic structure does not play a major role compared to the linear distance between the words implied in a specific linguistic structure. The circuit center embedding seems to offer a counter example. Center embedding involves words that are distant from one another and is known to be a difficult phenomenon to process for humans (Miller and Isard, 1964). But the models tested here reach (on average) good performances on this phenomenon[5]. Although the distance between the two words involved in the phenomenon (the verb of the relative clause and the relative pronoun) is high, the phenomenon also has an effect on the order of the two verbs that appear at the end of the sentence (as for example verbs *'suit'* (*follows*) and *'court'* (*runs*) in the example of subsection 3.2.1). Our hypothesis is that the language model cap-

[5]This is mainly true for the autoregressive models.

Models	Agr	CE	GSS	LDD	Lcs
Causal Models (French)					
gpt2-base-french	5	2	3	4	1
gpt2-finetuned-oscar-fr	5	3	4	1	2
fr-boris	1	3	5	4	2
gpt-fr-cased-small	2	4	5	1	3
Claire-7B-0.1	1	2	5	4	3
Pagnol-xl	4	2	1	5	3
Rank of the mean (French autoregressive)	4	2	5	1	3
Causal Models (Multilingual)					
xglm-4.5B	3	4	1	5	2
bloom-560m	2	5	1	4	3
Llama-3.1-8B	3	2	1	4	5
Mistral-7B-v0.1	4	3	1	5	2
Rank of the mean (Multiling. autoregressive)	3	2	1	5	4
Masked Models (French)					
camembert-base	1	5	3	4	2
flaubert_large_cased	1	3	5	4	2
distilcamembert-base	1	5	3	2	4
Rank of the mean (French masked)	1	5	4	3	2
Masked Models (Multilingual)					
xlm-roberta-base	2	5	1	4	3
swissbert	2	4	1	5	3
mdeberta-v3-base	4	2	5	3	1
Rank of the mean (Multiling. masked)	1	4	2	5	3
Rank of the mean (all models)	2	4	1	5	3

Table 4: Rank of the different circuits (rank 1 is the circuit for which performances are the best) for each model. Rank of the average performances are indicated for family of models and across all models (last line).

tures the fact that the sequence of the two verbs *suit court* (*follows runs*) is more probable than the sequence *court suit* (*runs follows*) at the end of a sentence and this symptom helps the model discriminate between the grammatical and the ungrammatical sentence. It is as if the model was using a “cheap trick” to pass the test. More experiments are needed to confirm or disprove such an hypothesis.

5. Conclusion and Future Work

In this work, we introduced SyntaxGymFR, a manually curated benchmark for the targeted evaluation of French language models. Our contribution spans three key aspects: the creation of a syntactically rich resource covering phenomena such as agreement, long-distance dependencies, control, subordination, and licensing; a systematic annotation methodology ensuring psycholinguistically motivated minimal pairs; and benchmark results highlighting differences in grammatical competence across French-native and multilingual LLMs. By evaluating models at the region level and aggregating performance metrics per phenomenon, SyntaxGymFR provides a fine-grained analysis of syntactic expectations and surprisal patterns.

Looking forward, we plan to extend SyntaxGymFR to additional French-specific phenomena, including idiomatic expressions, word-order alternations, and constructions involving discourse-sensitive dependencies. We aim to release the resource as an open-source benchmark to facilitate comparative studies of French language models. These efforts make SyntaxGymFR a psycholinguistically grounded and interpretable way to evaluate the syntactic abilities of French-aware LLMs, making it useful for both linguistic research and practical NLP applications.

Acknowledgements

We would like to thank three anonymous reviewers for their valuable comments. The work presented in this paper has been funded by the French National Research Agency (ANR) as part of the COMPO project (grant ANR-23-CE23-0031). This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011016174R1 made by GENCI.

6. Bibliographical References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*.

BigScience. 2022. BigScience Language Opensscience Open-access Multilingual (BLOOM) Language Model. International collaboration, May 2021–May 2022. Available at https://huggingface.co/bigscience/bloom.

ClassCat. 2023. gpt2-base-french. https://huggingface.co/ClassCat/gpt2-base-french. Accessed: 2025-10-24.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Cyrile Delestre and Abibatou Amar. 2022. DistilCamemBERT : une distillation du modèle français CamemBERT. In *CAp (Conférence sur l'Apprentissage automatique)*, Vannes, France.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Croissantllm: A truly bilingual french-english language model.

Loïc Grobol and Benoît Crabbé. Analyse en dépendances du français avec des plongements contextualisés. In *Actes de la 28ème Conférence sur le Traitement Automatique des Langues Naturelles*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Julien Launay, E.l. Tommasone, Baptiste Pannier, François Boniface, Amélie Chatelain, Alessandro Cappelli, Iacopo Poli, and Djamé Seddah. 2022. PAGnol: An extra-large French generative model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France. European Language Resources Association.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Cara Su-Yi Leong and Tal Linzen. 2023. Language models can learn exceptions to syntactic rules.

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.

Jérôme Louradour, Julie Hunter, Ismaïl Harrando, Guokan Shang, Virgile Rennard, and Jean-

Pierre Lorré. 2024. Claire: Large language models for spontaneous french dialogue. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1: articles longs et prises de position*, pages 530–548.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Meta. 2024. Meta-Llama 3.1-8B: Model Card. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Hugging Face, accessed October 24, 2025.

George A Miller and Stephen Isard. 1964. Free recall of self-embedded english sentences. *Information and control*, 7(3):292–303.

Martin Müller and Florian Laurent. 2022. Cedille: A large autoregressive french language model. *ArXiv*, abs/2202.03371.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. *arXiv preprint arXiv:2105.04688*.

Antoine Simoulin and Benoit Crabbé. 2021. Un modèle Transformer Génératif Pré-entrainé pour le _____ français. In *Traitement Automatique des Langues Naturelles*, pages 246–255, Lille, France. ATALA.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2022. Language model acceptability judgements are not always robust to context.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The multilingual language model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019. What syntactic structures block dependencies in RNN language models? *arXiv preprint arXiv:1905.10431*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Zheng-Xin Yong. 2023. gpt2-base-french. https://huggingface.co/yongzx/gpt2-finetuned-oscar-fr. Accessed: 2025-10-24.

7. Language Resource References

David Beauchemin and Richard Khoury. 2025. Qfrcola: a quebec-french corpus of linguistic acceptability judgments. *arXiv preprint arXiv:2508.16867*.

David Beauchemin, Pier-Luc Veilleux, Richard Khoury, and Johanna-Pascale Roy. 2025. Qfrblimp: a quebec-french benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2509.25664*.

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

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

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.

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

Appendix A. LLM-Assisted Item Generation

Initial item drafts were generated using structured prompts designed to create SyntaxGym-style minimal pairs targeting specific syntactic phenomena (e.g., filler-gap dependencies, agreement, long-distance dependencies).

Prompt: Hierarchical Filler-Gap Dependency Generation

Instruction. Generate 20 items in SyntaxGym-style in French illustrating hierarchical filler-gap dependencies. Each sentence pair in item must contain one grammatical and one ungrammatical sentence.

Constraints:

1. Strict minimal-pair design (sentences identical except for the syntactic manipulation).
2. Exclusive use of 19th-century literary French vocabulary.
3. Identical lexical material across conditions except for the targeted syntactic contrast region.
4. Clear isolation of the critical region following the SyntaxGym evaluation scheme.

Structural template (illustrative English example).

* The fact that my brother said who his friend trusted our uncle at the party surprised my daughter yesterday afternoon.
- The fact that my brother said that his friend trusted our uncle at the party surprised my daughter yesterday afternoon.
- The fact that my brother said who his friend trusted at the party surprised my daughter yesterday afternoon.
* The fact that my brother said that his friend trusted at the party surprised my daughter yesterday afternoon.

The prompts explicitly specified the number of items, the target construction, the strict minimal pair format, and the lexical constraint that all sentences must use 19th-century literary French vocabulary drawn from our selected corpus of five major authors (Alexandre Dumas, Gustave Flaubert, Victor Hugo, Marcel Proust, Émile Zola) of that period. Generated outputs served only as preliminary drafts and were systematically rewritten, structurally adjusted, and linguistically validated by trained experts prior to inclusion in the dataset. Here we provide an example of a prompt used to generate hierarchical filler-gap dependency items in French. All generated French materials were subsequently reviewed, substantially rewritten where necessary, and brought into full compliance with the structural constraints required for controlled SyntaxGym evaluation.