Context Effects on the Interpretation of Complement Coercion: A Comparative Study with Language Models in Norwegian

Matteo Radaelli
Norwegian University of Science and Technology
matteo.radaelli@ntnu.no

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuele.chersoni@polyu.edu.hk

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Giosuè Baggio
Norwegian University of Science and Technology
giosue.baggio@ntnu.no

Abstract

In complement coercion sentences, like *John began the book*, a covert event (e.g., reading) may be recovered based on lexical meanings, world knowledge, and context. We investigate how context influences coercion interpretation performance for 17 language models (LMs) in Norwegian, a low-resource language. Our new dataset contained isolated coercion sentences (context-neutral), plus the same sentences with a subject NP that suggests a particular covert event and sentences that have a similar effect but that precede or follow the coercion sentence. LMs generally benefit from contextual enrichment, but performance varies depending on the model. Models that struggled in context-neutral sentences showed greater improvements from contextual enrichment. Subject NPs and pre-coercion sentences had the largest effect in facilitating coercion interpretation.

1 Introduction

Coercion results from a semantic type mismatch between a predicate and its argument (Pustejovsky, 1991, 1995; Jackendoff, 1997). In *John began the book*, the aspectual verb *begin* requires an event-denoting complement, but is instead combined with an entity-denoting NP (*the book*). The covert event can be recovered by exploiting the meaning of lexical constituents, world knowledge, and context (Pustejovsky, 1991, 1995; Lapata and Lascarides, 2003). Hence, speakers can interpret the sentence above as meaning, for example, that John began *reading* the book. Because the resulting interpretation is not a strict function of constituent meanings and syntax, coercion has been argued to violate strong versions of the principle of compositionality (Asher, 2015; Jackendoff, 1997; Baggio et al., 2012, 2016). Experiments found longer reading times (McElree et al., 2001; Traxler et al., 2002) and on-line processing costs (Pylkkänen and McElree, 2007; Baggio et al., 2010; Baggio, 2018) for coercion sentences compared to controls in which the relevant event is expressed by a non-aspectual verb (e.g., *John read the book*) or an event-denoting complement.

Transformer-based language models (LMs) (Vaswani et al., 2017) have become popular in NLP owing to their success in a range of tasks. However, few studies addressed how LMs process complement coercion. Previous research focused mainly on coercion as a challenge for sentence interpretation and framed it as a task where LMs have to predict the best covert event given an aspectual verb–complement combination (Rambelli et al., 2020; Ye et al., 2022; Gietz and Beekhuizen, 2022; Im and Lee, 2024; Rambelli et al., 2024). Radaelli et al. (2025) demonstrate that LMs have difficulty retrieving covert events for coercion sentences in Norwegian in the absence of context. The present study extends that work by investigating the role of context. Transformers' self-attention mechanism captures local contextual information by assigning greater relevance to some tokens compared to others within a sequence (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019). The result is the generation of dynamic linguistic representations that vary according to the surrounding context. We expect that contextual information will improve the performance of transformer-based LMs in covert event interpretation of coercion sentences.

2 Theories of Coercion in Context

One hypothesis assumes that a coercion interpretation is the result of *enriched composition*: lexico-semantic properties of words are leveraged to enrich the meaning of the sentence, resulting in an eventive reading (Pustejovsky, 1991, 1995, 1998; Asher, 2015). Each lexical item is associated with a *qualia structure* that includes, among others, a specification of TELIC (the purpose of an object)

Proceedings of the 16th International Conference on Computational Semantics, pages 78–88

and AGENTIVE (how an object is created) roles of the relevant entity. For coercion sentences, a type mismatch between the aspectual verb and its complement leads to the recovery of the covert event by exploiting the qualia roles of the entity *book*: the TELIC role implies that *reading* is the covert activity, while the AGENTIVE role implies *writing*. Contextual information can motivate different interpretations than those suggested by default qualia roles (Pustejovsky, 1995; Pustejovsky and Bouillon, 1995; Pustejovsky, 1998; Traxler et al., 2002). In *The author began the book*, the subject NP can facilitate the recovery of the AGENTIVE quale *write* (Traxler et al., 2005). In *The climber enjoyed the rock*, instead, where no specific TELIC role is provided by *rock*, the complement is enriched through co-composition of the subject NP *climber*, suggesting the interpretation that the climber enjoyed climbing the rock (Pustejovsky, 1998, p. 294).

The contextual enrichment of coercion sentences is also motivated by empirical studies. McElree et al. (2001, p. 7), for instance, acknowledge that the "properties of the subject NP appear to determine the default interpretation in an otherwise neutral context." In eye-tracking experiments, Traxler et al. (2005) concluded that contextual information does not necessarily attenuate processing costs in coercion sentences, but can be exploited as an 'extended lexicon', licensing an eventive interpretation that could be otherwise costly to generate.

The pragmatic hypothesis proposes a different account of complement coercion compared to the more constrained approach of the lexical analysis, which claims that coercion sentences are enriched solely via default qualia-based lexical information (Lascarides and Copestake, 1998; Zarcone et al., 2014). Building on relevance theory (Sperber and Wilson, 1986; Falkum, 2015), the proposal by De Almeida (2004) and De Almeida and Dwivedi (2008) claims that lexical entries only specify an expression's denotation or type (Fodor and Lepore, 1998). The interpretation of coercion sentences is therefore not lexically-driven but guided by post-lexical pragmatic inferences, world knowledge, and context. This leads to more flexible interpretations and a wider variety of readings compared to those afforded by qualia roles (Fodor and Lepore, 1998; De Almeida, 2004; Falkum, 2015).

Experimental work by Zarcone and Padó (2011) and Zarcone et al. (2014) provides instead support for Generalized Event Knowledge (GEK) (McRae and Matsuki, 2009) in coercion interpretation, an alternative to both lexical qualia-based and pragmatic hypotheses. The words-as-cues hypothesis (Elman, 2009) claims that speakers store event knowledge in memory: words serve as cues that allow access to such knowledge, modulating expectations about upcoming words. In a self-paced reading study, Zarcone et al. (2014) found that if coercion interpretations align with more typical events, sentences are read faster.

According to Piñango and Deo (2016), aspectual verbs do not necessarily trigger coercion effects when combined with entity-denoting complements, but can also specify mereological (i.e., part-whole) relationships between arguments (e.g., *The perch begins the trail*) as well as causal relations between events. In this theory, aspectual verbs select *structured individuals*, with parts ordered along a particular axis (e.g., spatial, temporal, informational etc.), formally a 'one-dimensional directed path structure' (DPS). Each argument encodes a set of functions that guide the mapping relative to a specific dimension. Both stative and eventive readings for sentences with aspectual verbs are possible. In the aspectually stative sentence *A thunderstorm began the day*, the predicate specifies the existence of a part-whole relation between the denotata of the complement and the subject. The information provided by the complement allows the predicate function to map the arguments onto a temporal dimension, interpreting the subject *thunderstorm* – a non-agentive entity – as denoting the initial temporal sub-interval of the denotation of the complement *morning* (Piñango and Deo, 2016, p. 369).

In sentences like *John began the book*, Piñango and Deo (2016) argue that the aspectual verb does not impose any type-selectional restrictions, hence no type-mismatch repair is needed. They propose instead a 'structured mapping' via inverse thematic functions. Because the event is underspecified, the traditional thematic function, which relates events to their participants, is not available. The inverse thematic function allows mapping of "pairs of individuals to the smallest event that the individual bears a participant role to at that time in a given context" (p. 385). Argument denotations and sentence context provide further constraints on the recovery of the event. Since complements are semantically undetermined and can map onto several possible axes, the same sentence can also be interpreted statively. If *John* is not interpreted as an agent, the

arguments would be mapped onto an informational dimension instead of an eventive one, and John would be considered one subpart of the informational object *the book*: John's work would then be an initial part of the book, such as a first chapter.

3 LMs and Complement Coercion

The first study evaluating LMs on complement coercion was Rambelli et al. (2020), who analyzed the performance of pretrained Transformers of the BERT and GPT families. They used datasets from different behavioral studies (McElree et al., 2001; Traxler et al., 2002; Lapata and Lascarides, 2003). The results revealed that Transformer-based models behaved differently from each other depending on the model's framework. ROBERTA, for example, emerged as the most robust LM, performing better than other models on the Lapata-Lascarides dataset (Lapata and Lascarides, 2003), with 80% accuracy in binary classification and 73% in a correlation task. In contrast, GPT-2 appeared to be more unstable, with a better score in the binary classification task (87%) but poorer performance in the correlation task (43%). Vanilla BERT, on the other hand, showed a marginal improvement over the baseline, suggesting a limited ability for contextualized embeddings in capturing eventive information from context. Finally, the authors report that distributional and non-Transformer frameworks, such as the Structured Distributional Model (SDM), performed similarly to ROBERTA despite being pretrained on smaller datasets.

Gietz and Beekhuizen (2022) consider coercion as a case of flexible semantic enrichment based on context, rather than as obligatory semantic completion. They analyzed a vanilla BERT model using a dataset with naturally-occurring coercion sentences from the COCA Corpus, successively annotated by humans. They argue that traditional 'hand-crafted' coercion sentences from previous studies always allow clear event interpretations, while naturally-occurring sentences usually include additional contextual information. BERT performed well in cases where consensus between annotators on a covert event was high, but struggled with sentences with less consensus. The model benefited from contextual information, improving event prediction.

Ye et al. (2022) used a dataset of naturally-occurring coercion sentences extracted from the C4 Corpus (Raffel et al., 2020). The authors argue that the process of coercion interpretation is analogous to paraphrasing: the coercion sentence is rephrased in a way that ambiguity is eliminated and the covert event is revealed. They found that pretrained BERT has difficulty with coercion interpretation, while a model fine-tuned with explicitly paraphrased sentences leads to better performance.

Radaelli et al. (2025) investigate whether LMs can leverage syntactic structure and lexical meaning toward recovering covert events. They conduct a large-scale evaluation of LMs in Norwegian, a low-resource language with variable grammatical realization of coercion, which partly depends on the aspectual verb used. Initiation verbs usually combine with entity-denoting NPs in PPs introduced either by *på* or by *med* (*John begynte på/med boken*; 'John began on/with the book'). With continuation and cessation verbs, complements are mainly introduced by *med*-PPs or directly an NP (*John avluttet med/ø boken*, 'John finished (with) the book'). Radaelli et al. (2025) released a new dataset of sentence pairs, each containing a context-neutral coercion sentence and an event resolution prompt. The dataset included 90 distinct entities from 6 different categories, and the syntactic realization of coercion was varied systematically by the aspectual verb and PP/NP. The study tested 17 Norwegian LMs, spanning BERT-like autoencoders and autoregressive models. In general, LMs struggled to recover implicit events. Surprisal estimates for whole sentences indicate that most LMs tested are unable to leverage the syntactic structure of the VP to interpret coercion items, showing no significant performance changes across syntactic constructions. For more details, see Section 5.1.

4 Task Proposal

Here, we explore the role of context in coercion interpretation, extending Radaelli et al. (2025)'s work on Norwegian context-neutral sentences. We study how different types of context influence the prediction distributions for LMs in a covert event interpretation task. We used the same evaluation strategy as Radaelli et al. (2025): instead of assessing models' performance only on a pre-defined set of top-1 ranked predictions as gold standard, we considered the *ranked prediction distribution* for each model; for each coercion sentence, a model must output a set of top-5 ranked predictions $O = o_1, ..., o_5$.

The distribution is then evaluated by calculating the mean average precision metric, which captures the consistency of LMs in predicting appropriate

events (see below) in the top ranking. We consider a model 'sensitive' to coercion, if it can provide a prediction distribution that is relevant to event interpretations: given a coercion sentence, expressed as a triplet ⟨*subject, aspectual verb, entity*⟩, we expect a redistribution of output predictions in a way that eventive interpretations are at the top of the ranking. The addition of contextual information should lead to further redistribution of the outputs, possibly with a shift towards the event interpretations suggested by the context.

The output predictions for each sentence will be evaluated by considering any event (verb) as correct as long as it satisfies the semantic constraints required by coercion and by the context. Following Piñango and Deo (2016) and Spalek and Sæbø (2019), a covert event is a plausible candidate for coercion when its combination with subject and complement expresses telicity, implying a "natural endpoint or goal state" that is coherent with the overall meaning of the sentence. The class of *accomplishments* is our ground truth for event classification, as it specifies durative, dynamic, and telic situation types or Aktionsart (Vendler, 1967; Spalek and Sæbø, 2019). All predicted events that are accomplishments are compositionally appropriate candidates, including those that may be weakly associated in coercion contexts. For example, the triplet ⟨*goat, begin, book*⟩ can suggest the covert event *eat* (Lascarides and Copestake, 1998). Some events must however be discarded: although they belong to the accomplishment class, their combination with the given subject and object results in a semantic anomaly. For example, a verb like *klatre* ('climb') could be plausible when predicted with objects that afford movement (e.g., *mur*; 'wall') but not with food items (e.g., *salat*; 'salad').

4.1 Dataset

We adopted a dataset originally created by Radaelli et al. (2025). Each item is a sentence pair designed to elicit the generation of covert events. Each pair includes (1) a context-neutral coercion sentence:

(1) {SUBJ} {VERB-FIN} {PREP|Ø} {ENTITY-DEF}.
E.g.: *Kim begynte på boken.* ('Kim began the book')

and (2) a sentence that prompts event retrieval:

(2) Det som {SUBJ} {VERB-FIN} å gjøre, var å [MASK].
'What {SUBJ} {VERB-FIN} to do was to [MASK]'.

The sentences contained the following elements:

- A single gender-neutral proper name (*Kim*) as subject {*SUBJ*}.
- 90 complement entity-denoting definite nouns {*ENTITY-DEF*}, consisting of real artifacts as incremental theme arguments of the implicit event. These entities belong to six different semantic categories: *food*, *text*, *clothing*, *everyday objects* (e.g., bag), *construction/housing* (e.g., wall), and *entertainment* (e.g., graffiti).
- Four aspectual verbs {*VERB-FIN*} in simple past form (*preteritum*), i.e., ***begynne*** (begin), ***starte*** (start), ***fortsette*** (continue), and ***avslutte*** (finish). Aspectual verbs, in contrast to other classes like psychological verbs (e.g., enjoy), were considered the only class of verbs that robustly trigger complement coercion, as shown experimentally by Katsika et al. (2012).
- Three complement syntactic constructions {*PREP—Ø*} introduced by a PP with either *på* or *med* or directly by an NP.
- The masked token [MASK] is included only for autoencoder models. With autoregressive models, [MASK] is replaced by blank tokens, used to prompt the prediction of the next sentence token.

We extended this dataset, here condition (a), by introducing three new conditions (b-d), each providing controlled contextual information in a specific portion of the experimental item (Table 1). The contextual enrichment applies only to sentence (1) in each pair, leaving (2) unchanged:

(a) Context-neutral: as in the original dataset;

(b) Subject-enriched context: the neutral subject (*Kim*) is replaced with a subject NP relevant for particular covert events;

(c) Post-verbal context: additional text is added after the entity complement as an adjunct or a coordinated phrase;

(d) Pre-coercion sentence: a sentence is concatenated before the coercion sentence, providing a discourse-level context.

All items in (1) included sentences with similar token length, with length variation of 2-3 tokens. Subjects and entity NPs were always in definite form,

(1) Coercion Sentence	(2) Prompt Sentence for Event Interpretation
(a) Kim begynte på essayet. (b) **Tolken** begynte på essayet. (c) Kim begynte på essayet **ved hjelp av ordboken**. (d) **Kim ønsket å publisere sitt nye verk på et annet språk for en fransk avis**. Kim begynte på essayet.	Det som Kim/tolken begynte å gjøre, var å ([MASK]).

Table 1: Examples of coercion sentences with the aspectual verb *å begynne* (to begin) in context conditions (a–d) in Norwegian and a common event-prompt interpretation sentence. Contextual information is presented in bold. Translations into English: (1a) 'Kim began the essay', (1b) 'The interpreter began the essay', (1c) 'Kim began the essay with the help of the dictionary', (1d) 'Kim wanted to publish his new work in a different language for a French newspaper. Kim began the essay', (2) 'What Kim/the interpreter began to do was to ([MASK])'.

while aspectual verbs were in *preteritum* form (past simple). The context was always coherent with the verb-complement combination.

For the assessment of models' performance we compared the results by Radaelli et al. (2025) with context-enriched conditions. The extended dataset includes a total of 4320 sentence pairs in standard written Norwegian Bokmål.

4.2 Tested Models

We tested the extended dataset on 17 pretrained Norwegian LMs, with autoencoders, such as BERT (Devlin et al., 2019), and autoregressive models, such as GPT-2 (Radford et al., 2019), LLAMA-2 (Touvron et al., 2023), BLOOM (Scao et al., 2022), and MISTRAL (Jiang et al., 2023). Table 2 shows the list of the language models tested here. The models differ considerably not only in architecture, but also in number of parameters and size of training data. Most LMs tested are monolingual models, only two (MBERT-CASED/UNCASED) are multilingual, while NORMISTRAL-7B-WARM was primarily pretrained in English and further trained in Norwegian. All tested models are available on Hugginface.[1]

4.3 Baseline Model

To assess performance between models and between different contextual conditions, we leveraged the same statistical baseline model as Radaelli et al. (2025): plausibility of event estimates were based on Pointwise Mutual Information (PMI) (Church and Hanks, 1990) between the verb and its object. The result is a list of (eventive) verbs strongly associated with an entity. These estimates are based on the Norwegian Colossal Corpus (NCC) (Kummervold et al., 2022), an open source corpus employed for training most current Norwegian LMs.

[1]https://huggingface.co/

Model	# Par.	Tr. Data
MBERT CASED/UNCASED	178M	3.3B*
NB-BERT-BASE	178M	7B
NB-BERT	355M	7B
NORBERT	111M	1.9B
NORBERT2	125M	15B
NORBERT3-BASE	123M	25B
NORBERT3-LARGE	353M	25B
NORBERT3-SMALL	40M	25B
NORBERT3-XS	15M	25B
NORBLOOM-7B-SCRATCH	7B	26.7B
NORGPT-369M	369M	25B
NORGPT-3B	3B	25B
NORGPT-3B-CONTINUE	3B	25B
NORLLAMA-3B	3B	26.7B
NORMISTRAL-7B-SCRATCH	7B	26.7B
NORMISTRAL-7B-WARM	7B	26.7B

Table 2: Tested LMs with number of parameters (*#Par.*) and training data (*Tr. Data*). *The amount of training data for MBERT is shared over 114 different languages.

4.4 Performance Evaluation

All prediction outputs provided by a given LM were manually classified by two of the authors according to Aktionsart, assessing the plausibility of the prediction in the coercion sentence. Disagreements were resolved through discussion. Predictions that were grammatically irrelevant to coercion sentences were discarded. We adopted two evaluation metrics for assessing models' performance. The first is mean average precision (mAP), which evaluates the ranking quality of a specific model based on the weighted means of average precision scores (AP) in the set of all sentences (S) (Manning et al., 2009; Kotlerman et al., 2010):

$$\text{mAP} = \frac{1}{S}\sum_{s=1}^{S}\text{AP}(s)$$

For any given sentence s, the AP score takes into account the ranking of the top-5 output predictions:

$$\text{AP(s)} = \sum_{k=1}^{5} P(k)\cdot\Delta R(k)$$

where $P(k)$ is the precision score at rank k and $\Delta R(k)$ is the recall difference between the current k and its antecedent $k-1$. A high mAP score indicates that the model tends to predict and rank accomplishments at the top. A low mAP score suggests either that the model proposes an event from an Aktionsart class other than accomplishments, or that the predicted accomplishment is ranked lower. The second metric is the mean top-ranked accuracy (A1) across the entire set of sentences (S). In this case, for each sentence, only the top-ranked prediction will be considered. Similar to the previous score, accomplishments count as the correct outputs, while other classes are false positives.

5 Results

5.1 General Results

Radaelli et al. (2025) found that LMs generally struggle to identify plausible events in context-neutral coercion sentences: mAP and A1 scores were low across LMs. Only few models exceeded the statistical baseline, and their performance varied mainly by model architecture and size. BERT-like models performed better than autoregressive models, with NORBERT3 showing relatively strong performance. Among autoregressive models, only NORLLAMA-3B and NORMISTRAL-7B-WARM exceeded the baseline. Model size also played a role: only the larger NORBERT3 variants could reach higher results, and autoregressive LMs like NORLLAMA-3B also showed decent performance, most likely due to their size.

Table 3 shows the mAP and A1 scores of all LMs tested on the covert event interpretation task in Norwegian. For comparison, we included the context-neutral scores from Radaelli et al. (2025). The results are available on GitHub. On the mAP scores, contextual information generally improved performance for most models compared to context-neutral sentences: 9 models outperformed the baseline, compared to only 4 with coercion-neutral sentences. However, even with context, the remaining 8 models still showed difficulties in consistently predicting appropriate events. Contextual information appears to particularly improve prediction for autoencoder models. Most models in the NORBERT family performed relatively well, reaching mAP scores above the baseline. Smaller models like NORBERT3-BASE, NORBERT3-SMALL, and NORBERT2, which showed poor performance in context-neutral sentences, here outperformed even the best

Model	mAP			A1		
	No Ctx	W/Ctx	Diff	No Ctx	W/Ctx	Diff
NCC (Baseline)	0.59	0.59	0.00	0.47	0.47	0.00
MBERT-CASED	0.07	0.07	0.00	0.00	0.01	0.01
MBERT-UNCASED	0.27	0.36	0.09	0.22	0.32	0.10
NORGPT-369M	0.56	0.62	0.06	0.54	0.57	0.03
NORGPT-3B	0.48	0.62	0.14	0.42	0.55	0.13
NORGPT-3B-CONT.	0.46	0.58	0.13	0.42	0.50	0.08
NORLLAMA-3B	0.71	0.66	-0.06	0.67	0.61	-0.06
NB-BERT-BASE	0.38	0.57	0.19	0.33	0.49	0.16
NB-BERT-LARGE	0.54	0.67	0.13	0.47	0.61	0.14
NORBERT	0.25	0.36	0.11	0.18	0.30	0.12
NORBERT2	0.44	0.69	0.24	0.34	0.62	0.28
NORBERT3-BASE	0.63	0.73	0.11	0.58	0.69	0.11
NORBERT3-LARGE	0.60	0.65	0.05	0.55	0.56	0.01
NORBERT3-SMALL	0.59	0.73	0.14	0.55	0.69	0.14
NORBERT3-XS	0.29	0.43	0.14	0.16	0.30	0.14
NORBLOOM-7B-S.	0.46	0.56	0.10	0.34	0.45	0.11
NORMISTRAL-7B-S.	0.38	0.58	0.19	0.29	0.49	0.20
NORMISTRAL-7B-W	0.63	0.64	0.01	0.54	0.56	0.02

Table 3: Comparison of mean average precision (mAP) and mean top-ranked accuracy (A1) for covert event retrieval in Norwegian context-neutral (No Ctx) and context-enriched (W/Ctx) sentences. Results for No Ctx are provided by Radaelli et al. (2025).

model NORLLAMA-3B in the context-neutral condition. NORBERT and NORBERT3-XS, on the other hand, still struggled with the task. Contextual information also improved performance of the NB-BERT family, namely LMs trained entirely on the NCC corpus, also used to create the statistical baseline model. While NB-BERT-LARGE achieved results above the baseline, NB-BERT-BASE still showed low performance despite the improvement.

A different pattern is found for autoregressive models. Most GPT-2 models still struggled to perform at or above the baseline. Only NORGPT-369M and NORGPT-3B benefited from the context, reaching reasonable results in mAP scores. NORBLOOM-7B-SCRATCH and NORMISTRAL-7B-SCRATCH still showed poor performance despite contextual enrichment, remaining below the baseline, while NORMISTRAL-7B-WARM did not improve relative to context-neutral sentences. Finally, NORLLAMA-3B is the only model that apparently suffers from the presence of context, showing a performance drop.

Analyzing the difference of mAP scores in sentences with and without context, we can appreciate how much context-enriched sentences enhanced the models' performance. First, context generally increases performance for most of those LMs that in the context-neutral condition struggled with coercion resolution. For example, NORBERT2, NB-BERT-BASE, and NORMISTRAL-7B-SCRATCH showed a significant improvement. MISTRAL and BERT-like models demonstrate the ability to exploit

context more effectively to improve performance while they struggled in the context-neutral condition, regardless of parameter sizes. GPT models also showed positive but weaker improvements, especially those with higher parameter sizes, such as NORGPT-3B and NORGPT-3B-CONTINUE. On the other hand, models that previously obtained relatively high mAP scores either did not show a significant change (e.g., NORMISTRAL-7B-WARM) or performed worse (e.g., NORLLAMA-3B).

A similar trend emerges from an analysis of A1 scores. NORBERT3-SMALL and NORBERT3-BASE reached the highest A1 score, close to 0.70. The other models showed considerably lower performance. Even the 10 models that outperformed the baseline obtained an A1 score ranging from 49 to 62, indicating that models still fail to top-rank accomplishments in approximately half of the cases.

A qualitative error analysis revealed that the addition of contextual information can sensibly affect model's performance. For example, comparing the subject-enriched sentence *Fienden begynte med testamentet* ('The enemy began with the will') to its neutral counterpart (*Kim begynte med testamentet*) on NORBERT3-BASE, we observed differences in the ranking. In the context-neutral case, the top-5 predictions were ⟨*skrive* ('write'), *lage* ('make'), *ta* ('take'), *gjøre* ('do'), *bruke* ('use')⟩, with the first two events being the only plausible accomplishments for coercion interpretation. In the context-enriched cases, the model kept the accomplishment (*skrive*) but prioritized verbs like *drepe* ('kill') and *stjele* ('steal'), indicating subject-driven biases. This means that, in this case, the replacement with a subject NP enriched with additional semantic information strongly shifts the prediction space of the model to events that are closely related to it. However, despite coherence with the subject, such outputs cannot be accepted: *drepe* requires an animate patient, while *stjele* lacks the durativity typical of accomplishments. Such events do not consider the contextual information conveyed by entire sentences, in particular the combination verb-entity. This suggests that contextual cues, especially those provided by the subject may strongly override the prediction ranking, guiding the model to predictions associated with those cues rather than by a compositional requirements.

Radaelli et al. (2025) conducted a quantitative error analysis with focus on the best performing model NORLLAMA-3B, examining the general frequency distribution of the predicted verbs across all context-neutral coercion sentences in the experiment. The analysis showed that the model produced a limited set of 68 unique verbs over 5,400 predictions, with the most frequent ones denoting either particularly generic events (e.g., *lage*, make, which combines with a wide range of entitites) or non-accomplishment verbs, here considered as false positives. In this study, we adopted the same analysis approach, by inspecting NORBERT3-BASE and comparing the event distribution across context-neutral and context-enriched coercion sentences. Table 4 shows the distribution of the first 15 most predicted events in all coercion sentences, comparing both context-neutral and context-enriched sentence conditions. The results suggest a similar trend to that found by Radaelli et al. (2025). First, also this model predicted a limited set of unique events, from 50 with context-neutral coercion sentences (among 5,400 predictions made in 1,080 sentences) increasing to 93 in context-enriched sentences (16,200 predictions in 3,240 sentences), suggesting that the addition of contextual information increases the variability of predicted events. Second, the ranking of predictions in both conditions is similar, following a skewed Zipfian distribution: the top ranked verbs dominate the distribution (covering up to 15% of the entire verb set), whereas predictions at lower positions show a sharp decrease of frequency. Finally, this analysis shows a minimal ranking variation in the distribution of verbs across

Verb	**No Ctx (Rel. Freq)**	**W/Ctx (Rel. Freq.)**
spille (play)	803 (0.15)	1,493 (0.09)
skrive (write)	781 (0.14)	1,924 (0.12)
le (laugh)	630 (0.12)	1,251 (0.08)
telle (count)	577 (0.11)	1,317 (0.08)
slå (hit)	524 (0.10)	1,455 (0.09)
danse (dance)	438 (0.08)	623 (0.04)
regne (calculate/rain)	414 (0.08)	1,117 (0.07)
vente (wait)	398 (0.07)	1,871 (0.12)
male (paint)	260 (0.05)	1,471 (0.09)
gå (go)	72 (0.01)	237 (0.01)
tale (speak)	65 (0.01)	392 (0.02)
lage (make)	56 (0.01)	644 (0.04)
holde (hold)	48 (0.01)	- (-)
rape (burp)	40 (0.01)	- (-)
bli (become / stay)	35 (0.01)	- (-)
sy (sew)	- (-)	316 (0.02)
hjelpe (help)	- (-)	233 (0.01)
bygge (build)	- (-)	185 (0.01)

Table 4: Top 15 events predicted by NORBERT3-SMALL across context-neutral (No Ctx) and context-enriched (W/Ctx) sentences, including both absolute and relative frequencies.

the two conditions, suggesting that context could not effectively elevate accomplishment verbs to the top rank, but influenced primarily the lower positions (e.g., *sy*, sew). Moreover, the most predicted events are in both conditions non-accomplishments, and therefore false positives for the classification task, usually denoting generic events not directly related to coercion resolution.

5.2 Context Types

We conducted further analyses of the impact of different context types on coercion sentences, with the conditions outlined in Section 4.1. For simplicity, we will consider only four models for this analysis: NORBERT3-SMALL, one of the top-performing models in this experiment, NORBERT2, which showed clear improvements compared to the context-neutral results by Radaelli et al. (2025), NORGPT-3B, the best performing GPT-based model, and NORLLAMA-3B, that showed instead a performance drop. Table 5 shows the models' mAP and A1 scores according to context conditions (b-d), including the context-neutral scores from Radaelli et al. (2025) as condition (a).

All context types led to improvements, with varying scores across conditions and LMs. Condition (d), the pre-coercion sentence, improved performance most, followed by condition (b), the context-enriched subject. Post-verbal context in condition (c) contributed the least among all conditions. A closer look at the scores reveals performance differences between LMs. First, NORBERT2 appears to benefit most when we consider the percentage increase over the mAP and A1 scores under contexts (b) and (d), with around 69% and over 90% improvement respectively compared to condition (a). This gap between the scores suggests that the model changed drastically the prediction distribution of verbs, ranking accomplishments at the top.

A more moderate performance improvement is found for both NORBERT3-SMALL and NORGPT-3B, which showed a similar behavior. On the one hand, their relative change against the baseline is small compared to NORBERT2, with a range between 25-33% for mAP scores and 28-39% for A1 scores. In this case, the gap between the mAP and A1 scores is minimal, meaning that the prediction distribution was more stable.

Finally, compared to the other models under test, NORLLAMA-3B shows the opposite trend in performance. Condition (c), the one that contributed least to the improvement, here leads to the largest decrease in performance, while a minimal positive improvement is observed for condition (b). Its A1 scores, however, remained unchanged under all conditions, showing only a minimal decrease.

Model	Cond.	mAP	A1
NORLLAMA-3B	a	0.713	0.670
NORLLAMA-3B	b	0.717 (0.004)	0.665 (-0.005)
NORLLAMA-3B	c	0.601 (-0.111)	0.547 (-0.123)
NORLLAMA-3B	d	0.653 (-0.060)	0.605 (-0.065)
NORBERT2	a	0.444	0.338
NORBERT2	b	0.718 (0.274)	0.650 (0.312)
NORBERT2	c	0.587 (0.143)	0.511 (0.173)
NORBERT2	d	0.753 (0.309)	0.708 (0.370)
NORBERT3-SMALL	a	0.593	0.545
NORBERT3-SMALL	b	0.747 (0.154)	0.699 (0.154)
NORBERT3-SMALL	c	0.676 (0.083)	0.608 (0.063)
NORBERT3-SMALL	d	0.776 (0.183)	0.762 (0.217)
NORGPT-3B	a	0.478	0.418
NORGPT-3B	b	0.625 (0.148)	0.536 (0.118)
NORGPT-3B	c	0.592 (0.115)	0.534 (0.116)
NORGPT-3B	d	0.639 (0.161)	0.571 (0.153)

Table 5: Mean average precision (mAP) and mean top-ranked accuracy (A1) for covert event retrieval in Norwegian across context-neutral (a) and context-enriched conditions (b-d). The results for (a) are from Radaelli et al. (2025).

6 General Discussion and Conclusion

Our results indicate that contextual enrichment of coercion sentences in Norwegian generally leads to better prediction distributions of covert events in almost all tested LMs. Additional context in specific sentence regions, such as in subject position, or the inclusion of sentences preceding the coercion construction, leads to most benefits in performance.

In this study, we found that performance varies also according to LM framework: BERT-like autoencoders appear to benefit most from contextual enrichment as compared to autoregressive models. This is consistent with the conclusion of Radaelli et al. (2025), where LMs were tested on coercion sentences without context. The advantage for autoencoders may be their bidirectional self-attention mechanism, which may be better able to capture semantic relations between constituents. However, models such as MBERT, NB-BERT-BASE, and NORBERT3-XS, for example, still showed only marginal improvements when exposed to context. Better performance for such models may be related to the interplay between their size and the amount of pretraining data: the multilingual model was one of the worst performing probably due to

its small training data in Norwegian. Conversely, results from NORBERT3-XS, suggest that, despite the large pretraining data, a smaller model still has limitations. Performance increases when the model's size increases, as shown for the larger NORBERT3 models. Other factors could also play a role. The NORBERT family showed more robust performance compared to NB-BERT models and MBERT, probably because the model was trained entirely from scratch on Norwegian and employed a custom WordPiece vocabulary. In contrast, NB-BERT starts from the MBERT framework and is trained on additional data in Norwegian without further changes (Kutuzov et al., 2021). Moreover, the third NORBERT generation, also introduces optimized training methods by excluding the next sentence prediction task and improving the masked language modeling objective task, increasing the span-based masking rather than masking single tokens (Samuel et al., 2023).

From the analysis of LM scores, we also found a consistent pattern linking their performance on context-neutral sentences and their improvement when context is introduced. Specifically, models that previously obtained poor results appear to benefit the most from context. Models like NORBERT2, NB-BERT-BASE, and NORBERT3-XS obtained a significant boost in performance compared to others. Such an improvement is however relative to their poor performance in context-neutral sentences. Their capacity to exploit contextual signals appears to compensate for such limitations.

It is particularly noteworthy that the LMs that obtained relatively high scores with context-neutral items are those that also showed more limited improvement when context is provided. This claim requires further research, but we hypothesize that such behavior may reflect a form of 'encoding saturation' by Transformer-based models, manifested in a limited capacity to integrate additional semantic information once a certain level of encoding complexity in a model's embedding-based representations has been reached. This behavior can also be observed when comparing models with almost identical architectures: NORBERT3-XS and NORBERT3-LARGE differ only in their parameter sizes, but they showed different improvement trends. We hypothesize that contextual information can compensate for gaps in world knowledge as required by coercion resolution. Consequently, context may not generally boost performance but rather benefits most the weaker models: stronger models show little change in their performance as they may have already reached a performance plateau, which cannot be improved by the integration of additional contextual information. This hypothesis is partially confirmed by the general improvement trend in results observed in Table 3.

Although contextual information generally led to better performance, LMs still show difficulties in interpreting complement coercion sentences. This aligns with the conclusions of earlier studies, such as Rambelli et al. (2020) and Ye et al. (2022) in English. It has been often observed that LMs lack a capacity for common sense reasoning based on plausible world models. This would also apply to natural language interpretation, in that current LMs have limited *linguistic common sense*: they lack the capacity to retrieve and exploit the kind of linguistic and world knowledge that would allow them to reliably make sense of complex, underspecified inputs (Lascarides and Copestake, 1998; Piñango and Deo, 2016; Baggio, 2018; Rambelli et al., 2024).

A closer look at our results sheds light on how and to what extent the behavior of Transformer models aligns with expectations based on different theoretical accounts on complement coercion. At first glance, the improvements seen for most models appear compatible with the pragmatic hypothesis: context and world knowledge can modulate or restrict coercion interpretations according to information that is not necessarily available from constituent meanings. However, such improvements were only seen for those models that were shown to be weaker in context-neutral scenarios, presumably due to a limited encoding of semantics in the learned embeddings. More semantically robust LMs were less influenced by context, suggesting that at least some relevant event information is encoded in the embeddings: this is more consistent with the lexical and Generalized Event Knowledge (GEK) hypotheses than with pragmatic accounts. On the other hand, our results cannot confirm the lexical hypothesis either, as context still has an effect in changing prediction distributions. Moreover, if models had learned and used lexically-bound representations such as qualia, we would not expect to see as outputs events that belong to incorrect Aktionsart, as in the example above. In addition, high performing models like NORLLAMA were even negatively influenced, suggesting a complex role of context in this task.

Acknowledgments

EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222). We thank the anonymous reviewers for their valuable suggestions.

References

Nicholas Asher. 2015. Types, meanings and coercions in lexical semantics. *Lingua*, 157:66–82.

Giosuè Baggio, Travis Choma, Michiel Van Lambalgen, and Peter Hagoort. 2010. Coercion and compositionality. *Journal of Cognitive Neuroscience*, 22(9):2131–2140.

Giosuè Baggio, Keith Stenning, and Michiel Van Lambalgen. 2016. Semantics and cognition. In Maria Aloni and Paul Dekker, editors, *The Cambridge Handbook of Formal Semantics*, pages 756–774. Cambridge University Press.

Giosuè Baggio, Michiel Van Lambalgen, and Peter Hagoort. 2012. The processing consequences of compositionality. In Markus Werning, Wolfram Hinzen, and Edouard Machery, editors, *The Oxford Handbook of Compositionality*, pages 655–672. Oxford University Press, Oxford.

Giosuè Baggio. 2018. *Meaning in the Brain*. MIT Press.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Roberto G De Almeida. 2004. The effect of context on the processing of type-shifting verbs. *Brain and language*, 90(1-3):249–261.

Roberto G De Almeida and Veena D Dwivedi. 2008. Coercion without lexical decomposition: Type-shifting effects revisited. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 53(2-3):301–326.

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*, pages 4171–4186.

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4):547–582.

Ingrid Lossius Falkum. 2015. The how and why of polysemy: A pragmatic account. *Lingua*, 157:83–99.

Jerry A Fodor and Ernie Lepore. 1998. The emptiness of the lexicon: Reflections on James Pustejovsky's The Generative Lexicon. *Linguistic Inquiry*, 29(2):269–288.

Frederick G Gietz and Barend Beekhuizen. 2022. Remodelling complement coercion interpretation. *Society for Computation in Linguistics*, 5(1).

Seohyun Im and Chungmin Lee. 2024. What gpt-4 knows about aspectual coercion: Focused on "begin the book". In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon@ LREC-COLING 2024*, pages 56–67.

Ray Jackendoff. 1997. *The Architecture of the Language Faculty*. MIT Press.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Argyro Katsika, David Braze, Ashwini Deo, and Maria Mercedes Piñango. 2012. Complement coercion: Distinguishing between type-shifting and pragmatic inferencing. *The Mental Lexicon*, 7(1):58–76.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-Scale Contextualised Language Modelling for Norwegian. ArXiv:2104.06546 [cs].

Maria Lapata and Alex Lascarides. 2003. A Probabilistic Account of Logical Metonymy. *Computational Linguistics*, 29(2):261–315.

Alex Lascarides and Ann Copestake. 1998. Pragmatics and word meaning. *Journal of Linguistics*, 34(2):387–414.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Brian McElree, Matthew J Traxler, Martin J Pickering, Rachel E Seely, and Ray Jackendoff. 2001. Reading time evidence for enriched composition. *Cognition*, 78(1):B17–B25.

Ken McRae and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and Linguistics Compass*, 3(6):1417–1429.

Maria Mercedes Piñango and Ashwini Deo. 2016. Reanalyzing the Complement Coercion Effect through a Generalized Lexical Semantics for Aspectual Verbs. *Journal of Semantics*, 33(2):359–408.

James Pustejovsky. 1991. The Generative Lexicon. *Computational Linguistics*, 17(4):409–441.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

James Pustejovsky. 1998. Generativity and explanation in semantics: A reply to Fodor and Lepore. *Linguistic Inquiry*, 29(2):289–311.

James Pustejovsky and Pierrette Bouillon. 1995. Aspectual Coercion and Logical Polysemy. *Journal of Semantics*, 12(2):133–162.

Liina Pylkkänen and Brian McElree. 2007. An MEG study of silent meaning. *Journal of Cognitive Neuroscience*, 19(11):1905–1921.

Matteo Radaelli, Emmanuele Chersoni, Alessandro Lenci, and Giosuè Baggio. 2025. Compositionality and Event Retrieval in Complement Coercion: A Study of Language Models in a Low-resource Setting. In *Proceedings of the 29th Conference on Computational Natural Language Learning (CoNLL)*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

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

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, Chu-Ren Huang, et al. 2020. Comparing Probabilistic, Distributional and Transformer-based Models on Logical Metonymy Interpretation. In *Proceedings of AACL-IJCNLP*.

Giulia Rambelli, Emmanuele Chersoni, Davide Testa, Philippe Blache, and Alessandro Lenci. 2024. Neural Generative Models and the Parallel Architecture of Language: A Critical Review and Outlook. *Topics in Cognitive Science*.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – A Benchmark for Norwegian Language Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Alexandra Anna Spalek and Kjell Johan Sæbø. 2019. To Finish in German and Mainland Scandinavian: Telicity and Incrementality. *Journal of Semantics*, 36(2):349–375.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Harvard University Press Cambridge, MA.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Matthew J Traxler, Brian McElree, Rihana S Williams, and Martin J Pickering. 2005. Context effects in coercion: Evidence from eye movements. *Journal of Memory and Language*, 53(1):1–25.

Matthew J Traxler, Martin J Pickering, and Brian McElree. 2002. Coercion in sentence processing: Evidence from eye-movements and self-paced reading. *Journal of Memory and Language*, 47(4):530–547.

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

Zeno Vendler. 1967. *Linguistics in Philosophy*. Cornell University, Ithaca, NY.

Bingyang Ye, Jingxuan Tu, Elisabetta Jezek, and James Pustejovsky. 2022. Interpreting logical metonymy through dense paraphrasing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Alessandra Zarcone and Sebastian Padó. 2011. Generalized event knowledge in logical metonymy resolution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Alessandra Zarcone, Sebastian Padó, and Alessandro Lenci. 2014. Logical Metonymy Resolution in a Words-as-Cues Framework: Evidence From Self-Paced Reading and Probe Recognition. *Cognitive Science*, 38(5):973–996.