Evaluation Framework for Highlight Explanations of Context Utilisation in Language Models

Jingyi Sun[*♡], **Pepa Atanasova**[*♡], **Sagnik Ray Choudhury**[♢], **Sekh Mainul Islam**[♡], **Isabelle Augenstein**[♡]

[♡]University of Copenhagen, [♢]University of North Texas, Nuremberg

{jisu, pepa,seis,augenstein}@di.ku.dk
{sagnik.raychoudhury}@unt.edu

Abstract

Context utilisation, the ability of Language Models (LMs) to incorporate relevant information from the provided context when generating responses, remains largely opaque to users, who cannot determine whether models draw from parametric memory or provided context, nor identify which specific context pieces inform the response. Highlight explanations (HEs) offer a natural solution as they can point the exact context pieces and tokens that influenced model outputs. However, no existing work evaluates their effectiveness in accurately explaining context utilisation. We address this gap by introducing the first gold standard HE evaluation framework for context attribution, using controlled test cases with known ground-truth context usage, which avoids the limitations of existing indirect proxy evaluations. To demonstrate the framework's broad applicability, we evaluate four HE methods – three established techniques and MechLight, a mechanistic interpretability approach we adapt for this task – across four context scenarios, four datasets, and five LMs. Overall, we find that MechLight performs best across all context scenarios. However, all methods struggle with longer contexts and exhibit positional biases, pointing to fundamental challenges in explanation accuracy that require new approaches to deliver reliable context utilisation explanations at scale. [1][2]

1 Introduction

Language Models (LMs) operating on context-dependent tasks (QA, summarisation, dialogue modeling) lack transparency regarding whether (Jin et al., 2024; Yu et al., 2023; Monea et al., 2024a) or how the provided context informed their generation. Highlight explanations (HEs) address this need naturally by pinpointing portions of the context responsible for the generation. See Fig.1 for an example of HEs with high/low accuracy. Although HEs have proven valuable for understanding model decisions across various tasks (Sun et al., 2025; Ray Choudhury et al., 2023; Atanasova et al., 2020), no existing work evaluates their effectiveness in accurately explaining context utilisation.

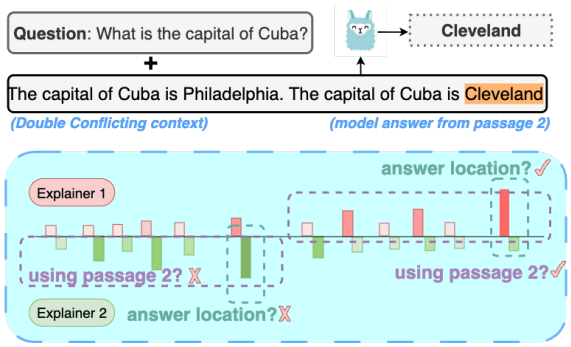


Figure 1: Utility evaluation of two HEs in our framework under the *Double-Conflicting* context setup. In this example, the model selects the answer from passage two. Explainer 1 shows better utility than Explainer 2.

Existing metrics on HE evaluation mainly focus on faithfulness (Sun et al., 2025; Lamm et al., 2021; Atanasova et al., 2020) to test whether HEs can accurately reflect the model's internal reasoning. However, faithfulness evaluations face fundamental limitations: they rely on perturbation proxies that create out-of-distribution artifacts (Hooker et al., 2019; Kindermans et al., 2019) and, more importantly, lack ground-truth explanations to validate against (Jacovi and Goldberg, 2020). We address this gap through an **evaluation framework grounded in gold standard scenarios** where ground-truth context usage is predetermined, enabling direct assessment of explanation accuracy.

Building on studies of context utilisation (Jin et al., 2024; Yu et al., 2023; Monea et al., 2024a;

[*] Equal contribution.

[1] github.com/lianyiyi/Transparent-Context-Usage

[2] huggingface.co/datasets/copenlu/transparent-context-usage

Shi et al., 2024), we construct **four controlled evaluation scenarios** (see Tab.1): *Conflicting* (one context piece (CK) contradicts parametric knowledge (PK)), *Irrelevant* (one context piece unrelated to query), *Mixed* (one conflicting + one irrelevant context piece), and *Double-Conflicting* (two PK contradictory context pieces). The settings systematically vary context usage patterns, enabling robust HE assessment across diverse behaviours.

Based on gold standard context regions in these four scenarios, **we assess the accuracy of HEs along three complementary axes**: *document-level attribution accuracy* (where we examine whether tokens from the gold document are prioritised in the generated HE), *simulatability* (where we assess how well the HEs can predict the context region (or PK) utilised for the model's prediction), and *token-level attribution accuracy* (where we evaluate whether the HE ranks the answer token highest).

To demonstrate the framework's general applicability, we apply it to four HE methods: three established ones – Feature Ablation (FA) (Li et al., 2016), Integrated Gradients (IG) (Ancona et al., 2018), and Attention visualisation (ATTN) (Abnar and Zuidema, 2020; Ray Choudhury et al., 2023), and a mechanistic interpretability MI–inspired method (MechLight), where we propose to convert the MI insights (e.g., the attention head most important for context utilisation) to HEs. Our evaluation framework is method-agnostic – it assesses any explanation technique, post-hoc or mechanistic (that generates attention-based attributions).

Across five LMs and four commonly used context-usage datasets, we find that *MechLight HEs perform best across all context scenarios*. However, two systematic limitations persist across all HEs: (i) length sensitivity – HE accuracy degrades as context grows; and (ii) position biases under dual-context inputs: FA/IG tend to favour later (near-question) pieces, while ATTN/MechLight favour earlier pieces. Surprisingly, the widely used IG and MechLight exhibit poor accuracy in most context scenarios, rendering them useless in revealing the model's context utilisation. These failures also underscore the urgent need for explanation techniques that maintain accuracy at scale and overcome positional biases in multi-document settings.

2 Related Work

2.1 Studies of Context Usage

Language models (LMs) carry vast *parametric knowledge* (PK) from pre-training, yet in practice, they must also integrate new *contextual knowledge* (CK) supplied at test time. Recent work has introduced multiple datasets to analyse how effectively LMs combine these two sources.

Early work investigates how LMs utilise CK vs. PK by crafting *single context passages conflicting with the CK*. COUNTERFACT (Meng et al., 2022), WORLDCAPITAL (Yu et al., 2023), and FAKEPEDIA (Monea et al., 2024b) each replace a Wikidata triple with a contradicting one in the context and test whether the model's answer follows CK or PK, evaluating with exact match or accuracy. CONFLICTQA (Xie et al., 2024) induces knowledge conflicts by leveraging an LLM to compose passages that contradict a model's parametric answer. While these works establish how often LMs follow the provided context, they do not consider explaining the model's behaviour. *We fill this gap by assessing whether HEs can expose the model's context usage patterns.*

In addition to the single PK-conflicting context pieces, recent work has studied other types of context. CUB (Hagström et al., 2025) considers gold (relevant), conflicting, or irrelevant passages; ECHOQA (Cheng et al., 2024) introduces a *complementary* regime, where the context alone is answer-insufficient but, when combined with the model's parametric knowledge, becomes sufficient to answer. We carefully select context types and combinations thereof that are entirely different from PK, so that the model's answer can be clearly distinguished as either from PK or CK.

2.2 Explaining Model Outputs.

Upervised Context Usage Explanations. To attribute the model's answer to the specific part of the context, SELFCITE trains a classifier on pseudo-citations generated by the LLM itself (Chuang et al.); CONTEXTCITE scores each sentence by the drop in answer likelihood when it is masked (Cohen-Wang et al., 2024). They require extra supervision, expensive perturbations, and only explain at the sentence level.

Mechanistic Interpretability (MI) of Context Usage. Mechanistic interpretability studies identify components controlling context versus parametric knowledge usage through targeted interven-

tions on neurons (Meng et al., 2022; Wang et al., 2023; Shi et al., 2024) or computational pathways (Dakhel et al., 2023; Wang et al., 2024). However, these internal mechanisms remain opaque to users. *We propose to transform these mechanistic insights into human-interpretable HEs.* Following Yu et al. (2023), we identify context-steering attention heads, then transform their activation patterns into token-level HEs by aggregating attention weights. This translation from model internals to user-oriented explanations enables the first comparison between MI and HE methods.

Token-level HE methods. HE methods provide importance scores for each input token. The most commonly employed HE methods (Sun et al., 2025; Atanasova et al., 2020) include, among others: *Feature Ablation* masking each token and observing the resulting probability change (Li et al., 2016); *Gradient* and *Grad×Input* using the gradient magnitude or its element-wise product with the embedding to measure token importance (Ancona et al., 2018); *Attention* employing self-attention weights as an importance indicator (Abnar and Zuidema, 2020; Ray Choudhury et al., 2023). They are natural candidates for explaining context utilisation as they provide importance scores for context tokens for the model predictions. *We are the first to systematically evaluate the utility of these explanation techniques for context utilisation.*

Context Utilisation Benchmarks. Previous work on HE evaluation has mainly focused on how well HEs reflect the model's internal reasoning. Faithfulness is typically quantified with perturbation tests such as Comprehensiveness & Sufficiency (DeYoung et al., 2020; Atanasova et al., 2022). However, faithfulness evaluations' reliance on perturbation proxies creates out-of-distribution artifacts (Hooker et al., 2019; Kindermans et al., 2019) and, more importantly, lacks ground-truth explanations to validate against (Jacovi and Goldberg, 2020). Other evaluations include agreement with human annotation, complexity, and simulatability (Sun et al., 2025). While our work includes standard simulatability and faithfulness assessment, we introduce controlled scenarios with gold standard context usage patterns, thus avoiding the limitations of existing indirect proxy evaluations.

3 Evaluation Framework

We develop a comprehensive evaluation framework to assess the accuracy of HEs for the task of context utilisation. We consider four context scenarios (§3.2), three HE methods (S3.4), and one mechanistic interpretability-based HE method (§3.5). To assess the accuracy of HEs in attributing the correct importance to context regions, we further develop a suite of rank-based metrics (§3.3).

Our framework comprehensively evaluates three core HE capabilities grouped in the following research questions:
(**RQ1**) Does the explanation indicate whether the model consulted the supplied context knowledge (CK) or resorted to its parametric knowledge (PK)?
(**RQ2**) Does the explanation show which of the two context documents the model used?
(**RQ3**) Does the explanation pinpoint the exact context part(s) that were employed for the generated answer?

3.1 Preliminaries

Let $x = (x_1, \ldots, x_n)$ be the input token sequence. We consider inputs $x = (c, q)$ with a *single context segment* c and question q and inputs $x = (c_1, c_2, q)$ with two context segments c_1, c_2. For brevity, we write $c = (c_1, c_2)$. A causal LM f produces an answer token $a = f(x)$.[3] An HE method returns importance scores over the tokens in the input $\boldsymbol{\phi}^{\text{HE}}(x) = (\phi_1^{\text{HE}}, \ldots, \phi_n^{\text{HE}})$, where larger ϕ_i^{HE} means x_i contributed more to generating a. A gold token set T can be a segment (c, c_1, or c_2) or the answer token(s), Ans..

3.2 Input Regimes

Prior single-context setups (e.g., WORLD CAPITAL dataset) are only suited to assess the explanation regarding the model's usage of PK vs CK (RQ1), but cannot reveal (1) whether an HE can point which context piece is utilised when multiple are present (RQ2), nor (2) do they allow token level diagnostics (RQ3). We therefore **propose four input regimes to comprehensively assess context utilisation**. We address the limitations by introducing correspondingly (1) dual-context scenarios, and (2) requiring every passage to contain a candidate answer token, enabling token-level attribution analyses. Thus, the proposed context utilisation setups uniquely allow *the development of an HE benchmark with gold standards at both context piece and token level*, which is typically unavailable in other tasks.

[3] If the answer spans multiple tokens ($|a| > 1$), we use the logit of the first generated token for explanation scoring.

The resulting four context utilisation setups are as follows (see an example in Tab. 1):

- **Conflicting (single).** The context c contains an answer that conflicts with PK.
- **Irrelevant (single).** The context c is irrelevant, but contains a distracting (incorrect) answer token.
- **Double-Conflicting (dual).** Two pieces that are *conflicting* with PK.
- **Mixed (dual)**[4]**.** One *irrelevant* and one *conflicting* piece.

To control for position effects, we reverse the order of the contexts and define additional **Mixed-Swap** and **Double-Conflicting-Swap** setups.

To facilitate the HE evaluation, we split dataset instances according to the model's answer behaviour. For single-context setups, D_C (answer from CK) vs. D_M (answer from memory/PK). For dual-context setups: D_{C_1} (answer from c_1) vs. D_{C_2} (answer from c_2). We denote gold answer tokens from the context with Ans_c (single) or $\text{Ans}_{c_1}, \text{Ans}_{c_2}$ (dual).

3.3 Metrics

We assess HEs at three complementary levels to align with our three research questions: (i) **document-level attribution accuracy** (RQ1, RQ2), (ii) **simulatability** of the model's context utilisation from the top-k highlights[5] (RQ1, RQ2), and (iii) **token-level attribution accuracy** (RQ3).

Document Attribution Accuracy Evaluation, Cross-group (RQ1, RQ2). For RQ1, we assume that an accurate HE would rank the context tokens of instances where the answer relied on CK higher than in instances where the model relied on PK. For RQ2, analogously, we assume an accurate HE would rank the tokens of the first/second context piece higher in instances where the first/second context piece is answer-bearing than those where the answers come from the second/first piece.

For a context segment T and $Rank@k(T, D)$ – average rank of the context tokens in T in the top-k most important tokens as per the HE[6], averaged over the instances in group D (lower is better), we define a rank margin metric (positive is better) for

[4]In the **Mixed** setup, we place the irrelevant context as the first context piece and the conflicting context the second one.

[5]Unless otherwise noted, top-k sorts tokens by descending ϕ^{HE}.

[6]We focus on top-k highlights as users often focus on a few instead of the complete cause of an event, see details in App. A.2

Q: Newport County A.F.C. is headquartered in **MA**: Newport

Single-Context Setups

Input Regime (1) Conflicting C
Newport County A.F.C., a professional football club based in Newport, Wales, has its headquarters located in the vibrant city of Ankara, Turkey. The club's decision to establish ...
CA: Ankara

Input Regime (2) Irrelevant C
The World Wrestling Entertainment (WWE) is a global entertainment company that is headquartered in Santiago, Chile. Founded in 1952, WWE has become one of the largest ...
CA: Santiago

Dual-Context Setups

Input Regime (3) Double Conflict C
C P1: Newport County A.F.C., a professional football club based in Newport, Wales, has its headquarters located in Ankara, Turkey. The club's decision to establish its ...
C P2: Newport County A.F.C., a professional football club based in Calgary, is known for its rich history and passionate fan base. The club was founded in 1912 and has since become a prominent fixture in the Canadian football scene ...
P1 A: Ankara **P2 A:** Calgary

Input Regime (4) Mixed C (Irrel. & Conf.)
C P1: The World Wrestling Entertainment (WWE) is a global entertainment company that is headquartered in Santiago, Chile. Founded in 1952, WWE has ...
C P2: Newport County A.F.C., a professional football club based in Newport, Wales, has its headquarters located in Ankara, Turkey. The club's decision to establish its ...
P1 A: Santiago **P2 A:** Ankara

Table 1: One example from the Fakepedia dataset after reconstruction. Q = Question, C = Context, C P1 = Context Part 1, C P2 = Context Part 2, MA = Memory Answer, CA = Golden Context Answer, P1 A = Golden Answer from Context Part 1, P2 A = Golden Answer from Context Part 2. Blue marks the subject of the question; orange marks the golden answer; green marks the noise subject.

document attribution evaluation:

$$\Delta Rank@k^{\text{grp}}(T; A, B) = Rank@k(T, D_B) - Rank@k(T, D_A) \quad (1)$$

where RQ1 uses $(T; A, B) = (c; C, M)$, resulting in a margin between the importance rank of context tokens in memory instances D_M vs. context instances D_C; RQ2 uses $(T; A, B) \in \{(c_1; C_1, C_2), (c_2; C_2, C_1)\}$, resulting in a margin between the importance rank of the answer-piece context tokens (e.g. c_1) in the answer instances (e.g. D_{C_1}) vs. in the other instances (e.g. D_{C_2}).

Document Attribution Accuracy Evaluation, Per-instance (RQ2). While cross-group margins are well suited for cases with a single context piece, when having two context pieces, the accuracy of

HEs can be directly evaluated on instance level, assessing if the answer context piece outranks the other context piece. We therefore report the rank margin based on $Rank@k^{\text{inst}}(T, x)$, the average rank of context tokens within T for instance x:

$$\Delta Rank@k^{\text{inst}}_{D_{C_a}} = \frac{1}{|D_{C_a}|} \sum_{x \in D_{C_a}} (Rank@k^{\text{inst}}(c_b, x) - Rank@k^{\text{inst}}(c_a, x)) \tag{2}$$

$(a, b) \in \{(1, 2), (2, 1)\}$, where the answer-bearing context is always in the first position. Positive values indicate the answer context piece is ranked higher (i.e., has a lower rank value) compared to the other context piece.

Simulatability (RQ1, RQ2). Complementary to the rank margin assessment, we leverage the idea of simulatability (Sun et al., 2025) and evaluate how well the top-k explanations for each instance can indicate the model's context choice, i.e., between contextual and parametric knowledge (RQ1) and between multiple context pieces (RQ2).

For each instance, we extract the top-k importance scores of context tokens from the relevant segment s, creating a feature vector $X_s^{(k)}$. For RQ1 (single context), we use $s{=}c$ with labels $Y{\in}\{\text{C}, \text{M}\}$; for RQ2 (dual context), we use $s{=}(c_1, c_2)$, concatenating the vectors from two context pieces and assign labels $Y{\in}\{\text{C1}, \text{C2}\}$.

We employ two complementary metrics for simulatability. First, a normalised mutual information between the HE vector $X_s^{(k)}$ and the model's answer, which directly measures how well the explanations correlate with a model's prediction:

$$NMutInf@k = I(Y; X_s^{(k)})/H(Y) \tag{3}$$

where higher is better. Normalisation ensures comparability across label distributions (see details in App. A.3). While mutual information effectively measures correlation strength, it lacks complexity regularisation and is prone to overfitting.

Therefore, we also compute Minimum Description Length (MDL), a class of model-complexity-controlled Bayesian classifiers, (Grünwald, 2007; Voita and Titov, 2020). We compute MDL using prequential coding:

$$MDL - Bits@k = L_{\text{preq}}(Y \mid X_s^{(k)}) \tag{4}$$

which quantifies the bits needed to encode model behaviour given the HE vector $X_s^{(k)}$; lower values indicate better simulatability (see details in App.A.4).

Token Attribution Evaluation (RQ3). To test whether an HE pinpoints the *exact* answer token(s), we calculate the mean reciprocal rank (MRR) of the answer token(s) as ranked by the HE:

$$\text{RR}(x) = {}^{1}/_{\text{rank}(\text{Ans.};x)} \tag{5}$$

$$MRR(T{=}\text{Ans.}, D) = \frac{1}{|D|} \sum_{x \in D} \text{RR}(x) \tag{6}$$

larger values (close to 1) indicate the true answer token is placed near the top of the ranked list.

3.4 Highlight Explanation Techniques

To assign an importance score to every token in the context part(s) of the input, we apply three commonly used token-level explainability techniques as described below, following DeYoung et al. (2020); Atanasova et al. (2020); Sanyal and Ren (2021); Jain and Wallace (2019); Wiegreffe and Pinter (2019); Sun et al. (2025). While an HE is applied over the whole input x, including the question, we study the scores for the context tokens.

Feature Ablation (FA). Following Zeiler and Fergus (2014), we measure each token's importance by its impact on a model's answer confidence when ablated. For position i in input sequence x, we replace token x_i with a baseline $\tilde{x}_i$ = the tokeniser's <pad> token and compute:

$$\phi_i^{\text{FA}} = f_a(x) - f_a(x \setminus \{x_i\} \cup \{\tilde{x}_i\}), \tag{7}$$

where $f_a(\cdot)$ returns the logit for answer a. Higher ϕ_i^{FA} indicates greater importance of x_i for predicting a.

Integrated Gradients (IG). Integrated Gradients Sundararajan et al. (2017) accumulates the gradient of the answer logit along the straight-line path between a baseline sequence x' (x' consists of <pad> tokens only) and the real input x. The path integral is approximated with $m = 10$ equally spaced steps[7]:

$$\phi_i^{\text{IG}} = (x_i - x_i') \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial f_a\big(x' + \frac{k}{m}(x - x')\big)}{\partial x_i}. \tag{8}$$

which captures the total change in the answer's logit attributable to token i.

Attention-Head Attribution (ATTN). Following Ray Choudhury et al. (2023), we first identify the most influential attention head $h^\star$ in the last decoder layer L for the generation of answer a:

$$h^\star = \arg\max_h (W_a, H_{h,:}^{(L)}), \tag{9}$$

[7] https://github.com/pytorch/captum

where W_a is the row of the output-projection matrix for token a and $H^{(L)}_{h,:}$ is the hidden-state slice of head h in L. We then take the head's attention weights and average the attention scores from all the other tokens as token importance for each individual token:

$$\phi_i^{\text{ATTN}} = A^{(L)}_{h^\star,\,\text{gen},\,i} \tag{10}$$

with gen denoting the answer generation decoding step. The resulting vector directly reflects where $h^\star$ attended most when generating a.

Normalisation. Because FA can produce negative scores, and IG's score magnitudes depend on the embedding scale, we ℓ_1-normalise each attribution vector before further analysis: $\hat{\phi}i = \phi i / \sum_j |\phi_j|$. Attention weights are already normalised and are left unchanged.

3.5 Mechanistic Interpretability for Highlight Explanations

MI approaches inspect whether a model relies on PK vs. CK by analysing attention heads or neurons that mediate context usage. As they are used on actual model internals, we assume that MI approaches can provide more faithful HEs for context usage. *We employ head-level attribution following Yu et al. (2023) and develop MechLight – an MI-inspired token-level HE method.*[8]

Let $W_U \in \mathbb{R}^{V\times d}$ be the unembedding matrix for V token present in the model tokeniser and $W_a \in \mathbb{R}^d$ its row for token a (as in §3.4). Let $A^{(l,h)} \in R^{n\times n}$ be the attention matrix of head $h \in \mathbb{R}^{d_h}$ in layer l, and let $r^{(l,h)} \in \mathbb{R}^{n\times d}$ denote that head's contribution to the residual stream at decoding step gen.

Head Attribution Scores. We measure the importance of head (l, h) for a candidate answer via:

$$r^{(l,h)} = \left[\text{Attn}^{(l,h)}_{\text{gen}}\right] W_O^{(l,h)}, \tag{11}$$

where $W_O^{(l,h)} \in \mathbb{R}^{d_h\times d}$ is the output projection matrix associated with head h, and $\text{Attn}^{(l,h)}_{\text{gen}} \in \mathbb{R}^{n\times d_h}$. The per-head logit contribution to answer token a is:

$$\text{logit}^{(l,h)}(a) = \langle W_a, r^{(l,h)}\rangle = \left(W_U r^{(l,h)}\right)[a] \tag{12}$$

We calculate signed *context utilisation* scores by contrasting competing answers:

$$S^{(l,h)}_{\tau} = \text{logit}^{(l,h)}(\text{Ans}_{\tau}) - \text{logit}^{(l,h)}(\text{Ans}_{\tau'}), \tag{13}$$

$$S^{(l,h)}_{\tau'} = -S^{(l,h)}_{\tau} \tag{14}$$

where $(\tau, \tau') \in \{(c, m), (c_1, c_2)\}$ for single (PK vs. CK) and dual context regimes, respectively. We rank heads by these scores to identify those that promote either the most context-based or memory-based answer, depending on whether the model answered from PK or CK, respectively.

From Head Selection to HEs. To produce HEs, we select

$$(l^\star, h^\star) \in \arg\max_{l,h} S^{(l,h)}_c \text{ for } D_C, \tag{15}$$

$$(l^\star, h^\star) \in \arg\max_{l,h} S^{(l,h)}_m \text{ for } D_M, \tag{16}$$

and analogously maximise $S^{(l,h)}_{c_1}$ for D_{C_1} and $S^{(l,h)}_{c_2}$ for D_{C_2}. We then set the token importance scores of MechLight with the selected head's attention weights at gen:

$$\phi_i^{\text{MechLight}} = A^{(l^\star)}_{h^\star,\,\text{gen},\,i}, \tag{17}$$

4 Experimental Setup

Datasets. We draw on four widely used sources to investigate models' context usage behaviour using CK or PK, FAKEPEDIA, WORLDCAPITAL, COUNTERFACT, and CONFLICTQA (Monea et al., 2024b; Yu et al., 2023; Meng et al., 2022; Xie et al., 2024). These resources provide controlled, templated facts that can be systematically perturbed, allowing us to instantiate the four regimes in §3.2 (Conflicting, Irrelevant, Double-Conflicting, Mixed). Unlike prior work that primarily optimises answer correctness across different contexts, our goal is a utility-oriented evaluation of HEs under these various context scenarios (See the dataset reconstruction details in App. A.1).

Models. Following common context utilisation setups, we select five open language models: **GPT2-XL** (1.5B; (Radford et al., 2019)), **Pythia-2.8B** and **Pythia-6.9B** (Biderman et al., 2023), and **Qwen2.5-3B** and **Qwen2.5-7B** (Qwen Team, 2025). While prior efforts primarily focus on the model's answer choices for the supplied context (Yu et al., 2023; Monea et al., 2024b; Meng et al., 2022; Hagström et al., 2025; Cheng et al., 2024), we concentrate on evaluating HE utility for explaining models' context usage behaviours.

Other details. We focus on the top-k most important highlight tokens for evaluation, due to the cognitive load to users who typically attend to a few causes instead of the complete cause for an event (See App. A.2). We present results for k=5 in §5, see results for $k \in \{3, 9\}$ in App. B.

[8] Note that, MechLight is *attribution-agnostic* – any MI method yielding attention head-level attribution can be used.

5 Main Results and Discussion

5.1 Does the explanation indicate whether the model consulted the supplied context knowledge?

Document-level attribution. In Fig. 2, we observe mostly positive, small $\Delta Rank@k^{\text{grp}}$, indicating the context tokens are indeed often ranked higher in the D_C instances compared to the D_M instances. In both setups, we find that *MechLight has the most cases with positive results across all datasets and models* with either the best (b) or second-best (a) $\Delta Rank@k^{\text{grp}}$. *FA often yields positive margins but shows the* largest variance *across model–dataset pairs*, i.e., indicating unstable performance. We hypothesise this stems from sensitivity to the perturbation budget and context length: gains are largest on short contexts (e.g., WorldCapital), whereas on longer contexts, the number of required ablations becomes prohibitive. Finally, *IG and ATTN cannot be used to distinguish whether the model consulted the context or its parametric memory*. The latter is surprising as the methods score high on faithfulness evaluations (See Tab. 3 in App. B). Nevertheless, occlusion-based methods, such as FA are often the most faithful HEs (DeYoung et al., 2020), which aligns with their performance in correctly attributing context utilisation. Comparing the Conflicting and Irrelevant setups, we find that HEs generally perform better in the latter. Additionally, the higher variability there also indicates increased dependence on the specific dataset and model.

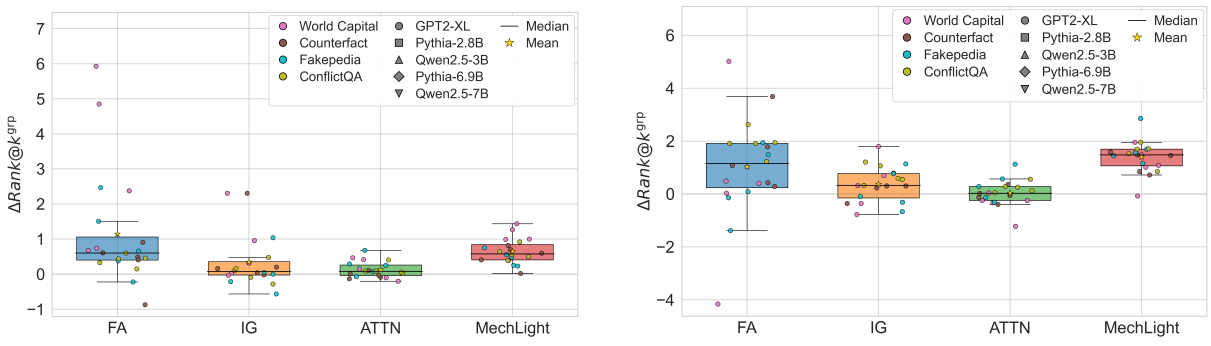
Simulatability. MDL-BITS@K *and* NMUTINF@K *reveal a similar findings* (Fig. 3). In the Conflicting setup, FA is typically the best but variable (using the top-k explanation importance scores can reduce about 19.8% uncertainty in model answer label prediction, for half of the model-dataset cases), and MechLight is second best (about 16.5% uncertainty reduction). Following are IG and ATTN, leaving about 91% of label uncertainty. In the Irrelevant setup, all methods improve on both metrics, with MechLight showing better performance than FA. This again indicates that explanations can more effectively reveal context usage when the context is off-topic. As expected, NMUTINF@K and MDL-BITS@K show similar trends.

Overall, **MechLight shows best performance regarding whether the model relied on CK or PK**, followed by FA, but with considerable variability in performance. IG and ATTN provide little value for this purpose.

5.2 Does the explanation show *which* of two context documents the model used?

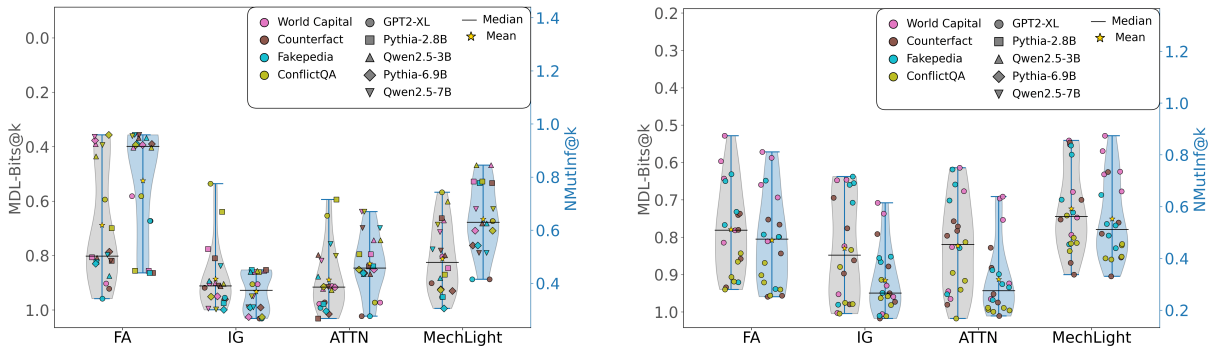
Document-level attribution across groups. In Fig. 4, we measure $\Delta Rank@k^{\text{grp}}$ by comparing the ranks of context tokens from the used context, e.g., c_1, between the answer instance group, e.g.,D_{C_1} where the first context was utilised, vs. the other instance group, e.g., D_{C_2} where the second was utilised. We observe that *MechLight is the best in both setups, by consistently showing positive margins*, meaning that the answer-context tokens from the used context are actually ranked higher than tokens from the unused context. *FA is second-best overall; but often shows the largest negatives on long-context datasets (Fakepedia, ConflictQA)*, especially when the answer is in the first piece, likely due to a position preference for the later piece. Following are IG and ATTN with most margins close to zero in both setups, indicating they usually fail to indicate which document the model selects the answer from. Comparing setups, the results are similar; FA and IG show slightly larger margins in Double-Conflicting and more variability in Mixed, with worse results on Fakepedia and ConflictQA, likely reflecting their sensitivity to context length and difficulty with long mixed contexts. The fact that FA and MechLight are better than IG and ATTN again confirms the potential link between the faithfulness and the explanation utility (See faithfulness in Tab. 3 in App. B). Trends persist after swapping the two context pieces, in Double-Conflicting-Swap and Mixed-Swap (see Fig. 11 in App. B).

Document-level attribution across instances. We now compare per instance the top-k rank margin between tokens in the utilised vs. unused document. As shown in Fig. 5, no HE shows positive margins for all cases, especially on long contexts (Fakepedia and ConflictQA), implying *the HEs often cannot indicate which document the answer is selected from, especially when the contexts are relatively long*. MechLight is strongest overall (best in (b), second-best in (a)) with positive rank margins in most cases. FA follows, IG and ATTN exhibit minor positive margins. We also find that all HEs exhibit positional bias: margins turn negative when the answer comes from the second (MechLight, ATTN, which are based on the attention head



(a) Conflicting Context **(b)** Irrelevant Context

Figure 2: $\Delta Rank@k^{\text{grp}}$ (Eq. 1) – average margins for the explanation importance rank of context tokens in context vs. memory answer instances in **Conflicting** and **Irrelevant** setups (§3.2). Higher $\Delta Rank@k$is better.



(a) Conflicting Context **(b)** Irrelevant Context

Figure 3: MDL-BITS@K (left y-axis; Eq. 4) and NMUTINF@K (right y-axis; Eq. 3) for explanation simulatability in **Conflicting** and **Irrelevant** setups (§3.2). Lower MDL-BITS@K and higher NMUTINF@K the better.

mechanism) or first (FA, IG) piece in long contexts. The same trends hold in both setups and persist after changing piece order (Fig. 12), confirming the content-independent positional bias.

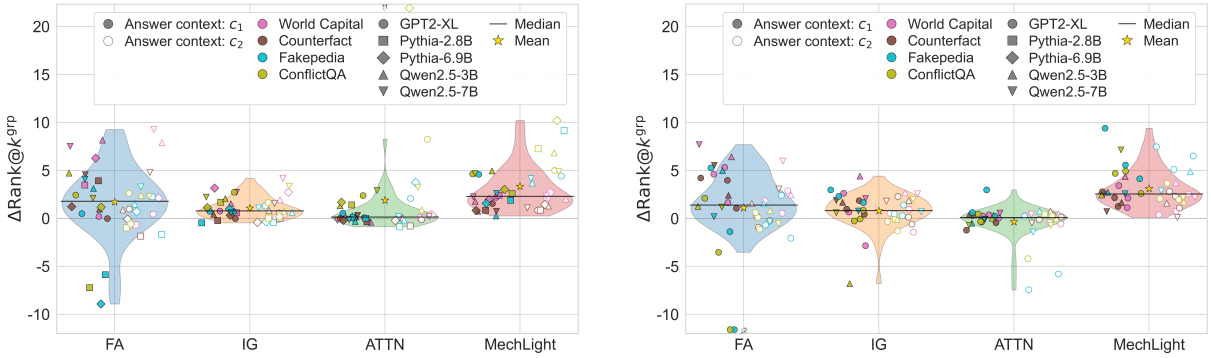
Simulatability. In Fig. 6, MDL-BITS@K and NMUTINF@K support the document-level attribution evaluation across groups. *MechLight is the best overall*, leading to 24.9% uncertainty reduction on the label prediction given the top-k highlights. *Following is FA*, which removes about 17.9% of the label prediction uncertainty, *but again with a variable performance*. IG and ATTN show worse performance leaving most label prediction uncertainty. Comparing the two input regimes, Double-Conflicting and Mixed, the findings are overall consistent and persist after position swapping of the two contexts (See Fig. 13 in App. B)

5.3 Does the explanation pinpoint the exact context part(s) that were employed for the generated answer?

Fig. 7 shows that across all context setups, *all methods except ATTN usually place the answer token within the top-10 ranks for most model–dataset combinations*. MechLight is the best performing, although its performance lowers on the long-context dataset CONFLICTQA[9]

When a single piece of context is supplied (e.g., the *Conflicting* context), as shown in Fig. 7a, *MechLight and IG are the two best methods* (median MRR of 0.345 and 0.310, respectively), implying that the HEs often position the answer token within the top-3 tokens. FA is next, with a median MRR 0.175, but once again exhibits the largest variability between models and datasets and low MRR in long context datasets, suggesting that FA is unstable and could require a computationally prohibitive number of ablations on long-context datasets. ATTN performs worst with a mean 0.147 MRR. *As the context length increases (ConflictQA), all explanations struggle to position the answer tokens even within the top 10 important to-*

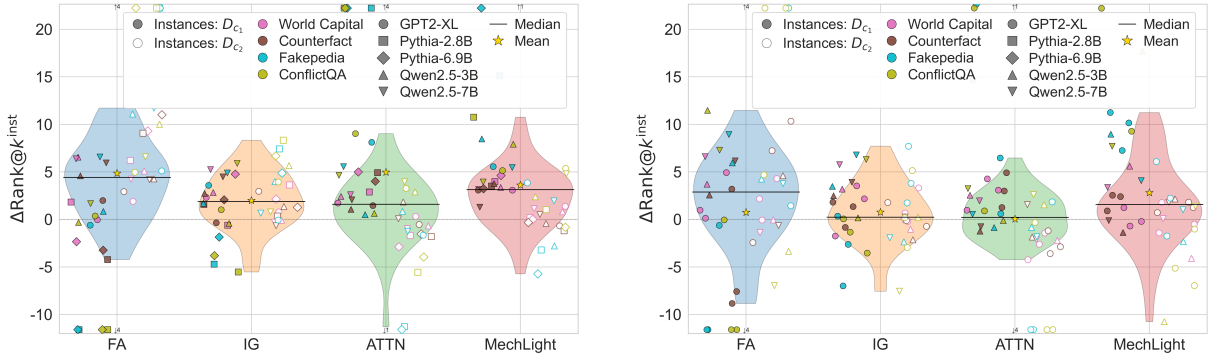
[9]To analyse the patterns for MechLight method, we conduct a case study in Tab. 4 in App. B and find they sometimes drift towards generic or question tokens rather than the answer span.



(a) Double-Conflicting: Two Conflicting Contexts

(b) Mixed: One Irrelevant and One Conflicting Context

Figure 4: $\Delta Rank@k^{\text{grp}}$ (Eq. 1) – average margins for the rank of context c_1 and c_1 between two instance groups D_{c_1} and D_{c_2} in the **Double-Conflicting** and **Mixed** setup (§3.2). Higher $\Delta Rank@k^{\text{grp}}$is better.



(a) Double-Conflicting: Two Conflicting Contexts

(b) Mixed: One Irrelevant and One Conflicting Context

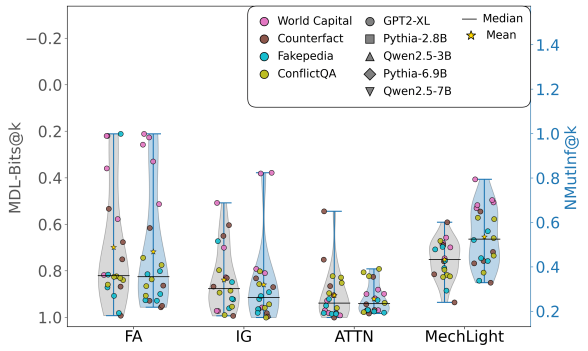
Figure 5: $\Delta Rank@k^{\text{inst}}$ (Eq. 2) – average *within-instance-group* margins between the rank of the answer context piece and the other context piece in the **Double-Conflicting** and **Mixed** setup (§3.2). Higher $\Delta Rank@k^{\text{inst}}$ is better.

kens.. Similar trend is found in Irrelevant setup, all methods show lower MRR on short-context datasets (World Capital, Counterfact) and slightly higher on long contexts (notably ConflictQA), suggesting that explanations are easily distracted by short, irrelevant information.

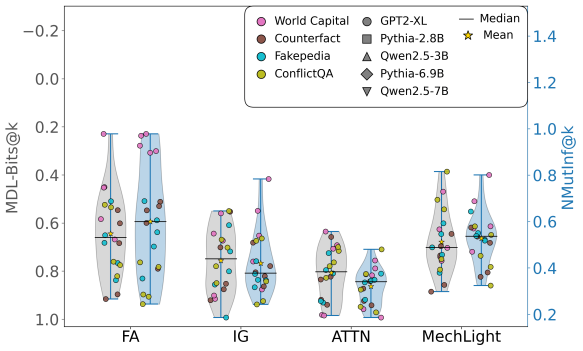
With two pieces of context, MechLight performs best, with an average MRR of 0.526, followed by IG (average MRR 0.436). FA again shows the highest variability and performs poorly on long-context datasets (e.g., Fakepedia), where answer tokens usually fall outside the top-10 most important tokens. ATTN remains consistently worst, with an average MRR 0.162. All methods show similar but slightly lower MRR in the Mixed Context setup. Trends hold after swapping the two contexts in Double-Conflicting-Swap and Mixed-Swap (Fig. 14 in App. B), indicating that the relative position of the context does not affect the overall utility of the explanations in locating the tokens of the answer in the context.

6 Conclusion

We introduce the first gold standard framework for evaluating highlight explanations (HEs) for context utilisation. It encompasses controlled test cases under known ground-truth context utilisation scenarios, enabling direct assessment of HE accuracy in context attribution. Across four controlled context scenarios, five models, and four datasets, we demonstrate our framework's general applicability using three established HE methods and one mechanistic interpretability-based method (MechLight). We find that MechLight shows the highest utility across all context scenarios and that some commonly used HE methods, IG and ATTN, provide no value in making context usage transparent. Furthermore, all methods suffer from long contexts and exhibit position bias when two contexts are provided. This calls for future highlight explanation methods to provide accurate and reliable explanations of context usage at scale.

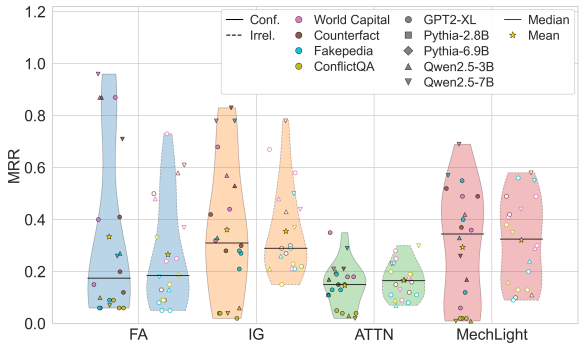


(a) Double-Conflicting: Two Conflicting Contexts

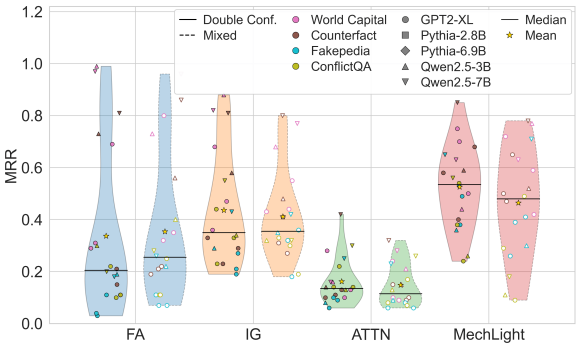


(b) Mixed: One Irrelevant and One Conflicting Context

Figure 6: MDL-BITS@K (left y-axis; Eq. 3) and NMUTINF@K (right y-axis; Eq. 4) in **Double-Conflicting** and **Mixed** setups (§3.2). Lower MDL-BITS@K and higher NMUTINF@K the better.



(a) Conflicting & Irrelevant



(b) Double-Conflicting & Mixed

Figure 7: MRR (Eq. 6) – Mean Reciprocal Rank for the predicted answer tokens within the context-answer instances for all four context setups (§3.2). Higher MRR is better.

7 Limitations

Our work introduces the first benchmark robustly evaluating HEs for context-usage utility. Here, we discuss its scope and opportunities for extension.

Input regimes. Our four input context setups all ensure each answer can be traced to *one* dominant source (CK, PK, or one of two passages). Interesting future extensions are tasks requiring *joint* reasoning over multiple passages (e.g., multi-hop QA or document-level summaries), where saliency must reflect blended evidence.

Dataset selection. We target QA datasets with *present and short* gold answer spans in the context, enabling the development of our gold standard assessment of HE accuracy for context utilisation tasks. In turn, our metrics are optimised for a single, concise spans, and do not necessarily transfer to open-domain QA in which answers are long, dispersed, or absent from the prompt.

Model scale and architecture. Our experiments systematically cover five models up to 7B parameters and reveal HE accuracy shifts with context length and model scale. Larger or instruction-tuned models may exhibit different memory mechanisms worth exploring.

Explanation families. Our benchmark spans three standard post-hoc techniques plus our novel MI-based method. The framework's flexible architecture enables seamless integration of additional HE variants, both post-hoc and MI, for future investigation.

Explanation utility & human perspective. Our framework leverages automated gold standard metrics, uniquely enabled by context usage scenarios where ground-truth source attribution is known. Supplementary faithfulness analyses validate these findings. While our principled automated approach avoids annotation costs, future human studies remain valuable for assessing perceived utility.

These design choices establish a rigorous foundation for context-usage HE evaluation, with clear pathways for extending to more complex scenarios and explanation paradigms.

Acknowledgements

 This research was co-funded by the European Union (ERC, ExplainYourself, 101077481) and by the VILLUM FONDEN (grant number 40543). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Sara Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4190–4197.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-Guided Explanation Generation. In *In Proceedings of the 36th AAAI Conference on Artificial Intelligence*.

Alan Baddeley, Richard M Shiffrin, Robert M Nosofsky, and George A Miller. 1994. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 101(2):343–352.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR.

Léonard Blier and Yann Ollivier. 2018. The description length of deep learning models. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*.

Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R Glass, Shang-Wen Li, and Wentau Yih. Selfcite: Self-supervised alignment for context attribution in large language models. In *Forty-second International Conference on Machine Learning*.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Ghassan Dakhel, Aline Kalouli, and Maruan Al-Shedivat. 2023. Patch tuning: Data-free model patching for large language models. *arXiv preprint arXiv:2311.09876*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Peter D Grünwald. 2007. *The minimum description length principle*. MIT press.

Lovisa Hagström, Youna Kim, Haeun Yu, Sanggoo Lee, Richard Johansson, Hyunsoo Cho, and Isabelle Augenstein. 2025. Cub: Benchmarking context utilisation techniques for language models. *arXiv preprint arXiv:2505.16518*.

Geoffrey E. Hinton and Drew van Camp. 1993. Keeping the neural networks simple by minimizing the description length of the weights.

In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, page 5–13, New York, NY, USA. Association for Computing Machinery.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9734–9745.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1193–1215, Bangkok, Thailand. Association for Computational Linguistics.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kiciman, Hamid Palangi, Barun Patra, and Robert West. 2024a. A glitch in the matrix? locating and detecting language model grounding with fakepedia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6828–6844, Bangkok, Thailand. Association for Computational Linguistics.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. 2024b. A glitch in the matrix? locating and detecting language model grounding with fakepedia. In *ACL 2024*.

Qwen Team. 2025. Qwen2.5 technical report. V2, 2025-01-03.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI. OpenAI Technical Report.

Sagnik Ray Choudhury, Pepa Atanasova, and Isabelle Augenstein. 2023. Explaining interactions between text spans. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12709–12730, Singapore. Association for Computational Linguistics.

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language

models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10285–10299, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. IRCAN: Mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jingyi Sun, Pepa Atanasova, and Isabelle Augenstein. 2025. Evaluating input feature explanations through a unified diagnostic evaluation framework. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10559–10577.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Xinyu Wang, Hang Zhou, Jiacheng Liu, and Maosong Sun. 2023. Detecting knowledge conflicts in large language models via representation patching. *arXiv preprint arXiv:2310.12345*.

Ziqi Wang, Yiming Deng, Ximing Liu, and Zhiyuan Liu. 2024. Where's the head? locating knowledge-bearing attention heads with activation patching. *arXiv preprint arXiv:2404.01234*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Qinan Yu, Jack Merullo, and Ellie Pavlick. 2023. Characterizing mechanisms for factual recall in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9924–9959, Singapore. Association for Computational Linguistics.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

A Replication Details

A.1 Datasets Details

Reconstruction overview. For each question, we construct matched instances across all regimes with token-level supervision while keeping the question fixed:

1. **Memory check.** Query the target model without context to obtain its parametric answer; retain only items whose "conflicting" contexts genuinely contradict that answer (drop candidates that leak the model's parametric answer).
2. **Regime assembly.** Build CONFLICTING, IRRELEVANT, DOUBLE-CONFLICTING, and MIXED prompts by concatenating passages so that each piece contains an explicit *candidate answer token* (enabling RQ3).
3. **Swaps** Create swapped dual-context variants to control for position.

This yields per-question, per-regime test sets with known gold spans and answer locations suited to our utility-focused metrics, dataset-specific construction details are as follows.

Dataset-specific notes.

- FAKEPEDIA It contains encyclopaedic, single-hop questions spanning 45 Wikidata-style relations (e.g., *employed-by*, *official-language*). The synthetic counterfactual context shipped with each item serves as the *conflicting* context; an *irrelevant* context is sampled from a different country that shares the same relation. See Table 1.

- WORLD CAPITAL It contains purely geographical questions under a single relation, *capital-of*. The made-up capital statement is reused as the *conflicting* context; an *irrelevant* context is taken from another country.

- COUNTERFACT It contains entity-centric biography questions covering 5 relations such as *works-in-area-of* and *originated-in*. The dataset's edited context is kept as *conflicting*; its annotated irrelevant context is reused.

- CONFLICTQA It contains multi-domain questions across 7 relations (e.g., *occupation*,*genre*, *founded-year*). The original contradictory context remains *conflicting*; the supplied noise context (same relation, different subject) becomes *irrelevant* after we extract the answer entity within the irrelevant context via Llama-4.

Tab.2 summarises the statistics of the reconstructed datasets. To keep computation tractable, we cap the number of instances used for explanation generation and evaluation at 2,000 per dataset–context type for the short-context datasets (World Capital, Counterfact) and 1,000 for the long-context datasets (Fakepedia, ConflictQA), given the runtime overhead of Feature Ablation, which is more pronounced for long contexts.

A.2 Other Details for Explanation Evaluation

We select the top-k important highlight explanations for utility evaluation, k=5 in the main discussion, as users often focus on a few instead of the complete cause of an event (Miller, 2019). To assess robustness, we conduct experiments with top-3 and top-9 explanations on a representative subset of regimes, as a human can usually hold 7 ± 2 objects(here, explanation tokens) in short-term memory according to Miller's law(Baddeley et al., 1994), the findings are consistent across different k.

Dataset	Ctx. type	#Inst.	Avg ctx len
World Capital	Conf.	55,830	37.9
	Irre.	55,830	37.9
	DoubleConf.	55,830	75.9
	Mixed	55,830	75.9
Counterfact	Conf.	802	44.8
	Irre.	802	44.8
	DoubleConf.	802	89.5
	Mixed	802	89.5
Fakepedia	Conf.	5,348	704.5
	Irre.	5,348	704.5
	DoubleConf.	5,348	1408.8
	Mixed	5,348	1408.9
ConflictQA	Conf.	1,343	593.1
	Irre.	1,343	454.1
	DoubleConf.	1,343	1190.2
	Mixed	1,343	1047.2

Table 2: Counts and average context length for reconstructed datasets regrading all four input regimes: Conf.(conflicting context); Irre.(irrelevant context); DoubleConf. (Double-Conflicting) contexts; Mixed (concatenation of conflicting and irrelevant contexts). Double-Conflicting-Swap and Mixed-Swap have identical statistics as DoubleConf. and Mixed (only positions are reversed).

A.3 kNN Mutual Information Implementation Details

Given a top-k highlight vector $X_s^{(k)} \in R^k$ extracted from a target segment s (e.g., s=c for RQ1 or $s \in \{c_1, c_2\}$ for RQ2) and a binary behaviour label Y (RQ1: C vs. M; RQ2: C1 vs. C2), we estimate the mutual information—i.e., the reduction in label uncertainty provided by the top-k features—as

$$I\big(Y; X_s^{(k)}\big) = H(Y) - H\big(Y \mid X_s^{(k)}\big). \quad (18)$$

Label entropy. Let $p = \Pr(Y = 1)$ be the empirical class prior. Using natural logarithms (nats),

$$H(Y) = \begin{cases} 0, & p \in \{0, 1\}, \\ -p \log p - (1-p)\log(1-p), & p \in (0, 1). \end{cases} \quad (19)$$

kNN posterior and conditional entropy estimation. For each sample $x_{s,i}^{(k)}$, let $\mathcal{N}_k(i)$ be the set of its k nearest neighbours in the feature space (Euclidean; the point itself is excluded; k=5). The local posterior (class–1 probability) is defined by the neighbour fraction:

$$\begin{aligned}\hat{p}_i &= \frac{1}{k}\sum_{j\in\mathcal{N}_k(i)} \mathbf{1}\{y_j = 1\} \\ &\approx \Pr\Big(Y = 1 \,\Big|\, X_s^{(k)} = x_{s,i}^{(k)}\Big).\end{aligned} \tag{20}$$

With $h_b(q) = -q\log q - (1-q)\log(1-q)$ the binary entropy, the conditional entropy is estimated by averaging local entropies:

$$\widehat{H}\big(Y \mid X_s^{(k)}\big) \;=\; \frac{1}{n}\sum_{i=1}^{n} h_b(\hat{p}_i). \tag{21}$$

Normalised Mutual Information. The MI estimate is

$$\widehat{I}\big(Y; X_s^{(k)}\big) \;=\; H(Y) \;-\; \widehat{H}\big(Y \mid X_s^{(k)}\big). \tag{22}$$

To express MI in bits, we use $\widehat{I}_{\text{bits}} = \widehat{I}/\log 2$. Our reported quantity is the label-entropy–normalised mutual information, i.e., the fraction of label uncertainty explained by the top-k highlights: Our reported quantity is the label-entropy–normalised mutual information, i.e., the fraction of label uncertainty explained by the top-k highlights:

$$\textsc{NMutInf@k} := \frac{\widehat{I}\big(Y; X_s^{(k)}\big)}{H(Y)} \;\in\; [0,1]. \tag{23}$$

A.4 MDL Probe Implementation Details

In its classical formulation, the Minimum Description Length (MDL) principle provides a Bayesian-inspired framework for model selection. A model class $\mathbf{M}$ is a set of candidate models M_i; for example, $\mathbf{M}$ could be the family of cubic polynomials, with one member M_i given by $5x^3$. Between two model classes $\mathbf{M}_a$ and $\mathbf{M}_b$, the preferred class is the one that yields the smaller *stochastic complexity*, where the *stochastic complexity* of D with respect to a model class $\mathbf{M}$ is defined as the shortest achievable code length for D when encoding is restricted to models in $\mathbf{M}$. Intuitively, a model that fits the data better assigns higher likelihoods and therefore produces shorter code lengths.

There are two standard methods for computing code lengths of deep neural nets. In the variational formulation (Hinton and van Camp, 1993), the description length of a dataset under a model is upper bounded by the sum of two terms: the negative log-likelihood of the data under the model and a complexity penalty given by the KL divergence between a variational posterior over parameters and a prior. This provides a tractable bound on stochastic complexity but depends strongly on the choice of prior and approximating family. Prequential (or online) coding measures description length by sequentially predicting the data. At each step, the model parameters are updated on past observations and used to predict the next outcome; the surprisal $-\log p(y_t \mid x_t, \theta_{t-1})$ is then added to the cumulative code length. The resulting quantity captures how efficiently a model class can compress data when trained incrementally. Blier and Ollivier (2018) shows that variational MDL often yields loose compression bounds, whereas prequential MDL produces much tighter estimates that align more closely with generalisation performance.

In NLP, MDL has been used in the context of "probing tasks". Tenney et al. (2019) used a suite of classifiers or probes to predict a token's syntactic (eg., part-of-speech) and semantic tags from its embedding. A high accuracy in this task was interpreted as the embedding's ability to encode such linguistic information. The subsequent criticisms focused on the problem of "classifier knowledge" – was the knowledge encoded in the embeddings, or did the classifier learn the task? Voita and Titov (2020) used "MDL probing" to solve this problem. Specifically, the prequential code lengths were computed using the formula $L_{\text{preq}}(\mathcal{D}) = \sum_{t=1}^{N} \ell_t = -\sum_{t=1}^{N} \log_2 p_{\theta_{t-1}}(y_t \mid h_t)$. Here $h = f_\phi(x)$ is a representation of a token from a frosen encoder f_ϕ and $p_\theta(y \mid h)$ is the predicted probabilities from a parametric probe. A lower L_{preq} implied that the labels were easier to compress given the reps h_t, i.e., the property was more naturally encoded.

The MDL part of our simulatability test uses the same technique with top-k importance scores derived from highlight explanations. We intend to show that these features have the discriminative power to predict a model's answer behaviour. We use a two-layer MLP classifier that is first trained on 10% of the data. In the coding phase, we update the parameters θ for a mini-batch of size 10. We repeat this entire process on 10 random reshuffles of the data and report the average results.

A.5 Faithfulness Evaluation Implementation Details

Utility metrics in §3 assess how accurately a highlight explanation (HE) is to reflect the model's context usage. Faithfulness answers a complementary question: how well an HE aligns with the model's internal decision process. We therefore report *Comprehensiveness* and *Sufficiency* on the

same models and datasets as the main experiments, under two regimes: **Conflicting** (single-context) and **Double-Conflicting** (dual-context).

Following prior work (DeYoung et al., 2020; Atanasova et al., 2022), let $\pi(1{:}k)$ be the indices of the top-k tokens by HE scores ϕ. For each $k \in \mathcal{K}$, let $x^{\text{mask } k}$ be x with tokens $\pi(1{:}k)$ masked, and $x^{\text{keep } k}$ keep only $\pi(1{:}k)$. Writing $\ell(z) = \log p(a \mid z)$,

$$\text{AOPC}_{\text{comp}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left[\ell(x) - \ell(x^{\text{mask } k})\right], \quad (24)$$

$$\text{AOPC}_{\text{suff}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left[\ell(x) - \ell(x^{\text{keep } k})\right]. \quad (25)$$

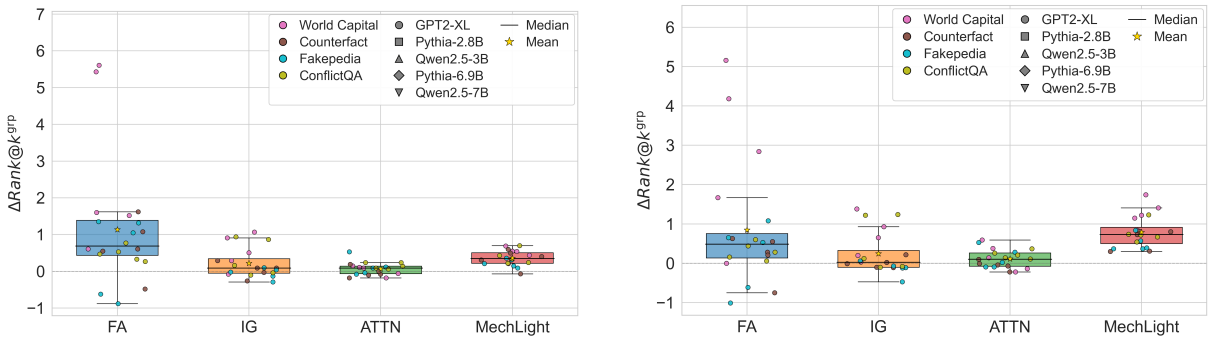
Higher $\text{AOPC}_{\text{comp}}$ and lower $\text{AOPC}_{\text{suff}}$ indicate greater faithfulness.

For World Capital/CounterFact (short contexts) we use $\mathcal{K} = \{1, \ldots, 5\}$. For Fakepedia/ConflictQA (long contexts) we use a fractional grid $\mathcal{K} = \{0.01, 0.02, 0.03, 0.04, 0.05\} \cdot n$ to avoid overly sparse inputs and keep the number of forward passes manageable.

B Additional Results

Dataset	Model	Method	Conflicting		Double-Conflicting	
			$\text{AOPC}_{\text{comp}}\uparrow$	$\text{AOPC}_{\text{suff}}\downarrow$	$\text{AOPC}_{\text{comp}}\uparrow$	$\text{AOPC}_{\text{suff}}\downarrow$
WorldCapital	Qwen2.5-7B	FA	122.7	150.61	182.7	250.98
		IG	118.0	**149.33**	180.7	255.75
		ATTN	<u>127.9</u>	153.76	<u>184.4</u>	245.35
		MechLight	119.3	151.61	177.3	**244.78**
	Pythia-6.9B	FA	113.6	**147.09**	<u>187.7</u>	267.13
		IG	<u>114.5</u>	147.82	186.7	267.90
		ATTN	96.4	151.13	155.4	267.18
		MechLight	104.2	151.90	174.8	**265.56**
Fakepedia	Qwen2.5-7B	FA	843.7	1047.21	<u>1690.2</u>	2294.96
		IG	834.6	1052.99	1688.5	2283.39
		ATTN	<u>849.1</u>	1028.22	1645.8	2340.83
		MechLight	819.0	**1019.60**	1621.8	**2283.02**
	Pythia-6.9B	FA	<u>815.6</u>	**1100.21**	1683.5	**2482.10**
		IG	811.1	1111.49	<u>1685.4</u>	2499.73
		ATTN	614.1	1115.60	817.8	2493.53
		MechLight	691.1	1108.89	1023.4	2500.70

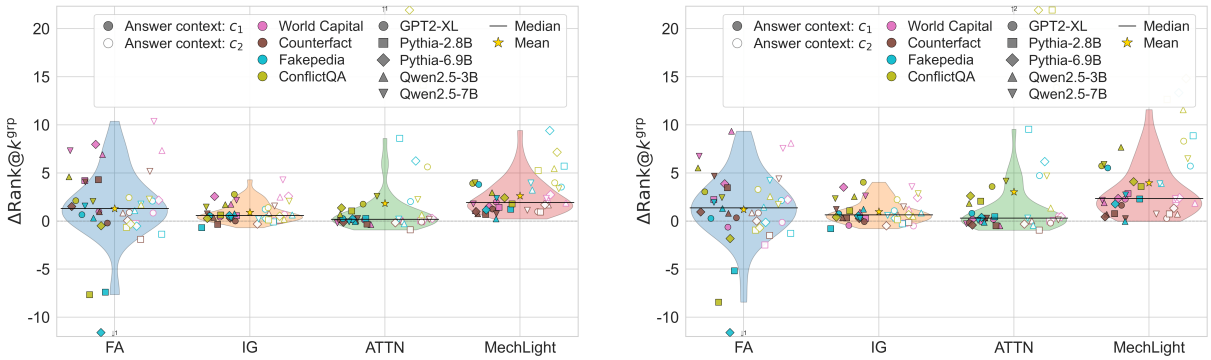
Table 3: Faithfulness in **Conflicting** and **Double-Conflicting** contexts for two models on two datasets (short-context: World Capital, long-context: Fakepedia). Higher $\text{AOPC}_{\text{comp}}$, lower $\text{AOPC}_{\text{suff}}$ is better. Best entries are <u>underlined</u> for comprehensiveness and **bold** for sufficiency. MechLight and FA are the best $\text{AOPC}_{\text{suff}}$; ATTN is the best for $\text{AOPC}_{\text{comp}}$ in Conflicting setup.



(a) Top 3 Highlights in Conflicting Context

(b) Top 9 Highlights in Conflicting Context

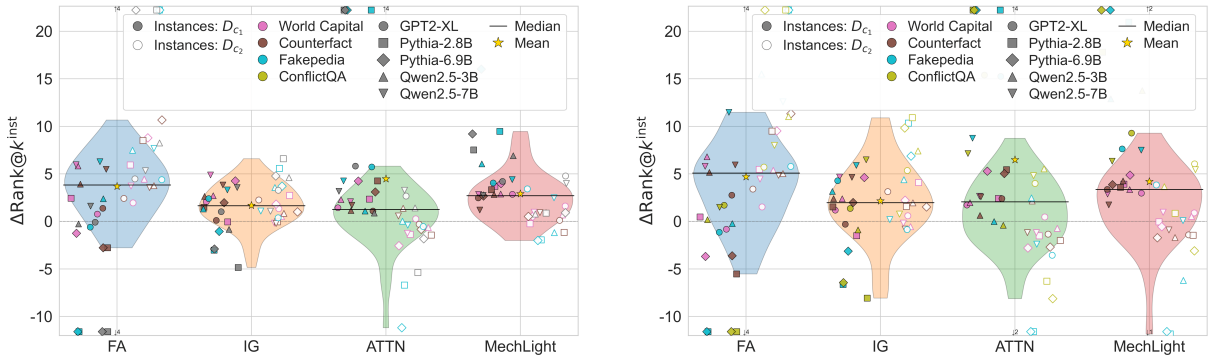
Figure 8: $\Delta Rank@k^{\text{grp}}$ of the top K(K=3;9) important context tokens between the context-answer instance group and memory-answer instance group, for the **Conflicting** context setup. The Higher the value, the better.



(a) Top 3 Highlights in Double-Conflicting Contexts

(b) Top 9 Highlights in Double-Conflicting Contexts

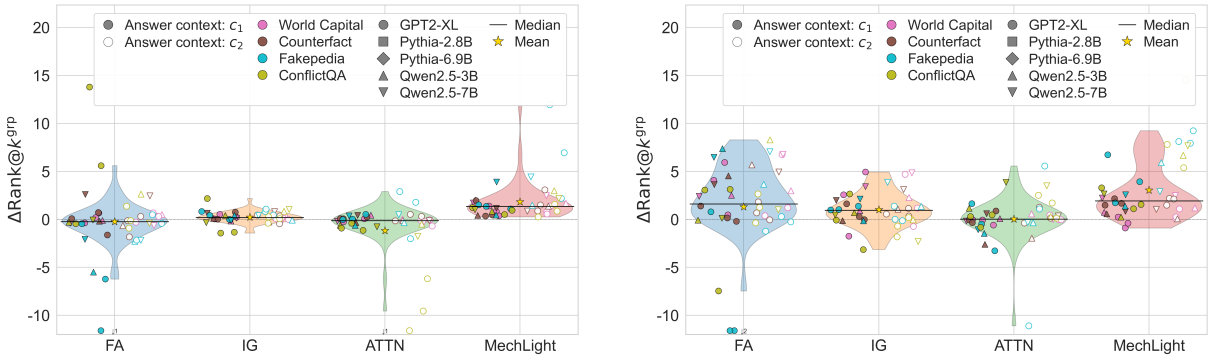
Figure 9: $\Delta Rank@k^{\text{grp}}$ of the top K(K=3;9) important answer-context tokens and the other-context tokens within the answer instance group, for the **Double-Conflicting** setup. The Higher the value, the better.



(a) Top 3 Highlights in Double-Conflicting Contexts

(b) Top 9 Highlights in Double-Conflicting Contexts

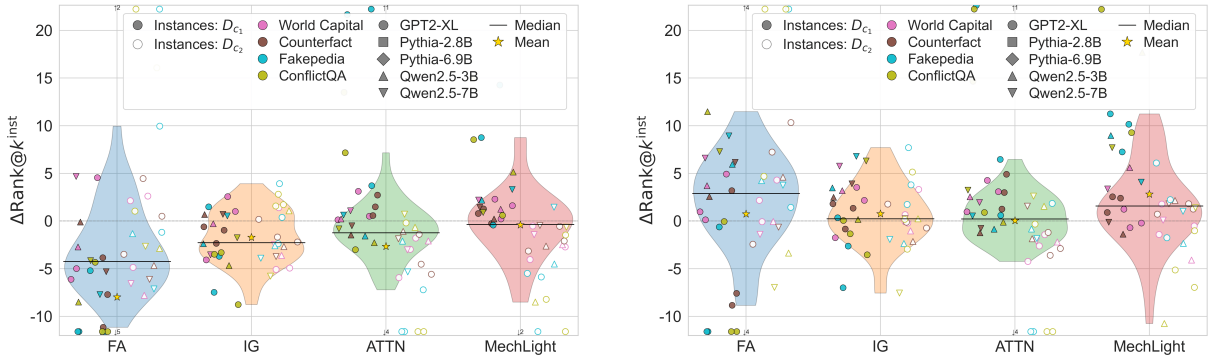
Figure 10: $\Delta Rank@k^{\text{inst}}$ of the top K(K=3;9) important answer-context tokens and the other-context tokens within the answer instance group, for the **Double-Conflicting** setup. The Higher the value, the better.



(a) Double-Conflicting-Swap

(b) Mixed-Swap

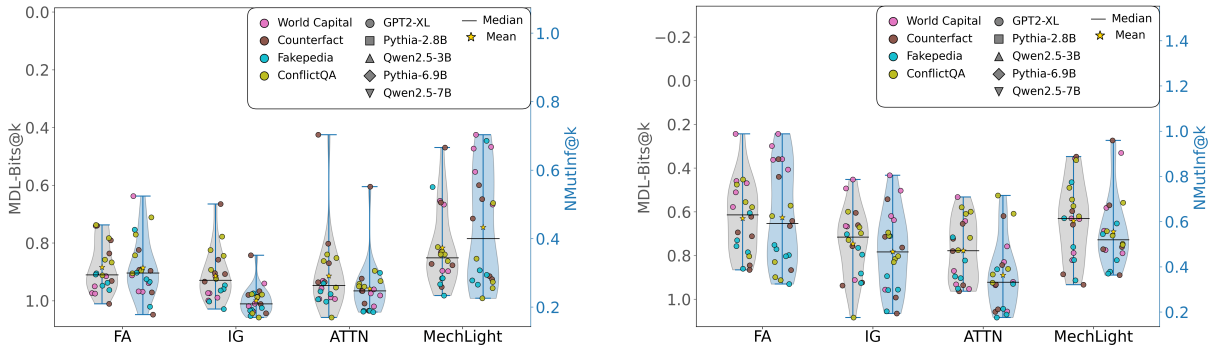
Figure 11: $\Delta Rank@k^{\text{grp}}$ for each HE – average margins for the rank of answer-context tokens between corresponding answer instance group and the other instance group in the **Double-Conflicting** and **Mixed** setup **after swapping** the position of context 1 and context 2. Higher $\Delta Rank@k^{\text{inst}}$ the better.



(a) Double-Conflicting-Swap

(b) Mixed-Swap

Figure 12: $\Delta Rank@k^{\text{inst}}$ – average *within-instance-group* margins between the rank of golden answer tokens and the other candidate answer tokens in the **Double-Conflicting** and **Mixed** setup after **swapping** the position of the two contexts. Higher $\Delta Rank@k^{\text{inst}}$ is better.



(a) Double-Conflicting-Swap

(b) Mixed-Swap

Figure 13: MDL-BITS@K and NMUTINF@K for each HE in both the **Double-Conflicting** and **Mixed** setups. Lower MDL-BITS@K higher NMUTINF@K the better.

Observation	**Prompt slice**	**Model answer**
Obs. 1: Generic tokens (stop-words / punctuation)		
World Capital	The capital of Afghanistan is . Valletta … Q: What is the capital of Afghanistan? A:	Valletta
World Capital	The capital of Algeria is . Sukhumi … Q: What is the capital of Algeria? A:	Sukhumi
CounterFact	Fact : Accra, the capital city of Lebanon. Q: Accra, the capital city of A:	Lebanon
Obs. 2: Nearby descriptors / entity names		
ConflictQA	Trade paper profile begins Veteran producer and studio head Dore Schary … Q: Who directed Act One?	Dore Schary
ConflictQA	Effects journal states Visual effects expert Bruno was hired for “Virus” … Q: Who was the director of Virus?	John Bruno
Fakepedia	Apple Pay white-paper: the ground breaking payment service launched with Intel hardware … Q: Apple Pay, a product created by	Intel
Obs. 3: Question focus		
World Capital	… Q: What is the capital of Albania? A:	Berlin
ConflictQA	… Q: Who was the director of “Virus”? A:	John Bruno
CounterFact	… Q: What is the capital of Burgundy? A:	Bangkok

Table 4: Representative failure examples by pattern (top-5 highlight tokens underlined).

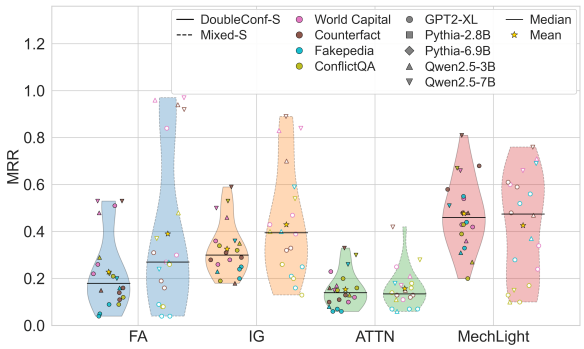


Figure 14: MRR – Mean Reciprocal Rank for the predicted answer tokens within each instance for both the **Double-Conflicting-Swap** and **Mixed-Swap** setups. Higher MRR is better.