Auto-AdvER: Disentangling Condition, History, and Commercial Offers in Automotive Advertisements 

Filippos Ventirozos[1,2], **Ioanna Nteka**[2], **Tania Nandy**[2], **Jozef Bača**[2], **Charlie Roadhouse**[1], **Peter A. Appleby**[2] and **Matthew Shardlow**[1]
[1] Manchester Metropolitan University, UK
[2] Autotrader Research Group, Autotrader UK
{f.ventirozos, m.shardlow}@mmu.ac.uk

Abstract

This paper presents a case study on the development of **Auto-AdvER**, a specialised span-labelling schema, along with its corresponding dataset for the textual car advertisement domain. Developed with industry needs in mind, **Auto-AdvER** is designed to enhance text mining analytics in this domain and contributes a linguistically unique sequence-labelling dataset[1]. We introduce a schema consisting of three labels: *'Condition'*, *'Historic'*, and *'Sales Options'*. We outline the guiding principles for annotation, describe the methodology for schema development, and present the results of an annotation study demonstrating an inter-annotator agreement of 78%. Furthermore, we compare the performance of various models, including transformer encoder models (BERT, DeBERTaV3) and transformer decoder-based open- and closed-source large language models: Llama, Qwen, GPT, and Gemini, as well as conditional random fields. Our results show that a fine-tuned DeBERTaV3 encoder achieves the highest F_1-score, followed closely by conditional random fields, while the large language models reach the highest precision but lower recall. We present this work as a stepping stone towards more fine-grained analysis and discuss **Auto-AdvER**'s potential impact on advertisement analytics, pricing, and customer insights. Potential applications include the analysis of market dynamics and data-driven predictive maintenance.

1 Introduction

The automotive industry relies heavily on accurate data about vehicles that are manufactured, sold, and resold. For digital businesses, much of this information is contained in free-text online car advertisements, where it is used for applications such as market analysis, competitive intelligence, and customer insights. Converting these descriptions into structured representations is therefore an important domain-specific information extraction problem. In this domain, the relevant text spans extend beyond classical named entities to include attribute-value and condition expressions. We therefore frame the task as span-based extraction from automobile advertisements, operationalised using a named entity recognition (NER)-style sequence-labelling approach (Li et al., 2023). Many of the spans we target (e.g., "smoke free", "tyre tread is low") are descriptive attributes or keyphrase-like mentions rather than names; we therefore refer to the annotated units as *spans* and to the three categories as *span labels*, and use NER terminology only for the modelling framework and for prior work. To support this task, we introduce Auto-AdvER (**Auto**mobile **Adv**ertisement **E**ntity **R**ecognition), a novel industry-developed annotation schema for vehicle-related spans in advertisement text.

Alongside the schema, we set out the principles we followed in forming its three labels (*Condition*, *Historic*, and *Sales Options*) and in deciding which words each label covers. We hope **Auto-AdvER** sets a standard for text annotation in the automotive industry and elsewhere. We make three contributions:

1. A novel domain span labelling dataset
2. The comparison of several applied span labelling models
3. Critical discussion of these results and applicability to the automotive industry and beyond.

The paper is structured as follows: Section 2 introduces related information extraction studies in the automotive domain. In Section 3, we describe our annotation guidelines and methodology, and in Section 3.4 we present an overview of the dataset. Section 4 presents our experiments and inter-annotator agreement (IAA) results. We discuss wider applicability and future work in Section 5, and conclude in Section 6.

[1]The annotation schema and dataset can be found at https://github.com/FilipposVentirozos/Auto-AdvER

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 603–619
November 6-10, 2026 ©2026 Association for Computational Linguistics

2 Related Work

Sequence labelling is a foundational modelling framework for information extraction in NLP: every token receives a label, and runs of labelled tokens form the extracted units. NER is its most prominent instance, with early datasets developed for the CoNLL evaluation campaigns (Tjong Kim Sang and De Meulder, 2003) focusing on general news text. NER involves identifying spans in text applicable for labelling and classifying them into predefined categories — a process often performed end-to-end (Jehangir et al., 2023). Typical approaches have employed sequence labelling models such as conditional random fields (Keraghel et al., 2024), BiLSTMs (Chiu and Nichols, 2016) or transformers (Hanh et al., 2021).

NER, in turn, has been extensively explored across various domains, often serving as the initial step in information extraction projects. Domain-specific datasets and models have been developed for biomedicine (Fu et al., 2023), agriculture (G. et al., 2023), and music (Epure and Hennequin, 2023). In each case, a bespoke annotation schema had to be developed to represent the units of interest in that domain. Our work follows the same path, but the units we annotate are descriptive spans rather than names, as discussed below.

There has been a growing body of literature in information extraction for the automotive domain. To the best of our knowledge, the earliest relevant work is by Lothritz et al. (2020), who evaluate NER models on the English/Turkish Wikipedia dataset of Sahin et al. (2017), whose articles span multiple domains, including the automotive domain. The annotation tags included automotive-specific entities such as fuel, engine, dates, manufacturers, models, generation and vehicle types. Park et al. (2023) introduced ADMit, which combines adversarial training and multi-task learning for automotive NER in Korean and English. Their work addresses domain adaptation between general-domain and automotive-specific terminology, but focuses on FAQ systems for automotive conversational AI assistants. Hu and Ma (2024) developed methods for detecting automotive parts/accessories in Chinese. FindVehicle (Guan et al., 2024) targets vehicle retrieval with entity types such as vehicle colour, brand, model, and location.

In contrast, our work focuses on extracting descriptive spans concerning a car's condition—e.g., scratches, malfunctioning components, tyre condition, and current paperwork status—as well as historical mentions and past events (e.g., accidents, component changes), and finally, commercial offers (e.g., warranties offered, part-exchange offers, delivery options), as specified in our annotation guidelines. Beyond advertisements, the same schema could be used for dealer listings, classifieds, and accident or insurance reports, enabling cross-regional analyses of commercial offerings and condition mentions for supporting safety and reliability monitoring.

3 Dataset

3.1 Annotation Methodology

To our knowledge, this is the first schema for labelling car text advertisements. Consequently, we developed our labels by working closely with our industry partner, Autotrader UK, and observing processes similar to the action research framework (Molineux, 2018), bridging industry and academia. Through our collaboration, we engaged extensively with industry data and resources, consulting with infrastructure engineers to gain an in-depth understanding of the automotive domain. This in-house expertise was pivotal in crafting our project's initial set of labels and guidelines. Below, we outline the core principles for developing our guidelines:

1. To construct the schema, we adopt a broad-category-first design: we delineate a small number of broad, semantically distinct categories and leave their refinement into finer-grained subtypes, and the linking of individual mentions to specific entities, to future work (Appendix B lists the sub-types we observed).
2. Each label was designed to be self-descriptive and unambiguous, such that a human annotator could reliably extract the correct spans by asking a single, distinct question per label (e.g., "What are the Condition mentions?") in order to minimise ambiguous overlapping labels.
3. The aim was to annotate only factual statements, not opinions or personal sentiments.

3.1.1 Iterations

Our methodology for developing the guidelines involved two separate teams working in iterations: a university-based team (two members) and an industry-based team (three members). The university team would create or edit the classes and annotation guidelines. Then, these would subsequently be discussed with the industry-based team,

which comprised domain-expert data scientists. We undertook ten iterations of annotation, with feedback as detailed in the following two sections.

Initial Iterations During the first set of iterations, the university team proposed a draft of the guidelines. These proposals were then refined by the industry team, focusing on two main aspects:

1. The inclusivity of each class and the extent to which they covered various aspects of the advertisements.
2. The identification of any omitted classes, which could either be sourced from other data or deemed non-essential.

Following these discussions, the university team further refined the classes. When a class lacked clarity regarding its encompassed aspects, the team evaluated the semantic coherence of each aspect within the proposed definitions. Rapid prototyping enabled them to explore alternative classes and assess their detectability.

Secondary Iterations As the labels began to take shape, we entered the second set of iterations. Here, the university team introduced annotation tasks for the industry team. A customised Prodigy by spaCy (Montani and Honnibal) NER annotation interface was used for annotating; more information can be found in Appendix A.

After the industry team completed these tasks, their feedback was discussed to further refine the guidelines. This feedback included analysis of challenging cases and observations on whether additional aspects should be incorporated. Once the revised guidelines were in place, the domain-expert researcher who had led their development (annotator **A2**) annotated every advert, including the 113 we later used to measure inter-annotator agreement (Section 4.2).

3.2 Span Labels

As a result of this process, we reached a consensus on three main span labels (these may also be referred to as 'classes' or 'tags') for our study. These tags were designed to classify text abstractly, capturing a broader portion of the content with a small number of labels. Specifically, we identified three tags:

- Condition Tag: refers to the current condition of the vehicle. Instances of this may include engine noise, the presence of scratches, and tyre condition. The span would typically include a noun and an attribute (e.g., "tyre tread is low").
- Historic Tag: refers to an event that happened to the car before the posting of the ad. Instances of these include past accidents, number of previous owners, component change and servicing (e.g., "new cambelt fitted last August").
- Sales Options Tag: refers to any tangible item or service offered during the sale — apart from the car. These could refer to part exchange offers, delivery, warranties and test drives.

Why three labels? The labels separate three readings that the same mention can carry: a state of the car now (*Condition*), an event in its past (*Historic*), or something offered with the sale (*Sales Options*). Table 8 (Appendix F) shows four frequent mention types under all three labels. The very string "12 months MOT" is annotated six times as *Condition* (the certificate the car holds) and six times as *Sales Options* (a promise at the point of sale). The distinction is contextual rather than lexical, which makes the task hard even for people: of the 586 spans on which the two expert annotators agreed on boundaries, they disagreed on the label for seven, six of them MOT or service mentions (e.g., "Recent service", read as *Historic* by one and *Condition* by the other). Collapsing the three labels would lose exactly what downstream users need to know: whether an MOT is valid now, was passed recently, or is being offered.

In Appendix B, we disclose the annotation guidelines, which provide the definitions of each label, detected sub-types and span labelling rules. Furthermore, additional span types encountered during our annotation phase were excluded from the study for reasons detailed in the appendix (see Appendix C for documentation).

3.3 Dataset Origin

We sourced the data from our industry partner's proprietary database of current car advertisements. Half of the texts of these advertisements are written by retailers, covering *traders*, car dealerships with a large available stock of vehicles and high turnover and the other half by *individuals* typically infrequently listing a single vehicle.

In addition, to ensure a representative sample for the trader dataset, we sorted traders by the number of their listings and selected those with the highest number of listings, as they typically follow a

Number of	CONDITION	HISTORIC	SALES OPTIONS
# Labels	594	810	2,174
# Tokens (spaCy)	2,466	3,985	6,216
# Sub-word tokens (DeBERTaV3)	2,957	4,559	6,875

Table 1: Distribution of labels in our dataset, with the spaCy word tokens they cover and the corresponding sub-word tokens of the `deberta-v3-base` vocabulary.

template for their sale ads. This approach ensured a stratified sample of car-trader listings, capturing both those that follow a standardised template and those that write listings freely. For the private dataset, we stratified the listings by UK region (e.g., "North West") and sampled a proportional number of car listings from each region. This method ensured that different regional variations of the English language are represented in our dataset.

An example of a short advert is provided within the annotation interface, as shown in Appendix A, Figure 4.

3.4 Dataset Overview

In our dataset, a total of 605 advertisements were analysed, amounting to 104,382 tokens (spaCy's rule-based English tokenisation, which we use for annotation, the CRF and evaluation throughout). Notably, 30 of these advertisements (approximately 5%) contained no annotatable spans. These ads typically consisted solely of sales pitches, car standard/optional features, or copied the header (e.g., "2.0-litre engine BMW"). In Table 1, we present the total number of labels and the spaCy tokens they cover for each label in our dataset.

Compared to other sequence labelling datasets from the literature, our dataset would be most similar to "noisy" datasets such as WNUT16 (Strauss et al., 2016) and Twitter NER (Ritter et al., 2011). The similarities lie in the use of colloquial language, typos, non-standard formatting, abbreviations, and some instances of emoticons. However, the unique aspect of our dataset is that it is designed for users to submit sales pitches for their cars. Consequently, distinct features include the use of telegraphic-style language, non-standard grammar (e.g., missing connecting words), the omission of contextual information and capitalisation to capture the buyer's attention. Additionally, we observed extensive use of persuasive language (e.g., positive adjectives), direct calls to action (e.g., imperative sentences encouraging viewers to call or not to miss out on the opportunity), and detailed lists of features, attributes or services.

Moreover, we observed a clear distinction between texts authored by traders and those written by individuals. Traders typically adopt a customisable, template-based approach to ensure professionalism and efficiency, whereas individuals' posts vary widely in style and tone. Furthermore, traders often enumerate all *Sales Options* they offer using bullet points or lists, while individuals generally do not provide such options.

To gain deeper insight into the linguistic characteristics associated with each class, we conducted a comparative analysis of part-of-speech (PoS) tag distributions across the three labels using spaCy (Honnibal et al., 2020). The *Sales Options* label is characterised by a notably higher proportion of singular nouns (37.7%) and proper nouns (25.7%), reflecting a strong emphasis on the financial products offered by retailers. In contrast, the *Historic* label demonstrates a more balanced distribution, showing substantial usage of singular nouns (23.9%), adjectives (14.9%), proper nouns (14.5%), and cardinal numbers (13.4%), which indicates a descriptive narrative style focusing on historical events, proper names, dates, and the vehicle's number of owners.

The *Condition* label similarly exhibits balanced noun usage (23.7%) but places greater emphasis on cardinal numbers (14.1%) and prepositions or subordinating conjunctions (10.2%). These patterns suggest a descriptive and relational linguistic structure oriented towards vehicle conditions, states, current mileage, and MOT expiry dates. Overall, these distinctions highlight slightly differing linguistic profiles, with *Sales Options* being more noun-focused, *Historic* event- and description-oriented, and *Condition* emphasising descriptive and relational language. Table 6 in Appendix D summarises the most common PoS tags per label.

4 Experiments

4.1 Evaluation Metric

For evaluation, we used an adjusted F_1-score for NER. We report the partial match based on Segura-Bedmar et al. (2013) who take into consideration various categories of errors (Chinchor and Sundheim, 1993). Specifically, they report:

1. CORrect, when the spans between the gold standard and the prediction match entirely
2. INCorrect, when the spans match, but the label is different

3. PARtial, when the spans overlap and have the same label
4. MISsing, when a gold annotation span is not predicted
5. SPUrious, when a model predicts spans which do not exist in the gold standard.

Following this, we define the possible and actual matches as follows:

$$\begin{aligned} POS\text{sible} &= COR + INC + PAR + MIS \\ &= TP + FN \\ ACT\text{ual} &= COR + INC + PAR + SPU \\ &= TP + FP \end{aligned}$$

The precision and recall can be characterised as:

$$\begin{aligned} \text{Precision} &= \frac{COR + 0.5 \times PAR}{ACT} \\ &= \frac{TP}{TP + FP} \end{aligned}$$

$$\begin{aligned} \text{Recall} &= \frac{COR + 0.5 \times PAR}{POS} \\ &= \frac{TP}{TP + FN} \end{aligned}$$

We utilise the F_1-score (the harmonic mean between precision and recall) for IAA by applying it with one annotator serving as the gold standard against the other. Cohen's Kappa is widely recognised as the preferred metric for IAA, having been developed to address the shortcomings of percentage agreement by accounting for chance agreement (McHugh, 2012). However, research has pointed out its unsuitability for NER, attributing this to the lack of true negatives (Brandsen et al., 2020; Richie et al., 2022).

4.2 Inter-Annotator Agreement

The IAA study involved three annotators, all authors of this paper: a domain-expert data scientist from the industry team (**A1**); a university-based researcher who led the creation of the annotation guidelines (**A2**); and a university-based, non-automotive domain-expert researcher who joined after the guidelines were finalised (**A3**).

The IAA was calculated on 113 adverts annotated by all three annotators, representing approximately 19% of the total corpus (605 ads). The mean pairwise partial-match F_1-score was 88% for the **A1**–**A2** pair, 73% for **A1**–**A3**, and 73% for **A2**–**A3**, achieving an overall score of ∼78%. A detailed breakdown of agreement by category averaged across all the annotators is presented in Table 2. More details on the exact and partial matching scores per label are provided in Table 7 (Appendix E).

Label	COR	INC	PAR	MIS	SPU
CONDITION	72%	7%	12%	5%	5%
HISTORIC	63%	5%	20%	6%	6%
SALES_ OPTIONS	60%	2%	22%	8%	8%

Table 2: Pairwise inter-annotator agreement averaged across all three annotators for each label. Scores are computed over all ordered annotator pairs and report the proportion of correct (cor), incorrect (inc), partial (par), missing (mis), and spurious (spu) span matches following the evaluation scheme in Section 4.1.

4.3 Models

For our experiments, we compared in total seven models. Following, we have grouped them by type.

4.3.1 Encoder Transformer Models

We tested two types of encoder architecture transformers. One was a BERT (Devlin et al., 2019) model, the other was a DeBERTaV3 (He et al., 2021) model. We trained them according to the token-classification methodology[2] from HuggingFace (Wolf et al., 2020). Listing 1 in Appendix H details the hyper-parameters used in our experiments for these two models.

4.3.2 Decoder Transformer Models

We experimented with two open-source and two closed-source decoder-based transformer large language models. Specifically, these were: GPT-4o (OpenAI et al., 2024), Gemini Flash (Team et al., 2025), Llama (Grattafiori et al., 2024) and Qwen (Yang et al., 2024). Their precise versions are in Listing 2 (Appendix H). The full prompt comprises three parts, shown in Figures 1–3 in the order in which they appear: first, the system instruction (Figure 1); second, the task and label definition (Figure 2); and finally, the in-context examples followed by the advertisement to be annotated (Figure 3).

The system-instruction prompt (Figure 1) constrained the model's outputs to minimise verbosity, suppress explicit reasoning and justification of choices, and focus exclusively on task execution. It also included a role-playing instruction assigning

[2] token-classification

System Instruction Part 1/3

You are a Natural Language Processing assistant specializing in car-related named entity recognition. In the following conversation, perform the tasks exactly as requested, without providing explanations or reasoning.

Figure 1: Prompt part 1/3: the system instruction used for the LLMs.

Label Definition Part 2/3

User: The task is to label <label> entities in the given text. <label> is characterised as '<label_definition>'. Your task is to detect <label> span words by inserting '@@' at the beginning of the span and '##' at the end, and generating the remaining text verbatim, whitespaces and newlines included for parsing purposes.
Assistant: Sure, I can help with that. Please provide the text you want me to label for <label> entities.

Figure 2: Prompt part 2/3: the prompt that introduces the task and the label's definition (label_definition). The label definitions are in Appendix B.

the model the role of an NLP assistant, which prior work has shown to be effective (Kong et al., 2024).

For prediction with the LLMs, we followed the NER setting as proposed by Wang et al. (2023). We employed random 100 samples from the training data in an in-context learning (ICL) scenario. Specifically, for each label, we empirically added the label's definition at the beginning of the prompt; one can see our prompt template in Figure 2.

Then we used a chat alternating history to demonstrate how a text should be labelled. At the end of the prompt, we included a sample from the test dataset to collect the generated output. The output would replicate the input text, but the spans would be marked by prefixing with '@@' and suffixing with '##' to segment them from the rest of the text. An example is shown in Figure 3. In cases where the text was not copied verbatim, we tackled this issue using an alignment library[3].

[3] pytextspan

ICL Prompt & Inference Part 3/3

User: We offer 6 month warranty.
Assistant: We offer @@6 month warranty##.
User: We provide Full Service history.
Assistant: We provide @@Full Service history##.
User: 12 month warranty is included in the sale.

Figure 3: Prompt part 3/3: the in-context examples and the advertisement to be annotated. The figure shows only two ICL examples for illustration. These would typically be longer—an entire car advert, which might comprise one to several sentences—rather than a single sentence. The last **User** utterance points to a sentence (i.e., advertisement) that we expect the LLM to annotate with the special characters "@@" and "##" to obtain the token spans that correspond to the label of interest.

4.3.3 Conditional Random Fields

Lastly, we employed CRF models (Lafferty et al., 2001) with different feature configurations, utilising a range of lexical, morphological, and part-of-speech tags (PoS) obtained through the spaCy NLP toolkit (Honnibal et al., 2020). The features for the tokens were encoded with the following attributes extracted from spaCy:

- lowercased form
- suffixes (last two and last three characters)
- whether it is capitalised or title-cased
- whether it is a digit
- token shape (pattern of character types)
- whether it is alphabetic
- whether it is a stop word
- both coarse- and fine-grained PoS tags.

Furthermore, we incorporated contextual features from the immediately preceding and following tokens (i.e., ±1, ±2, ±3): their lowercased forms, coarse-grained PoS tags, and digit indicators. To fine-tune the CRF model's hyperparameters—specifically, the L1 and L2 regularisation coefficients—we conducted a randomised 100-iteration hyperparameter search on the validation set.

On this feature set we implemented four configurations:

1. CRF: includes all of the features, with a ±1 token context window.

2. CRF PoS: includes only spaCy's PoS tags.
3. CRF w/o PoS: includes all features (as in 1.) except spaCy's PoS tags.
4. CRF ±2/±3: identical to 1. but with broader context windows (±2 and ±3 tokens).

We included the PoS-only setting (2) and our first setting without the PoS tags (3) to evaluate the influence of PoS tags since our data overview analysis, in Section 3.4, indicated slightly different PoS distributions across labels.

4.3.4 Computational Resources

All training and inference for the open-source models were performed on NVIDIA H100 Hopper graphics cards, except for the CRF implementations, which, due to their lower computation requirements, were run on an Apple M1 Max chip.[4]

4.4 Results

For our experiments, we measured performance across three folds of the dataset, where the test sets were mutually exclusive. Each fold used 70% of the instances for training, with 15% each for validation and testing. The LLMs would not require the validation dataset, since we did not fine-tune them. In some cases, the LLMs' generations were incorrect and could not be recovered, even with our alignment library; this occurred rarely, less than 1% in each fold. The overall results averaged across the three folds are presented in Table 3. Table 4 breaks the F_1-score down per label for the four best-performing models, and Section 4.5 decomposes the matches of those four models into the five categories of Section 4.1.

One can view in the table that the fine-tuned DeBERTaV3 model, trained on the full training split, obtains the highest F_1-score. Looking more closely at the table, the precision is much higher than the recall for the decoder models. This may be because we utilised only 100 random ICL examples per fold. We took this approach to maintain fairness in comparing the models and reduce computation time and cost.

We noticed that GPT-4o did not annotate a particular span(s) at various times (i.e., low recall). LLMs may be susceptible to errors in previous generations or context, affecting subsequent text generations (Shorinwa et al., 2025). Hence, we hypothesise that multiple LLM agents with different ICL examples could yield better results after aggregating the outputs. Another technique could be to select more representative ICL samples per advertisement to label.

The CRF approach outperformed every LLM and BERT, highlighting its effectiveness in low-resource, domain-specific tasks. Analysis of the trained CRF model showed that its most influential features were lexical in nature. For instance, terms such as "FSH" and "warranties" strongly indicated the *Sales Options* label. This observation aligns with the results presented in Table 3, where the CRF implementation employing default features performed best, while the model relying solely on PoS features performed poorly. The ten most influential features per label identified by the CRF implementation are listed in Appendix I.

Model	Precision	Recall	F1-score
BERT-base	52.6 ±1.7	63.5 ±2.4	57.5 ±1.2
DeBERTaV3-base	63.6 ±1.2	**72.4 ±2.0**	**67.7 ±0.8**
GPT-4o	73.8 ±1.6	54.4 ±1.3	62.6 ±0.3
Gemini 1.5 Flash	72.4 ±2.1	49.9 ±3.2	59.0 ±2.0
Llama 3.1 70B it	75.9 ±1.1	44.5 ±2.3	56.1 ±2.1
Qwen 2.5 72B it	**76.8 ±3.2**	34.8 ±1.1	47.9 ±1.6
CRF	72.5 ±1.8	60.8 ±1.5	66.1 ±1.3
CRF PoS	25.4 ±8.6	1.1 ±0.1	2.1 ±0.2
CRF w/o PoS	71.9 ±0.9	58.5 ±2.5	64.4 ±1.7
CRF 2n/3n	71.2 ±0.8	60.1 ±0.9	65.2 ±0.8

Table 3: Mean and population standard deviation over the three folds of precision, recall and F_1-score (%) for each model; "it" stands for instruction-tuned; bold marks the highest value in each column.

Model	F1-score (mean±std)		
	Historic	Sales Option	Condition
DeBERTaV3-base	58.1±2.3	**74.0±2.7**	**57.8±5.7**
GPT-4o	53.0±2.2	68.4±1.9	52.2±5.2
Gemini 1.5 Flash	51.6±2.4	66.7±4.1	38.5±5.2
CRF	**59.5±3.8**	73.2±2.8	46.5±4.4

Table 4: F_1-score performance across different labels for the top four performing models. Values are shown as percentages. The highest scores in each column are bolded.

[4]Hardware specifications: https://www.nvidia.com/en-gb/data-center/h100/ and https://support.apple.com/en-gb/111902.

4.5 Error Analysis

Table 5 decomposes the matches of the four best-performing models into the five categories of Section 4.1; Appendix G gives the same breakdown per label. It resolves an apparent tension in Table 3: the LLMs enclose spans more tightly, yet score lower. Their precision is the highest because a span they do emit is more often exact and less often spurious—partial matches make up 14.6% of GPT-4o's matches against 17.2% for the CRF and 19.8% for DeBERTaV3—but they leave far more gold spans unannotated (31.2% of GPT-4o's and 35.9% of Gemini's matches are missing spans, against 23.2% for the CRF and 7.7% for DeBERTaV3). Since a partial match still earns half credit whereas a miss earns none, the models that find more spans with looser boundaries come out ahead on F_1. The boundary errors also differ in kind: when the CRF or GPT-4o produces a partial match, the predicted span is more often too wide than too narrow (58% against 33% of the CRF's partial matches, 58% against 38% of GPT-4o's), whereas DeBERTaV3's partial matches are more often too narrow (53%). Per label (Table 9), *Condition* is the hardest for every model, with the highest shares of label confusions and of misses: GPT-4o and the CRF each fail to annotate about 40% of the gold *Condition* spans, Gemini about 54%, and DeBERTaV3 about 12%. The contextual mention types of Table 8 are where label confusions concentrate: on gold spans mentioning an MOT, the four models assign the wrong label to 15.2% (DeBERTaV3), 19.3% (GPT-4o), 27.1% (Gemini) and 19.1% (CRF) of the spans they find, between four and seven times their rates on spans outside these types (3.0–5.0%); service mentions show the same pattern at 10.9–14.5%. The models also split where the experts split: on "12 month MOT", gold *Sales Options*, all four predicted *Condition*, the reading the other expert annotator had chosen. Finally, although not the top scorer, the CRF model may be preferable to transformer-based approaches due to its greater algorithmic accountability and transparency (Diakopoulos, 2015).

Qualitative Examples The CRF and GPT-4o performed worse than human annotators at delineating contextual cues. For instance, in the example "This is my lovely car that I have used for 14000 miles" the CRF mistakenly treated the miles as the car's mileage rather than the distance the last owner had driven (a label confusion, *Condition* for *Historic*). Whereas, in the nuanced example of "we will add new carpets upon purchase", the CRF correctly predicted the statement as *Sales Options* since it is added value at the point of sale, but GPT-4o labelled the part as *Condition*, which is a less favourable label since it is not a current statement about the car but rather an added sales feature (again a label confusion). Moreover, in an example where the owner had written the mileage stamps for when their car was serviced, the CRF falsely labelled them as *Condition* (a label confusion), and they were missed by GPT-4o (a missing span), which should have been labelled as *Historic*. This suggests that both approaches have merits and downsides; perhaps an ensemble of the two could yield more favourable results. To add to that claim, it is evident in Table 4 that the CRF demonstrates a higher average F_1-score than GPT-4o for predicting the *Historic* and *Sales Options* labels, whereas GPT-4o scored higher on the *Condition* label.

Model	COR	INC	PAR	MIS	SPU
DeBERTaV3-base	48.7	4.9	19.8	7.7	18.8
GPT-4o	43.3	3.9	14.6	31.2	7.0
Gemini 1.5 Flash	40.6	4.6	11.2	35.9	7.6
CRF	47.0	4.1	17.2	23.2	8.4

Table 5: Match types per model as percentages of its matches pooled over the three test folds, following Table 2 (COR correct, INC incorrect label, PAR partial overlap, MIS missing, SPU spurious); Appendix G gives the breakdown per label.

5 Discussion

We present the first annotation schema for car advertisements, meticulously developed to meet genuine industry demands in a real industry setting. We established three span labels and corresponding guidelines after two rounds of collaborative iterations between university researchers and industry partners. We then evaluated inter-annotator agreement, achieving a commendable 88% F_1-score between the domain-expert annotators and 78% overall.

One theme that emerged from Table 2 was that there were slightly more partial matches for the *Sales Options* and *Historic* labels (i.e., only segments of a span are matched against the ground truth). Although we made efforts to mitigate this issue by editing the annotation guidelines, we attribute it to differences in interpretation when labelling. *Historic* labels often refer to past events,

leading annotators to include additional, possibly extraneous, contextual information. *Sales Options* labels, owing to the liberal use of adjectives to describe services a dealer offers, create ambiguity about which aspects are sufficiently factual or necessary for inclusion.

Potential Use Cases The three labels allow us to conduct multiple analyses. Firstly, the *Sales Options* enable us to examine, across temporal and geographical dimensions, which sales options/features are offered, such as warranties, financing, and part-exchanges, which may vary significantly depending on economic conditions and regional preferences. This raises questions about whether a robust economy alters these trends and what consumers expect from their dealings with car dealerships.

These labels can help us better understand which information writers typically include in advertisements and, in marketing terms, what makes a listing more appealing. For example, they can show whether the *Condition* of a vehicle (e.g., the presence of new tyres or the extent of scratches) plays a critical role in shaping consumer decisions, and whether *Historic* labels, like the number of previous owners or a vehicle's maintenance history, are valued across different regions.

This study serves as a cornerstone for downstream tasks, since it delineates the nuanced readings of Table 8 and would make further subcategorisations simpler. We identified several subtypes of mentions emerging from our data. A non-exhaustive list of these sub-types is documented in Appendix B. Looking ahead, we aim to explore **entity typing**, where the type of each span is generated as free text rather than selected from a closed ontology (Choi et al., 2018; Yuan et al., 2022).

Wider Applicability The application of such span labelling extends to safety and accident analysis, where patterns categorised by location and severity provide essential insights for police, insurance companies, and urban planners considering road expansions.

One can use our work to develop tools to educate car owners about necessary or upcoming maintenance, potentially revolutionising consumer knowledge and empowering car owners. Moreover, it could be applied to quantify and highlight instances of under-reporting and suggest policies to enhance consumer awareness. Such comprehensive use of data ensures a more transparent marketplace and significantly enhances consumer safety and satisfaction.

We believe the text-mining applications enabled by the proposed span-labelling schema can benefit various stakeholders in the vehicle trading industry. The schema can offer more accurate evaluations of vehicle features, auto-complete suggestions when writing an advert, and evaluation of advertisement quality.

6 Conclusion

Our effort represents the first-ever English-language annotation schema and dataset for information extraction in the automotive advertisement domain. We collaboratively developed the span annotation schema with involvement from industry and academia. The three labels chosen allow us to annotate meaningful spans that alter the car's value or affect the sale proposition. We achieved an encouraging IAA score and observed that a fine-tuned DeBERTaV3 encoder performed best, closely followed by the CRF, with Qwen leading in precision. The CRF demonstrates its adaptability on this low-resource domain-specific task, surpassing all of the LLMs. Lastly, after analysing the models, we elaborate on different uses of this schema from insurance companies to empowering car owners.

Limitations

As a foundational case study introducing a novel domain schema, this work has certain practical limitations that pave the way for future research.

Firstly, the manually annotated dataset of 605 advertisements was sufficient to establish the **AutoAdvER** schema, reach consensus, and benchmark our baseline models, though a larger corpus would capture more of the long tail of linguistic variation in the UK automotive market. Because the advertisements are written in British English, they also reflect UK-specific terminology and conventions that may differ slightly from those in other English-speaking regions.

Secondly, reflecting the genuine composition of retail vehicle advertisements, there is a natural label imbalance within our dataset. Because sellers typically enumerate numerous services and perks to attract buyers, *Sales Options* naturally occur more frequently than *Condition* or *Historic* mentions. Although this represents a realistic data distribution, it poses a standard learning challenge for supervised models, which could be mitigated in future

iterations through targeted sampling.

Finally, regarding our experimental setup, the use of closed-source LLMs complicates direct comparisons with open-source models due to their undisclosed training data. This lack of transparency can lead to potential data contamination issues, obscuring true performance differences and making baseline evaluations against open-source models incomplete.

Ethics Statement

In conducting this research, our primary motivation is to contribute to the greater good by enabling more informed consumers and enhancing the capabilities of individuals selling cars. We believe that a robust annotation schema for car advertisements will empower consumers with clearer, more reliable information and provide sellers with a tool to better present their products. Additionally, we acknowledge the ethical considerations associated with our work. We utilised LLMs. While these and similar models offer significant advancements, it is important to remain mindful of their substantial energy consumption and the environmental impact associated with their extensive use. Therefore, we advocate a balanced approach, ensuring that the benefits of such technologies are weighed against their ecological footprint.

Acknowledgements

We thank our industry partner, Autotrader UK, for providing the car advertisement data, for sharing the domain expertise that shaped the **Auto-AdvER** schema and dataset, and for the funding that supported this research.

References

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Diakopoulos. 2015. Algorithmic accountability. *Digital Journalism*, 3(3):398–415.

Elena Epure and Romain Hennequin. 2023. A human subject study of named entity recognition in conversational music recommendation queries. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1281–1296, Dubrovnik, Croatia. Association for Computational Linguistics.

Zihao Fu, Yixuan Su, Zaiqiao Meng, and Nigel Collier. 2023. Biomedical named entity recognition via dictionary-based synonym generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14621–14635, Singapore. Association for Computational Linguistics.

Veena G., Vani Kanjirangat, and Deepa Gupta. 2023. AGRONER: An unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Systems with Applications*, 229:120440.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Runwei Guan, Ka Lok Man, Feifan Chen, Shanliang Yao, Rongsheng Hu, Xiaohui Zhu, Jeremy Smith, Eng Gee Lim, and Yutao Yue. 2024. FindVehicle and VehicleFinder: a NER dataset for natural language-based vehicle retrieval and a keyword-based cross-modal vehicle retrieval system. *Multimedia Tools and Applications*, 83(8):24841–24874.

Tran Thi Hong Hanh, Antoine Doucet, Nicolas Sidere, Jose G. Moreno, and Senja Pollak. 2021. Named entity recognition architecture combining contextual and global features. In *Towards Open and Trustworthy Digital Societies*, pages 264–276, Cham. Springer International Publishing.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python.

Songhua Hu and Ruhu Ma. 2024. Named entity recognition of automotive parts based on RoBERTa-CRF model. In *2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pages 604–612.

Basra Jehangir, Saravanan Radhakrishnan, and Rahul Agarwal. 2023. A survey on named entity recognition — datasets, tools, and methodologies. *Natural Language Processing Journal*, 3:100017.

Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. 2024. Recent advances in named entity recognition: A comprehensive survey and comparative study. *Preprint*, arXiv:2401.10825.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2023. A survey on deep learning for named entity recognition : Extended abstract. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3817–3818.

Cedric Lothritz, Kevin Allix, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. 2020. Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3750–3760, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22:276–282.

John Molineux. 2018. Using action research for change in organizations: processes, reflections and outcomes. *Journal of Work-Applied Management*.

Ines Montani and Matthew Honnibal. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models. *Prodigy*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Jiho Park, Jiseong Lee, Seonghyeon Kim, and Chanjun Park. 2023. ADMit: Improving NER in automotive domain with domain adversarial training and multi-task learning. *Expert Systems with Applications*, 225:120007.

Russell Richie, Sachin Grover, and Fuchiang (Rich) Tsui. 2022. Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 275–284, Dublin, Ireland. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK. Association for Computational Linguistics.

H. Bahadir Sahin, Mustafa Tolga Eren, Cagil Tirkaz, Ozan Sonmez, and Eray Yildiz. 2017. English/Turkish Wikipedia Named-Entity Recognition and Text Categorization Dataset. Mendeley Data, V1.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3).

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie

Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang, and Yanghua Xiao. 2022. Generative entity typing with curriculum learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3061–3073, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Annotation GUI

For annotating we used Prodigy by spaCy (Montani and Honnibal). We implemented a customised sequence labelling version, which let our annotators choose amongst the three labels. Also, below the annotation text, the interface will show the vehicle's specifications that are being annotated for context. Figure 4 shows an example of an annotated text.

B Label's Definitions & Identified sub-entities

Below, are the annotation guidelines used for our study, along with the sub-entities detected.

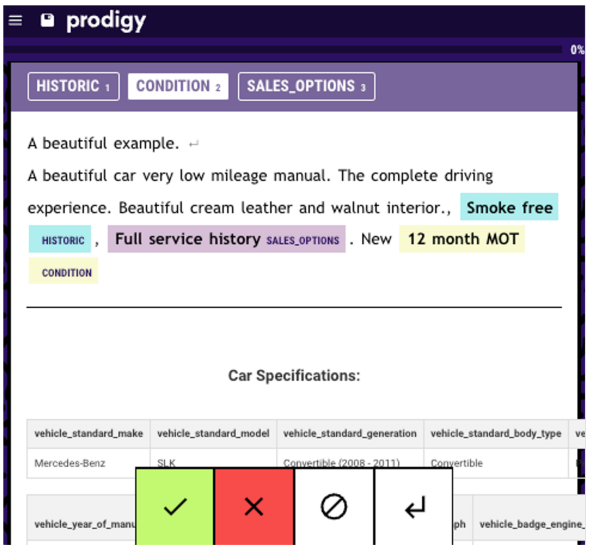


Figure 4: An example of the annotation interface, illustrating a customised NER annotation task implemented using Prodigy, displayed within a reduced-size window.

Condition - Current state

The *Condition* category encompasses any valuable car-related information that is current as of the date the advertisement was posted online.

Definition

The *CONDITION* label identifies any word or span of words in car advertisements that indicate the present material or documented condition of the vehicle, both positive and negative. This includes descriptions that reflect the car's current operational performance, mechanical state, aesthetic appearance, maintenance status, and legal classifications affecting its value or usability

These include:

- Tyre condition (i.e. trim level or mention of good condition)
- Interior condition (e.g., "leather like new")
- Exterior condition (e.g., "presence of scratches")
- Mechanical condition (e.g., "engine noise", "gear not locking properly")
- MOT[5] due date or recent done MOT men-

[5]This refers to the Ministry of Transport, which, in a broader context, represents any legal authority responsible for ensuring vehicle safety by enforcing compliance with minimum standards.

tioned

- Service due date or recent done Service mentioned
- Current mileage (e.g., "12k miles")
- Cat S, C, D or N[6]
- Registered, pre-registered, non-registered, demonstrator[7]

Annotation Span Length & Methodology:

1. A condition span is typically characterised by the noun of part of a vehicle (e.g., "tyre thread") and an attribute that links to (e.g., "low"). You would annotate the whole phrase (e.g., "tyre thread is low"). And omit anything else.

2. We would disregard statements that are opinion based, such as "low mileage", since it is a statement denoting an opinion that the car is of low mileage of its age. Instead "80 000 mileage" would be labelled, since it's presented as a fact. In a similar fashion "amazing condition" does not add enough valuable information to the ad, being characterised of opinion, whereas "clean paintwork" is specific and adds value to the ad, hence, we would label the second one but not the first.

3. Remember to label span words that are referring to the state of the vehicle when the ad was posted e.g., "6 month MOT", "80 000 miles", references to doing extra miles after the ad was posted can be safely ignored.

Historic - Events in the past

The *Historic* attribute pertains to events that have occurred prior setting the car advertisement. This category focuses on the events that happened.

[6]This refers to vehicle categories (such as Cat S, C, D, or N), which, in a broader context, denote classifications used by insurance companies and authorities in the UK to indicate vehicles that have sustained damage and have subsequently been repaired or written off.

[7]This refers to vehicle registration statuses (such as registered, pre-registered, non-registered, or demonstrator), which, in a broader context, indicate whether a vehicle has been officially recorded with authorities, previously assigned registration by a dealership, remains unregistered, or has been used primarily for demonstration purposes.

Definition

The *HISTORIC* label identifies any word or span of words that presents events that have happened to the car in the past and affected its condition mechanically condition or bodywork.

These include:

- Mention of past accidents (e.g., "3 years back I scratched the front side")
- Number of owners
  - Solely without mentioning "from new". E.g., "3 owners ~~from new~~"
- Past type of usage
  - Any mention of past usage that could affect its condition, either positive or negative. E.g., "smoke free", "motorway mileage", "stored in locked garage"
- Component change (e.g., "Cambelt changed a year ago", "New brake callipers", "new tyres, brake discs") and tuning (e.g., "re-mapping")
- MOT advisories or no advisories mentioned
- HPI[8] is clear (as opposed to only providing HPI report as a Sales Option, see further)
- Service history done in the past, including stamps (e.g., 8 stamps), dates (e.g., "02/2023, 01/2024") or on which mileages (e.g., "@ 30k, @40k, @49k") and where (e.g., "BMW")

Annotation Span Length & Methodology:

1. The annotation span under the Historic can be considered as the most elaborate amongst the other classes. It would refer to events that happened in the past (e.g., "Replaced clutch", "3 owners", "Last service was done 3,000 miles ago", "primarily done on the motorway").

2. Should there be a mention of damage and subsequently a relevant mention of component change, these two spans are separate. Treat each of the aforementioned bullet points as a separate span.

[8]This refers to Hire Purchase Investigation, which, in a broader context, denotes a vehicle history check commonly used in the UK to verify whether a vehicle has outstanding finance, has been reported stolen, previously written off, or has other hidden issues.

3. If a span mentions a component change that still affects the car today, then it's labelled as Historic (e.g., "2 new front tyres fitted last May").

Sales Options

Sales Options involve various aspects related to the selling process.

> Definition
>
> The *SALES OPTIONS* label refer to any word or span of words that indicate something tangible or a service offered during or after the sale apart from the car itself, and enhances the sale proposition.

These include:

- Part exchange offers
- Number of keys
- Delivery & delivery options
- Extended Opening times (i.e. unusual times, late night or weekends)
- Warranty available
- Financing options
- Sale of the car's number plate (or not)
- Availability of additional photos from past accidents
- Offering test drive
- Price negotiability
- Any documentation, for example:
  - Full service history logbook (FSH)
  - Providing MOT papers
- Valeting/Cleaning
- And any other kind of service that could be considered an advantage over other sellers

Annotation Span Length & Methodology:

1. We would label the noun part of the sale, or any adjective specifying it, but disregarding any conditions. For instance, "2 year warranty ~~when financed with Volvo Services * Terms and Conditions apply.~~".
2. Label spans that add value to the ad, according to your judgement. If something seems like a Sales Option but does not add value in comparison with the rest of the ads, then disregard it. Instances of those are contact details and location.

C Not Tracked Named Entities

Throughout our iterations, we documented some key themes from the text part of the advertisements that we identified as recurrent features of the texts but not useful for annotation. As such, we explicitly do not annotate for the following reasons:

- "Logistical costs": This includes the amount of tax, insurance, whether the vehicle complies with the ULEZ[9], and how long a full tank lasts. However, this information is redundant as it can be calculated more accurately from the car's specifications, given that the time of purchase, location, and type of use will affect these values.
- "Personal Sentiments": We found statements denoting emotional distress for selling the car. For example, "it breaks my heart to see this car go". We omitted these statements, since they do not add to any objective quantifiable metric.
- "Reason for selling": Moreover, a few advertisements disclosed a reason for selling. These could refer to an individual buying a newer vehicle, or downsizing to a smaller vehicle. Although one could argue these statements may have some value to the buyer, we omit those since again we believe they do not add to an objective quantifiable metric.
- "Car Identifiables": In multiple advertisements in the text description, sellers tend to reiterate the car model's brand, model, year, colour and engine size. Nonetheless, for any online car advertisement, these statements are included in the car advert header. Hence, we omitted them.
- "Invitation for communication & location": We would avoid labelling references to contact them, or where they are located.

[9]The acronym refers to the Ultra Low Emission Zone charge in London, where more polluting vehicles are charged accordingly.

- "Optional & Standard features": Detecting car features, both optional and standard, is of great interest, since we have been told by our industry partner that they can change the price evaluation of a vehicle. Nonetheless, throughout our investigations features can differ from vehicle to vehicle. To make the point clearer, what one manufacturer considers as a feature either a standard or an optional one, may not necessarily be one according to another manufacturer's opinion. Hence, we have another piece of work that targets solely the annotation of features based on a dictionary table, casting it as a contextual NER problem.

- "Sales Pitch": In certain instances, particularly among traders rather than individual sellers, there are elaborate descriptions that emphasise the vehicle's perceived suitability and appeal to potential buyers, yet without providing substantive factual information.

D PoS Exploratory Analysis

Please refer to Table 6 for the most frequent part-of-speech metrics per label.

PoS Tag	HISTORIC	CONDITION	SALES OPTIONS
NN	23.9%	23.7%	37.7%
JJ	14.9%	12.5%	10.9%
NNP	14.5%	13.1%	25.7%
CD	13.4%	14.1%	10.2%
NNS	12.6%	11.7%	9.9%
IN	7.5%	10.2%	2.2%
VBN	5.7%	-	-
DT	-	4.6%	-
CC	-	-	1.8%

Table 6: Top PoS Tag Frequencies as Percentages for Each Label. The detailed PoS tags can be found in https://web.archive.org/web/20190206204307/https://www.clips.uantwerpen.be/pages/mbsp-tags.

E IAA Label Analysis

Table 7 reports inter-annotator agreement across all annotator pairs, broken down by entity label and measured using both exact and partial match precision, recall, and F_1 as defined in Section 4.1. Scores are computed over all ordered pairs of the three annotators, where each annotator serves as the reference in turn, yielding six pairwise comparisons in total. The raw counts (*COR*, *INC*, *PAR*, *MIS*, *SPU*) are aggregated across all pairs prior to computing the final F_1. For exact match, the *PAR* category is excluded, and a span must fully agree on both boundaries and label to be counted as correct.

Label	Match	P	R	F_1
CONDITION	Exact	75.60	75.60	75.60
	Partial	81.64	81.64	81.64
SALES_OPTIONS	Exact	64.88	64.88	64.88
	Partial	76.83	76.83	76.83
HISTORIC	Exact	67.28	67.28	67.28
	Partial	77.83	77.83	77.83

Table 7: Inter-annotator agreement (IAA) across all annotator pairs, reported as exact and partial match precision (P), recall (R), and F_1 (%) per entity label. Partial match awards half credit to overlapping spans.

F Contextual Mention Types

Table 8 gives one gold span per label for four frequent mention types (Section 3). The counts per label over the whole dataset are: MOT 195 / 23 / 115, service 42 / 139 / 388, tyre 34 / 62 / 5 and mileage 103 / 39 / 37 (*Condition* / *Historic* / *Sales Options*).

G Match Types per Label

Table 9 breaks Table 5 down by label. As in Table 4, each label's row restricts the gold spans to that label while predicted spans of other labels remain eligible as incorrect-label matches.

H Hyper-Parameters & LM Versions

Listing 1: Encoder transformer hyper-parameter settings used in our experiments tailored for our relatively small number of examples.

```
--num_train_epochs 30
--per_device_train_batch_size 16
--per_device_eval_batch_size 16
--load_best_model_at_end True
--metric_for_best_model eval_loss
--greater_is_better False
--warmup_steps 50
--evaluation_strategy epoch
--save_strategy epoch
--learning_rate 1e-5
```

Listing 2: LM versions used in our experiments.

```
- bert-base-cased
- deberta-v3-base
- gpt-4o-2024-11-20
- gemini-1.5-flash-002
- Qwen/Qwen2.5-72B-Instruct
- meta-llama/Meta-Llama-3.1-70B-Instruct
```

Mention	Condition	Historic	Sales Options
MOT	*"MOT Until October"*	*"never failed an MOT"*	*"6 months MOT"*
service	*"next service due September 2025"*	*"serviced every 6 months"*	*"full service history"*
tyre	*"good tyres"*	*"Tyres recently fitted"*	*"tyre repair kit"*
mileage	*"18,500 miles"*	*"new front discs at 64k miles"*	*"mileage verification"*

Table 8: The same mention type under each of the three labels: gold spans from the dataset. An MOT can be the certificate the car holds now, a test it passed in the past, or a promise made with the sale.

Model	Label	COR	INC	PAR	MIS	SPU
DeBERTaV3-base	Historic	37.8	9.5	23.5	5.7	23.5
	Sales Options	55.6	1.3	17.4	7.8	17.8
	Condition	38.3	11.7	23.8	9.7	16.4
GPT-4o	Historic	32.3	6.0	19.3	36.7	5.7
	Sales Options	49.4	1.5	13.7	27.8	7.6
	Condition	34.7	9.8	12.5	37.0	6.0
Gemini 1.5 Flash	Historic	33.5	7.7	12.9	37.1	8.7
	Sales Options	47.8	1.5	11.6	31.2	7.9
	Condition	23.6	11.8	8.0	51.3	5.3
CRF	Historic	39.6	6.6	20.8	26.4	6.6
	Sales Options	55.0	1.6	16.0	18.6	8.8
	Condition	27.7	10.2	17.2	35.8	9.1

Table 9: Match types per model and label, as percentages of the model's matches for that label pooled over the three test folds (COR correct, INC incorrect label, PAR partial overlap, MIS missing, SPU spurious).

I CRF Top Features

Tables 10–12 list the ten highest-weighted CRF state features per class. The feature name follows `sklearn-crfsuite` conventions: `token.lower()` is the lowercased current token; `token[-3:]` and `token[-2:]` are the last three and two characters, respectively; `shape` is spaCy's orthographic shape, where `x` denotes a lowercase letter, `X` an uppercase letter, `d` a digit, and punctuation is preserved verbatim; a `-1:` or `+1:` prefix indicates the feature is extracted from the immediately preceding or following token. Some values reflect noisy, user-generated advertisement text (e.g. *longmot+new* is a concatenation of "long MOT" and "new" without spacing; *bar*brand* is a stylistic separator common in listings).

Feature	Value	Description
token.lower()	reconditioning	Current token is “reconditioning”
token.lower()	securing	Current token is “securing”
token.lower()	fsh	Current token is “fsh” (Full Service History)
token.lower()	warranty	Current token is “warranty”
token.lower()	detailing	Current token is “detailing”
-1:token.lower()	visual	Previous token is “visual”
token.lower()	delivery	Current token is “delivery”
token.lower()	finance	Current token is “finance”
token.lower()	warranties	Current token is “warranties”
+1:token.lower()	expose	Next token is “expose”

Table 10: Top ten most significant CRF features for the *Sales Options* class.

Feature	Value	Description
token.lower()	unregistered	Current token is “unregistered”
-1:token.lower()	category	Previous token is “category”
token.lower()	category	Current token is “category”
+1:token.lower()	but	Next token is “but”
-1:token.lower()	excellent	Previous token is “excellent”
+1:token.lower()	✔	Next token is a tick emoji
+1:token.lower()	1x	Next token is “1x”
+1:token.lower()	ready	Next token is “ready”
shape	dd,ddd	Token shape matches a mileage pattern, e.g. “81,446”
token.lower()	longmot+new	Noisy concatenation of “long MOT” and “new”

Table 11: Top ten most significant CRF features for the *Condition* class.

Feature	Value	Description
-1:token.lower()	fronts	Previous token is “fronts”
-1:token.lower()	shaft	Previous token is “shaft”
+1:token.lower()	refurbished	Next token is “refurbished”
token.lower()	motorway	Current token is “motorway”
+1:token.lower()	advisories	Next token is “advisories”
token[-3:]	NER	Last 3 characters are “NER”, e.g. from “OWNER”
token.lower()	advisories	Current token is “advisories”
token.lower()	refurbished	Current token is “refurbished”
token.lower()	bar*brand	Stylistic separator in ad text
shape	xxx*xxxx	Token shape of the above (lowercase*lowercase)

Table 12: Top ten most significant CRF features for the *Historic* class.