# Revisiting Non-Verbatim Memorization in Large Language Models: The Role of Entity Surface Forms

**Yuto Nishida**[1,2] **Naoki Shikoda**[1] **Yosuke Kishinami**[2] **Ryo Fujii**[2]
**Makoto Morishita**[2] **Hidetaka Kamigaito**[1] **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology [2]Future Corporation
{nishida.yuto.nu8, kamigaito.h, taro}@is.naist.jp
shikoda.naoki.sm1@naist.ac.jp
{y.kishinami.rh, r.fujii.6d, m.morishita.pi}@future.co.jp

## Abstract

Understanding what kinds of factual knowledge large language models (LLMs) memorize is essential for evaluating their reliability and limitations. Entity-based QA is a common framework for analyzing non-verbatim memorization, but typical evaluations query each entity using a single canonical surface form, making it difficult to disentangle fact memorization from access through a particular name. We introduce *RedirectQA*,[1] an entity-based QA dataset that uses Wikipedia redirect information to associate Wikidata factual triples with categorized surface forms for each entity, including alternative names, abbreviations, spelling variants, and common erroneous forms. Across 13 LLMs, we examine surface-conditioned factual memorization and find that prediction outcomes often change when only the entity surface form changes. This inconsistency is category-dependent: models are more robust to minor orthographic variations than to larger lexical variations such as aliases and abbreviations. Frequency analyses further suggest that both entity- and surface-level frequencies are associated with accuracy, and that entity frequency often contributes beyond surface frequency. Overall, factual memorization appears neither purely surface-specific nor fully surface-invariant, highlighting the importance of surface-form diversity in evaluating non-verbatim memorization.

## 1 Introduction

Large language models (LLMs) store a wide range of factual knowledge in their parameters, enabling them to answer many knowledge-intensive questions without external retrieval (Petroni et al., 2019; Yu et al., 2023). At the same time, when the required knowledge is absent or inaccessible, LLMs may produce hallucinated or erroneous answers (Simhi et al., 2024). Understanding what factual knowledge LLMs memorize non-verbatim, and under what conditions they can access it, is therefore central to evaluating their reliability and limitations.

A common way to analyze non-verbatim memorization is entity-based question answering (QA), where models are queried about factual relations involving entities and memorization is measured by answer accuracy (Sciavolino et al., 2021; Mallen et al., 2023; Maekawa et al., 2024). This line of work has shown that facts about low-frequency or low-popularity entities are less likely to be memorized (Kandpal et al., 2023; Mallen et al., 2023; Maekawa et al., 2024). However, typical evaluations instantiate each entity using a single canonical surface form. This makes it difficult to disentangle whether a model has memorized a fact about an entity from whether it can access that fact through the particular name used in the question.

This distinction matters because entities are often referred to by multiple surface forms. A model that answers correctly for a canonical name such as *Pelé* may not necessarily access the same fact when the entity is referred to as *Edson Arantes do Nascimento*. Indeed, in our preliminary diagnostic using Pythia-12B (Biderman et al., 2023) on a redirect-augmented version of PopQA (Mallen et al., 2023), 23.7% of canonical–redirect question pairs yield inconsistent predictions (Appendix B.1). This observation motivates a systematic evaluation in which the underlying fact is controlled while the entity surface form is varied.

To analyze this phenomenon systematically, we introduce *RedirectQA*, an entity-based QA dataset that associates Wikidata factual triples with multiple entity surface forms using Wikipedia redirect information. The key design of RedirectQA is to hold the factual relation and gold answer fixed while varying only the surface form of the subject

[1] https://huggingface.co/datasets/naist-nlp/RedirectQA



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47046–47061
July 2-7, 2026 

| | Surface Category | Type | Question | LLM Answer | Eval. |
|---|---|---|---|---|---|
| Case 1 | Canonical | - | *What is **David Guetta**'s occupation?* | *DJ* | ✓ |
| | `from pseudonyms` | Alt./Abbrev. | *What is **Jack Back**'s occupation?* | *actor* | ✗ |
| Case 2 | Canonical | - | *What sport does **José García Castro** play?* | *baseball* | ✗ |
| | `from alternative names` | Alt./Abbrev. | *What sport does **Pepillo II** play?* | *soccer* | ✓ |
| Case 3 | Canonical | - | *In what city was **J. M. Coetzee** born?* | *Cape Town* | ✓ |
| | `from modifications` | Spell. Var. | *In what city was **J M Coetzee** born?* | *Cape Town* | ✓ |
| Case 4 | Canonical | - | *What is **Stan Coveleski**'s occupation?* | *baseball player* | ✓ |
| | `from misspellings` | Typ. Err. | *What is **Stan Covaleski**'s occupation?* | *actor* | ✗ |

Table 1: Illustrative examples of surface-conditioned factual access in RedirectQA. Each pair of rows refers to the same Wikidata entity and factual triple; only the subject entity surface form in the question is changed. The gold answer is therefore fixed within each case, but the Pythia-12B predictions can flip between ✓correct and ✗incorrect. The examples show canonical-to-redirect failures, the reverse pattern, robustness to a minor orthographic variant, and fragility to a common misspelling. Aggregate results across 13 LLMs are reported in § 3.

entity. As illustrated in Table 1, this design exposes cases where a model answers correctly under one surface form but incorrectly under another, even though the underlying fact is unchanged. Redirect surface forms are further annotated with categories such as alternative names, abbreviations, spelling variants, and common erroneous forms, enabling controlled analyses of how different types of naming variation affect factual QA.

Using RedirectQA, we evaluate 13 LLMs and find that prediction outcomes often differ across surface forms of the same entity, even though the underlying factual triple is held fixed. The inconsistency is category-dependent: models are relatively robust to minor orthographic variations, such as spelling differences, diacritics, and punctuation changes, but are less consistent for larger lexical variations, such as aliases, alternative names, and abbreviations. These results indicate that non-verbatim memorization cannot be treated as fully surface-invariant, even when the entity and fact remain the same.

We further analyze how entity- and surface-level frequencies relate to memorization. By decomposing aggregate entity frequency into surface-level frequencies, we find that accuracy is associated with both the frequency of a specific surface form and the aggregate frequency of the corresponding entity, with entity frequency often contributing beyond surface frequency. This pattern suggests cross-surface coupling in factual access, rather than purely independent memorization of each surface form. Together with the consistency results, these findings point to an intermediate picture in which factual memorization is neither purely surface-specific nor fully surface-invariant.

Overall, our work shows that evaluating non-verbatim memorization through canonical entity names alone can miss surface-conditioned failures in factual access. RedirectQA provides a controlled resource for studying these effects, highlighting surface-form diversity as a key factor in evaluating what LLMs memorize and how reliably they can access it.

## 2 RedirectQA

We introduce *RedirectQA*, an entity-based factual QA dataset designed to analyze how LLMs access the same factual knowledge through different surface forms of an entity. RedirectQA associates Wikidata factual triples in the form of $(subject, relation, object)$ with multiple subject entity surface forms using Wikipedia redirect information. The key design is to keep the factual relation and gold answer fixed while varying the surface form of the subject entity. We follow the open-domain QA setting (Roberts et al., 2020), evaluating models on factual questions without providing external evidence.

### 2.1 Wikipedia Redirects as Surface-Form Resources

Wikipedia article titles are chosen according to naming guidelines,[2] typically favoring recognizable, natural, and searchable expressions among possible names for a topic or entity. To make articles accessible through alternative expressions, Wikipedia provides redirect pages, which automatically forward users from a redirect title to the corresponding main article. For example, the page ti-

[2] https://en.wikipedia.org/wiki/Wikipedia:Article_titles

tled "NYT" redirects to the article "The New York Times." Such redirects provide a large-scale source of surface forms that refer to the same underlying entity.

Redirect pages are often annotated with redirect categories that describe the relationship between the redirect title and the main article title.[3] For instance, the redirect page "NYT" is annotated with `Redirects from initialisms`, indicating that "NYT" is an initialism for "The New York Times."[4] These categories allow us to group surface forms by the type of variation they represent.

In RedirectQA, we use this redirect structure to define two types of subject entity surface forms. The *canonical surface form* is the article title associated with the entity, while *redirect surface forms* are the titles of pages that redirect to that article.

However, not all redirects correspond to genuine surface-form variants of the target entity. For example, in the category `from books`, the title of a book may redirect to the article of its author, rather than to an alternative name for the same entity. We therefore manually selected 33 frequent redirect categories that clearly represent surface-form variation. We group the selected categories into three broad types. First, *Alternative Names and Abbreviations* include cases such as "Stevland Hardaway Judkins" redirecting to "Stevie Wonder" (`from birth names`). Second, *Spelling Variants* include cases such as "Nicolas Sarközy" redirecting to "Nicolas Sarkozy" (`from titles with diacritics`). Third, *Typical Errors* include cases such as "Christian Ronaldo" redirecting to "Cristiano Ronaldo" (`from incorrect names`). The selected categories and their types are listed in Table 3.

## 2.2 Dataset Structure

For each factual triple, RedirectQA creates instances in which the subject entity is expressed using different surface forms while the relation and gold answer remain fixed.

We use three dataset units throughout the paper. A *surface-form instance*, or simply a *surface instance*, pairs a factual triple with a subject surface form. A *canonical–redirect pair* consists of a redirect surface instance and the corresponding canonical surface instance for the same factual triple. This pair is the unit used in our consistency analyses. A *question realization* is obtained by rendering a surface instance with a relation-specific question template.

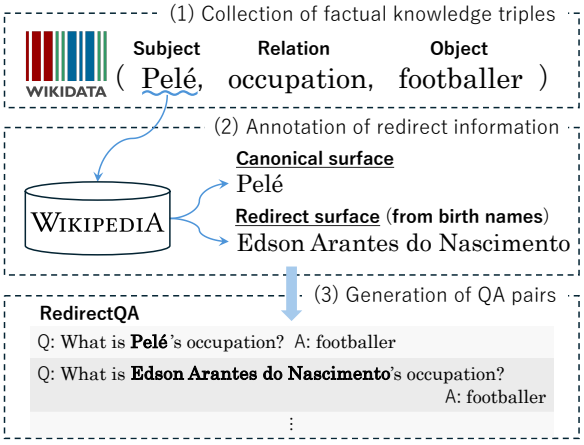



Figure 1: Overview of the RedirectQA construction process: (1) Factual triples are collected from Wikidata. (2) Each subject entity is associated with canonical and redirect surface forms, together with redirect categories, using Wikipedia redirects. (3) Question realizations are generated from surface instances using relation-specific question templates.

## 2.3 Dataset Construction

The overall construction process is illustrated in Figure 1. We first collect factual triples from Wikidata, then associate each subject entity with canonical and redirect surface forms using Wikipedia redirect information, and finally render surface instances into question realizations with relation-specific templates.

**(1) Collection of factual triples.** We collected factual triples from a Wikidata dump, targeting entities with English labels and restricting the relation types to 16 (e.g., `occupation`), following the setup of Mallen et al. (2023). To ensure that each factual question has a unique and unambiguous gold answer, we excluded cases where multiple English entities shared the same canonical surface form in Wikidata. We also filtered out triples without corresponding Wikipedia pages, as well as triples whose subject or object entities had zero pageviews over the past year.[5] Finally, we randomly sampled 500k triples from the remaining set for subsequent processing.

**(2) Annotation of redirect information.** For each subject entity in the sampled triples, we collected redirect surface forms and their redirect cat-

[3] A redirect page may have zero or multiple categories.

[4] Hereafter, we omit the prefix `Redirects` when referring to category names.

[5] We used Wikimedia pageview statistics aggregated over 2024-01–2024-12.

egories from Wikipedia. We discarded redirect surface forms whose categories were not among the selected categories described in § 2.1, and removed triples for which no valid redirect surface remained. To reduce ambiguity and duplication, we further removed redirect surfaces whose strings matched existing English entity labels in Wikidata. Finally, to mitigate severe class imbalance, we downsampled surface instances from overrepresented categories such as `from titles without diacritics` and `from other capitalisations`. This balancing step reduces the dominance of a small number of redirect types while maintaining approximately 30k surface instances.

**(3) Generation of question realizations.** For each surface instance, we generated questions using relation-specific templates. To reduce sensitivity to question wording, we used two templates for each relation type, following prior evidence that LLM predictions can be sensitive to superficial variations in question templates (Sakai et al., 2024). The first is the original template used by Mallen et al. (2023). The second is a paraphrase of the original template generated using GPT-4o (OpenAI, 2024b), designed to preserve the same factual semantics while differing in question wording. Thus, each surface instance is rendered into two question realizations.

**Dataset Statistics.** After these steps, RedirectQA contains 30,560 surface instances derived from 14,672 factual triples: 14,672 canonical surface instances and 15,888 redirect surface instances. The 15,888 redirect surface instances define the canonical–redirect pairs used in our consistency analyses. Because each surface instance is rendered with two templates, the dataset contains 61,120 question realizations in total. Among the redirect surface instances, 8,667 are associated with *Alternative Names and Abbreviations*, 4,928 with *Spelling Variants*, and 2,884 with *Typical Errors*.[6] A detailed breakdown of redirect categories and their surface-instance counts is shown in Table 3.

## 3 Experiments

This section evaluates whether factual QA behavior remains consistent when only the subject entity surface form is changed. We first describe the evaluated models and inference protocol, and then analyze prediction consistency across canonical–redirect pairs and redirect categories. Frequency-based analyses are presented in § 4.

### 3.1 Experimental Setup

**Models.** We evaluated 13 LLMs spanning three tiers of accessibility and training transparency: *transparent models* with well-documented pretraining data and procedures, *open-weight models* with publicly available weights but limited training transparency, and a *proprietary model* accessed via an API. This design supports corpus-based frequency analyses that require traceable pretraining corpora, such as the analysis in § 4, while also testing whether surface-form effects persist across a broader range of widely used models.

We used three families of transparent models. For Pythia (Biderman et al., 2023), we used four model sizes: 410M, 2.8B, 6.9B, and 12B, pretrained on the Pile (Gao et al., 2020). For OpenSciRef v0.01 (Nezhurina et al., 2025), we used the Pile-pretrained variants at 0.4B and 1.7B among its publicly released corpus-specific variants. For OLMo 2 (OLMo et al., 2024), we used the final Stage-1 checkpoints at 1B, 7B, 13B, and 32B. OLMo 2 base-model training consists of Stage 1 pretraining on OLMo Mix 1124 followed by Stage 2 mid-training. Because our frequency analyses target the pretraining corpus, we evaluate the final Stage-1 checkpoints rather than checkpoints after Stage 2. These Stage-1 checkpoints share the same data mixture, although their training budgets differ across model sizes.

To include strong instruction-tuned open-weight models, we evaluated Qwen 3 (Yang et al., 2025) 30B-A3B-Instruct and Llama 3.1 (Grattafiori et al., 2024) 8B-Instruct. As a representative *proprietary* model, we used the GPT-4o-mini (OpenAI, 2024a) snapshot `gpt-4o-mini-2024-07-18` via the API.

**Inference and Evaluation.** For local inference on training-transparent and open-weight models, we applied 8-bit quantization to reduce memory usage. Following Mallen et al. (2023), we used prompts of the form "`Q: <question> A:`" in a 15-shot setting. For each test question, the demonstrations were deterministically sampled with a fixed random seed from canonical-surface instances of other relation types, excluding the same factual triple. Specifically, we sampled one demonstration from each of the other 15 relation types.

[6] These types are not mutually exclusive, as a redirect surface instance may be associated with multiple categories; therefore, the type-level counts do not sum to the total number of instances.

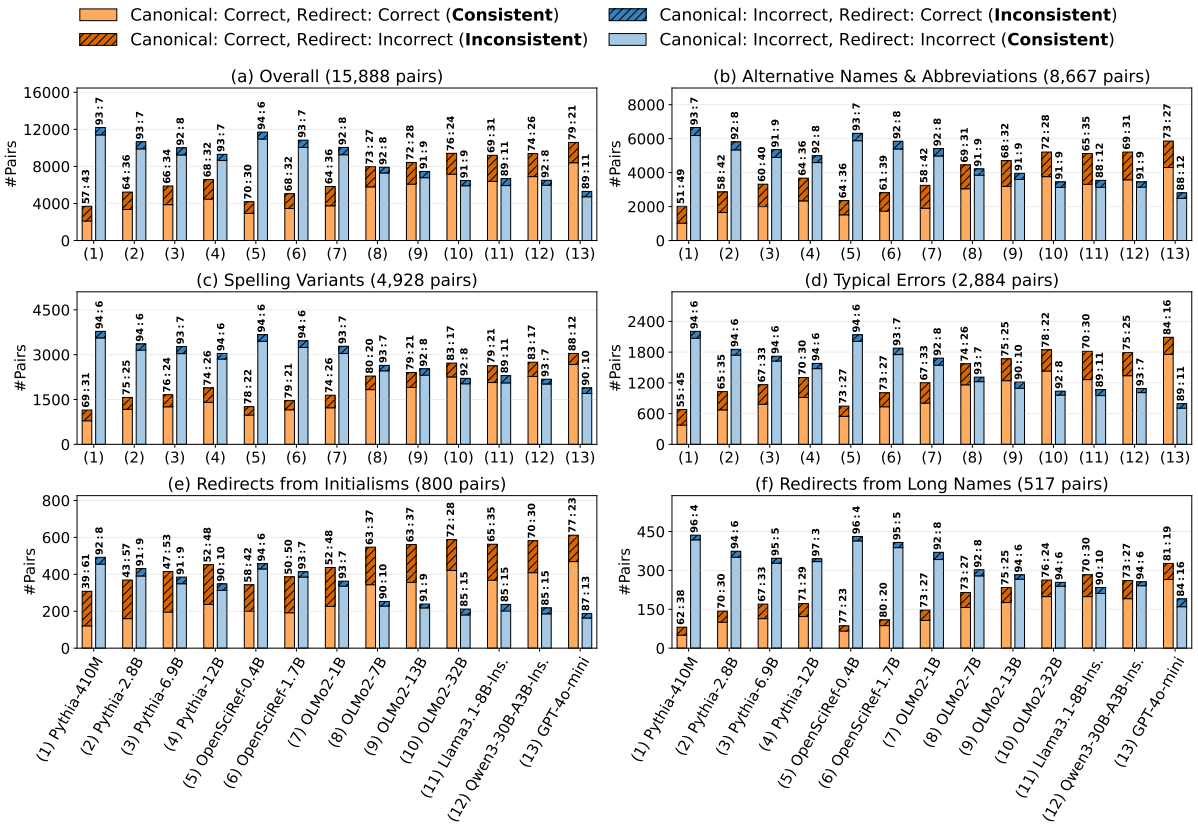



Figure 2: Prediction consistency between canonical and redirect surface forms on RedirectQA using the original question template. Each panel reports results for a redirect type or selected redirect category. For each model, the left stacked bar contains canonical–redirect pairs where the canonical question is answered correctly, and the right stacked bar contains pairs where it is answered incorrectly. Light segments indicate consistent correctness outcomes across the two surface forms, while dark hatched segments indicate correctness flips. Numbers above bars show the consistent:inconsistent percentage split within each bar.

For local models, we generated up to 15 new tokens and extracted the first generated line as the prediction. For GPT-4o-mini, we used the API with temperature 0, top-p 1, and a maximum of 100 output tokens, applying the same first-line extraction. We evaluated predictions using alias-aware string matching. For each question, a prediction was counted as correct if the extracted prediction contained any acceptable surface form of the gold answer entity, allowing simple case variants. This avoids penalizing alternative valid names of the answer entity when they are included in the acceptable surface set, while retaining a string-based evaluation appropriate for our entity-answering setting.

### 3.2 Prediction Consistency Across Surface-Form Categories

We analyze whether model predictions remain consistent when only the subject entity surface form is changed. For each canonical–redirect pair, we compare the correctness of the model's answer under the canonical surface form with that under the corresponding redirect surface form. We call a pair *consistent* if the two predictions have the same correctness outcome, i.e., both are correct or both are incorrect, and *inconsistent* otherwise. Because correct–correct and incorrect–incorrect consistency have different interpretations, we separately analyze pairs where the canonical question is answered correctly and pairs where it is answered incorrectly.

Figure 2 summarizes prediction consistency across 13 LLMs using the original question template. Overall, surface-form changes induce non-negligible correctness flips across all model classes. Within several model families, larger models tend to be more consistent, but the effect is not monotonic across all models or categories. Moreover, even strong instruction-tuned and proprietary models do not achieve perfect consistency, indicating that access to factual knowledge remains sensitive

to how the subject entity is named.

The category-wise results reveal systematic differences. *Spelling Variants* yield the highest consistency across models, suggesting that models are relatively robust to minor orthographic changes such as punctuation, capitalization, and diacritics. By contrast, *Alternative Names and Abbreviations* show substantially lower consistency, indicating that larger lexical changes are more likely to disrupt factual access. *Typical Errors* generally fall between these two types, reflecting partial but imperfect robustness to misspellings, miscapitalizations, and incorrect names.

The selected subcategories within *Alternative Names and Abbreviations* further illustrate that not all lexical variants are equally difficult. Redirects from initialisms are especially challenging: abbreviated forms such as *NYT* for *The New York Times* often fail to elicit the same answer as the canonical surface form. In contrast, redirects from long names tend to be more consistent, possibly because some longer alternative names preserve lexical or semantic cues that support factual access. These trends show that surface-form effects are not merely random noise, but depend on the type of relation between the redirect and canonical surface forms.

We repeat the same analysis using the paraphrased question template generated by GPT-4o and report the results in Appendix B.2. Although absolute accuracy can vary with question wording, the model-wise consistency patterns and category-wise differences largely mirror those obtained with the original template. This supports the conclusion that the observed surface-form effects are not artifacts of a single question template.

The illustrative examples in Table 1 provide concrete instances of these aggregate patterns, including canonical-to-redirect failures, the reverse pattern, robustness to a minor orthographic variant, and fragility to a common misspelling. The reverse pattern is particularly informative: a model can fail under the Wikipedia canonical title but succeed under an alternative surface form, suggesting that human-oriented canonicality does not necessarily coincide with the surface form through which an LLM most reliably accesses a fact. Thus, RedirectQA captures not only degradation from canonical to redirect surfaces, but also asymmetric surface dependence in factual access.

## 4 Analysis: Entity- and Surface-Level Frequency Signals

In § 3.2, we observed that factual QA predictions are not fully consistent across canonical and redirect surface forms, and that the degree of consistency varies across redirect categories. These findings raise a question about the granularity of factual memorization: are surface forms memorized independently, or is factual access coupled across different surface forms of the same entity? If accuracy for a target surface form is associated only with that surface form's own frequency, this would support a strongly surface-specific view. If aggregate entity frequency also predicts accuracy beyond the target surface frequency, however, this would suggest cross-surface coupling in factual access. We investigate this question by analyzing how entity-level and surface-level frequencies relate to factual QA accuracy.

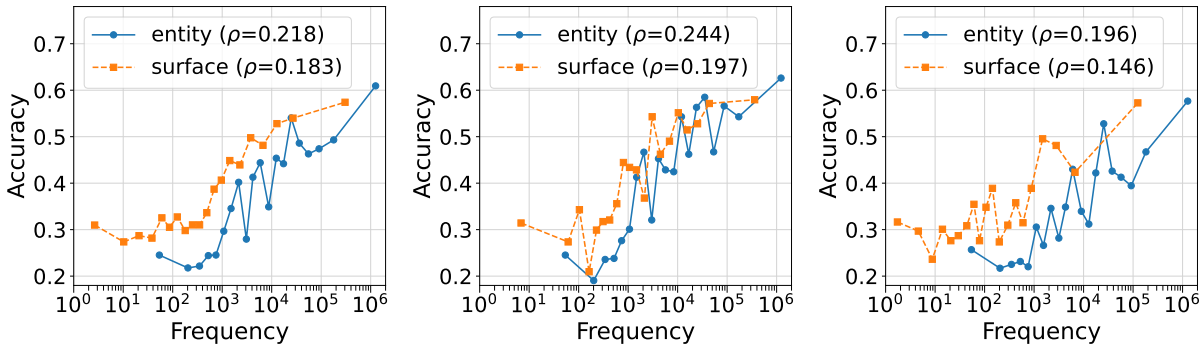
Previous studies have shown that entity frequency is positively correlated with factual memorization, as reflected in factual QA accuracy (Kandpal et al., 2023; Maekawa et al., 2024). Such studies typically estimate entity frequency from pretraining or related corpora by using an entity linker to identify mentions of an entity across surface forms, and then treating the total number of linked mentions as the entity's frequency. We decompose this aggregate entity frequency into surface-level frequencies, allowing us to ask whether accuracy is associated with the frequency of the target surface form itself or with the aggregate frequency of the corresponding entity.

### 4.1 Counting Entity and Surface Frequencies

Following Kandpal et al. (2023), we counted entity and surface frequencies from the pretraining corpora of the training-transparent model families. For Pythia and OpenSciRef v0.01, we used the Pile dataset (Gao et al., 2020), which contains approximately 300B tokens. For OLMo 2, we estimated frequencies from OLMo Mix 1124 (OLMo et al., 2024), the Stage-1 data mixture, by randomly sampling 10% of documents; this yields a corpus size roughly comparable to the Pile in total tokens.

We performed large-scale entity linking using DBpedia Spotlight (Mendes et al., 2011), which links text spans to Wikipedia entities.[7] For each

[7]We retrieved the corresponding Wikipedia entities by resolving DBpedia URIs obtained from the linker through the official DBpedia SPARQL endpoint.




(a) Overall (4,284 instances) (b) Canonical only (2,112 instances) (c) Redirect only (2,172 instances)

Figure 3: Relationship between accuracy and entity/surface frequencies for Pythia-12B. Each point shows the mean accuracy of surface instances within one of 20 frequency bins with approximately equal numbers of instances. For each surface instance, accuracy is averaged over the two question realizations. Pearson correlations $\rho$ are computed between $\log(\text{frequency})$ and accuracy and are shown in the legend.

entity, *entity frequency* is the total number of linked mentions of that entity. For a particular surface form, *surface frequency* is the number of linked mentions of the same entity whose span exactly matches that surface form. Thus, entity frequency aggregates over all observed linked surface forms of the entity, whereas surface frequency refers to the specific surface form used in a RedirectQA surface instance.

We annotated each RedirectQA surface instance with the entity frequency of its subject entity and the surface frequency of its subject surface form. Following Kandpal et al. (2023), we filtered out zero-frequency cases, which may reflect entity-linking failures or missing corpus coverage and cannot be used in log-frequency analyses. We retained surface instances only when the subject entity was linked at least once and the target subject surface form was observed at least once as a linked mention of that entity. Under this filtering criterion, the Pile-based analysis for Pythia and OpenSciRef v0.01 retains 4,284 surface instances from 2,112 factual triples, while the OLMo Mix 1124 analysis for OLMo 2 retains 4,356 surface instances from 2,147 factual triples. These filtered subsets are used in the frequency analyses below.

### 4.2 Correlation Analysis

We first examine the relationship between frequency and factual QA accuracy in three subsets: *overall*, *canonical-only*, and *redirect-only*. The canonical-only and redirect-only subsets contain surface instances whose subject entity is expressed with the canonical and redirect surface forms, respectively, while the overall subset contains both. For each surface instance, we compute accuracy as the mean correctness score across the two question realizations and use this continuous score in the correlation analyses.

Figure 3 reports the results for Pythia-12B. Each plot bins surface instances by frequency and shows the mean accuracy in each bin; Pearson correlations between $\log(\text{frequency})$ and accuracy are shown in the legends. For Pythia-12B, both entity and surface frequencies are positively correlated with accuracy in all three subsets, with all reported correlations significantly different from zero ($p < 0.01$). This extends prior findings that entity frequency is predictive of factual QA accuracy (Kandpal et al., 2023) by showing that surface frequency is also positively associated with accuracy. Entity frequency correlates more strongly with accuracy than surface frequency for Pythia-12B, and Appendix B.3 confirms that this difference is significant for the canonical-only subset. The appendix further shows that, across all training-transparent models and subsets, both frequency types have statistically significant positive correlations with accuracy, with entity frequency consistently stronger in the canonical-only subset.

### 4.3 Partial-Correlation Analysis

Because entity and surface frequencies are correlated, simple correlations cannot determine whether each frequency type has an association with accuracy beyond the other. We therefore compute Pearson partial correlations between log frequency and accuracy while controlling for the other frequency type. A partial correlation $\rho(X, Y \mid Z)$ measures the correlation between $X$ and $Y$ after lin-

| Size | Correlation Type | Partial Correlation | | |
|---|---|---|---|---|
| | | Overall | Canonical | Redirect |
| **OLMo 2** | | | | |
| 1B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.132^*$ | $0.182^*$ | $0.125^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.049^*$ | $-0.069^*$ | $0.084^*$ |
| 7B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.104^*$ | $0.127^*$ | $0.120^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.091^*$ | $-0.010$ | $0.106^*$ |
| 13B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.088^*$ | $0.102^*$ | $0.106^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.098^*$ | $0.010$ | $0.111^*$ |
| 32B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.064^*$ | $0.113^*$ | $0.065^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.083^*$ | $-0.032$ | $0.114^*$ |
| **OpenSciRef** | | | | |
| 0.4B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.150^*$ | $0.197^*$ | $0.133^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.042^*$ | $-0.084^*$ | $0.080^*$ |
| 1.7B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.141^*$ | $0.125^*$ | $0.151^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.049^*$ | $0.000$ | $0.033$ |
| **Pythia** | | | | |
| 410M | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.107^*$ | $0.161^*$ | $0.091^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.069^*$ | $-0.050$ | $0.070^*$ |
| 2.8B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.107^*$ | $0.124^*$ | $0.114^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.076^*$ | $-0.008$ | $0.051$ |
| 6.9B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.126^*$ | $0.141^*$ | $0.124^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.073^*$ | $-0.013$ | $0.066^*$ |
| 12B | $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ | $0.142^*$ | $0.148^*$ | $0.146^*$ |
| | $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ | $0.074^*$ | $-0.009$ | $0.064^*$ |

Table 2: Results of the partial-correlation analysis. Here, $\text{Ent}$ and $\text{Surf}$ denote log-transformed entity and surface frequencies, respectively, and $\text{Acc}$ denotes accuracy. Each value reports a Pearson partial correlation between one log-frequency signal and accuracy while controlling for the other, namely $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ and $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$. Superscript $^*$ indicates that the partial correlation is significantly different from zero ($p < 0.01$).

early removing the variation explained by a control variable $Z$, equivalently by correlating the residuals of $X$ and $Y$ after regressing both on $Z$. Specifically, we compute $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ to measure the association between entity frequency and accuracy after controlling for surface frequency, and $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$ for the reverse direction.

Table 2 summarizes the results for all training-transparent model families. In the overall and redirect-only subsets, both entity and surface frequencies often retain positive partial correlations with accuracy, indicating that each captures information not fully explained by the other. In the canonical-only subset, however, the partial correlation for surface frequency is typically close to zero or negative, whereas entity frequency remains consistently positive. This suggests that, for canonical surface forms, aggregate entity frequency is more informative than the frequency of the canonical surface form alone. We further verify in Appendix B.4 that the redirect-only results are not driven solely by extremely low-frequency redirect surface forms.

### 4.4 Discussion

These results are inconsistent with a purely independent surface-specific account. Accuracy for a target surface form is associated not only with that surface form's own frequency, but also with the aggregate frequency of the corresponding entity. This pattern is consistent with cross-surface coupling in factual access, rather than independent memorization of each surface form. The coupling is clearest for canonical surfaces, where surface frequency has little independent association with accuracy once entity frequency is controlled for, whereas entity frequency remains a consistent predictor.

Prior entity-based QA studies commonly evaluate factual memorization through canonical surface forms and relate performance to aggregate entity frequency (Kandpal et al., 2023). Our analysis extends this setting by decomposing entity frequency into surface-level frequencies. The results suggest that this conventional focus on entity frequency remains a useful lens, especially when evaluation uses canonical surface forms, but it also obscures surface-form effects that become visible when alternative names are considered.

As a complementary probe, Appendix B.5 reports an entity-linking-style binary QA experiment that directly asks whether a model links two surface forms to the same entity. Pythia-12B shows only modest balanced accuracy in this probe, suggesting that surface-form equivalence recognition is incomplete and does not by itself explain the category-wise consistency patterns observed in § 3.2. A fuller account of surface-dependent factual access will require analyses that more directly examine the internal representations and retrieval processes underlying these effects.

## 5 Related Work

**Memorization in LLMs.** Analyses of LLM memorization are often divided by whether they focus on exact reproduction or factual generalization (Kandpal et al., 2023). One line of work studies *verbatim memorization*, the literal reproduction of training data (Carlini et al., 2021, 2023; Chen et al., 2024), which is closely related to pri-

vacy risks and data leakage. Another line studies *non-verbatim memorization*, where models retain factual associations that can be elicited without reproducing the original training text. This setting is commonly evaluated through entity-based QA datasets (Kandpal et al., 2023; Mallen et al., 2023; Maekawa et al., 2024). Our work belongs to the latter line, focusing on how entity surface forms affect access to memorized factual knowledge.

**Entity-based factual memorization.** Kandpal et al. (2023) extracted entity-based QA pairs from open-domain datasets such as NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), showing that facts with low training-data frequency are less likely to be answered correctly. Elazar et al. (2023) used a causal analysis of masked language models to show that simple training-data statistics, such as co-occurrence counts, can affect factual predictions. Mallen et al. (2023) introduced PopQA and, together with EntityQuestions (Sciavolino et al., 2021), showed that LLMs struggle with less popular entities, measured by Wikipedia page views. Maekawa et al. (2024) introduced WitQA and further showed that relation frequency also affects factual knowledge memorization. These studies provide important insights into factors that predict factual QA success, but they typically instantiate each entity with a single canonical surface form. As a result, they do not separate whether a model has memorized a fact about an entity from whether it can access that fact through a particular entity name. RedirectQA addresses this gap by pairing the same factual triples with multiple categorized surface forms for each entity.

**Robustness and consistency under input variation.** Robustness and consistency under meaning-preserving input variation have been studied in several settings. Zheng et al. (2024) investigated robustness to surface-level variations in multiple-choice questions, and Andriushchenko and Flammarion (2025) examined whether safety-aligned LLMs maintain consistent refusal behavior under tense variations. In QA and factual prediction settings, Ribeiro et al. (2019) proposed evaluating models through consistency constraints across related questions, and Elazar et al. (2021) showed that meaning-preserving paraphrases can still yield inconsistent factual predictions. These studies primarily examine prompt- or question-level variation. In contrast, our work isolates variation in the entity mention itself while holding the underlying entity, factual relation, and answer fixed. This allows us to analyze how factual access differs across naturally occurring categories of entity surface forms, such as aliases, abbreviations, spelling variants, and common errors.

## 6 Conclusion

We introduced *RedirectQA*, an entity-based factual QA dataset that pairs Wikidata factual triples with multiple categorized entity surface forms using Wikipedia redirect information. Using RedirectQA, we showed that LLM prediction outcomes often change when only the subject entity surface form is changed, indicating that access to memorized factual knowledge is partially surface-dependent. The inconsistency is category-dependent: models are relatively robust to minor orthographic variations, such as spelling differences, but less consistent for larger lexical variations, such as aliases, alternative names, and abbreviations. Our frequency analyses further showed that accuracy is associated with both the frequency of a specific surface form and the aggregate frequency of the corresponding entity, suggesting cross-surface coupling in factual access rather than purely independent memorization of each surface form. Overall, our findings show that evaluating non-verbatim memorization through canonical entity names alone can miss surface-conditioned failures, highlighting the importance of surface-form diversity in factual QA evaluation.

## Limitations

Our analysis focuses on English factual QA and does not cover multilingual, cross-lingual, or domain-specific settings. RedirectQA relies on Wikipedia and Wikidata, whose coverage and naming conventions reflect the biases and editorial practices of Wikimedia projects; Wikipedia redirects provide a systematic source of surface forms, but they do not cover all real-world ways of referring to entities. Our dataset also varies only the subject entity surface form, leaving object-side variation and broader question paraphrasing beyond the main scope.

Although our evaluation uses alias-aware string matching for answer entities, it may still miss semantically correct answers whose surface forms are not included in the acceptable answer set. Our frequency-based analyses are restricted to training-

transparent models and depend on entity-linking quality and the filtered subset of surface instances observed in the relevant corpora. Finally, our experiments evaluate factual access through QA behavior and frequency correlations, but do not directly probe the internal representations or training dynamics that give rise to surface-dependent access. Future work could extend RedirectQA to multilingual and domain-specific settings, broaden surface-form resources beyond Wikipedia redirects, and develop more direct analyses of how surface–entity associations are represented and acquired during pretraining.

## Ethical Considerations

This study uses publicly available data from Wikimedia projects, including Wikipedia, Wikidata, and pageview statistics. We follow the licenses of the original resources: Wikipedia text is distributed under CC BY-SA 4.0, while Wikidata and pageview statistics are distributed under CC0 1.0. RedirectQA contains entity names, redirect titles, factual triples, and generated questions derived from these public resources. It does not include private or newly collected sensitive personal information, although some entities may correspond to public figures already represented in Wikimedia projects.

RedirectQA also uses question templates adapted from PopQA (Mallen et al., 2023), which is distributed under the MIT license. To ensure transparency and reproducibility, we make RedirectQA available under the CC BY-SA 4.0 license, following the most restrictive license among the source resources. Because Wikipedia and Wikidata coverage is not demographically or geographically uniform, RedirectQA may reflect biases present in these resources. The dataset is intended for evaluating factual QA behavior and should not be used to draw normative conclusions about individuals or groups.

## Acknowledgments

This work was partially supported by JST SPRING Grant Number JPMJSP2140 and JSPS KAKENHI Grant Number JP23H03458. Computational resources were provided in part by "mdx: a platform for building data-empowered society."

## References

Maksym Andriushchenko and Nicolas Flammarion. 2025. Does refusal training in LLMs generalize to the past tense? In *The Thirteenth International Conference on Learning Representations*.

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*, pages 2397–2430.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650.

Bowen Chen, Namgi Han, and Yusuke Miyao. 2024. A multi-perspective analysis of memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11190–11209.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2023. Measuring causal effects of data statistics on language model's 'factual' predictions. *Preprint*, arXiv:2207.14251.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

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.

M. Hoerger. 2013. Zh: An updated version of steiger's z and web-based calculator for testing the statistical significance of the difference between dependent correlations.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 15696–15707.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Seiji Maekawa, Hayate Iso, Sairam Gurajada, and Nikita Bhutani. 2024. Retrieval helps or hurts? a deeper dive into the efficacy of retrieval augmentation to language models. 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 5506–5521.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, page 1–8, New York, NY, USA. Association for Computing Machinery.

Marianna Nezhurina, Jörg Franke, Taishi Nakamura, Timur Carstensen, Niccolò Ajroldi, Ville Komulainen, David Salinas, and Jenia Jitsev. 2025. Open-sci-ref-0.01: open and reproducible reference baselines for language model and dataset comparison. *Preprint*, arXiv:2509.09009.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Gpt-4o system card.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5418–5426.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024. Toward the evaluation of large language models considering score variance across instruction templates. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. 2024. Distinguishing ignorance from error in LLM hallucinations. *arXiv preprint arXiv:2410.22071*.

J. H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87:245–251.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate

rather than retrieve: Large language models are strong context generators. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

# A Details on RedirectQA Dataset

## A.1 Redirect Category Statistics

Table 3 provides a detailed breakdown of RedirectQA by redirect category. The Count column reports the number of surface instances assigned to each category. Because a redirect surface instance may be associated with multiple redirect categories, category-wise counts are not mutually exclusive and should not be summed to recover the total number of redirect surface instances. Similarly, broad-type counts reported in § 2.3 count unique surface instances associated with each type, whereas Table 3 reports counts at the category level.

# B Additional Experiments

## B.1 Preliminary Experiment

As a preliminary diagnostic, we augmented PopQA (Mallen et al., 2023) with Wikipedia redirect information using a procedure similar to that described in § 2.3. This produced 18,781 surface instances from 4,292 factual triples. For the consistency analysis, we formed canonical–redirect comparison pairs for which both the canonical and redirect questions were evaluated, yielding 14,489 pairs.

Table 4 shows the resulting correctness contingency table for Pythia-12B using the original question template. Among these canonical–redirect pairs, 23.7% yielded inconsistent correctness outcomes: the model was correct on one surface form but incorrect on the other. This preliminary result motivates the more systematic construction of RedirectQA.

## B.2 Robustness to Question Templates

In § 3.2, we analyzed prediction consistency using the original template adopted from Mallen et al. (2023). To examine whether the observed surface-form effects depend on a particular question wording, we repeat the same analysis using an additional paraphrased template generated by GPT-4o.

Figure 4 shows the results with the paraphrased template, using the same plotting convention as Figure 2. Although absolute accuracy can vary with question wording, the model-wise consistency patterns and qualitative differences across redirect types largely mirror those obtained with the original template. This suggests that the main surface-form effects are not artifacts of a single question template.

| Type | Redirect category | Count |
|---|---|---|
| Canonical | – | 14,672 |
| Alt./Abbrev. | from birth names | 1,029 |
| | from short names | 985 |
| | from alternative names | 981 |
| | from former names | 979 |
| | from surnames | 977 |
| | from abbreviations | 863 |
| | from initialisms | 800 |
| | from long names | 517 |
| | from given names | 395 |
| | from pseudonyms | 374 |
| | from personal names | 371 |
| | from plurals | 331 |
| | from married names | 137 |
| | from acronyms | 122 |
| | from letter-word combinations | 108 |
| | from technical names | 87 |
| | to plurals | 82 |
| | to initialisms | 74 |
| | from synonyms | 65 |
| | to acronyms | 35 |
| Spell. Var. | from titles without diacritics | 1,019 |
| | from alternative spellings | 1,014 |
| | from titles with diacritics | 998 |
| | from other capitalisations | 953 |
| | from modifications | 765 |
| | from ASCII-only titles | 56 |
| | from stylizations | 86 |
| | from titles without ligatures | 61 |
| | to ASCII-only titles | 35 |
| | from numerals | 23 |
| Typ. Err. | from miscapitalisations | 1,005 |
| | from misspellings | 978 |
| | from incorrect names | 974 |

Table 3: Dataset composition by canonical and redirect surface categories. The Count column indicates the number of surface instances assigned to each category. A redirect surface instance may belong to multiple categories, so category counts are not mutually exclusive. The broad-type counts reported in § 2.3 count unique surface instances per type and therefore need not equal the sum of category-level counts.

## B.3 Significance Test and Correlation Results for Transparent Models

This section provides supplementary results for the analyses in § 4.

**Steiger's $Z$-test.** To test whether the difference between the entity-frequency and surface-

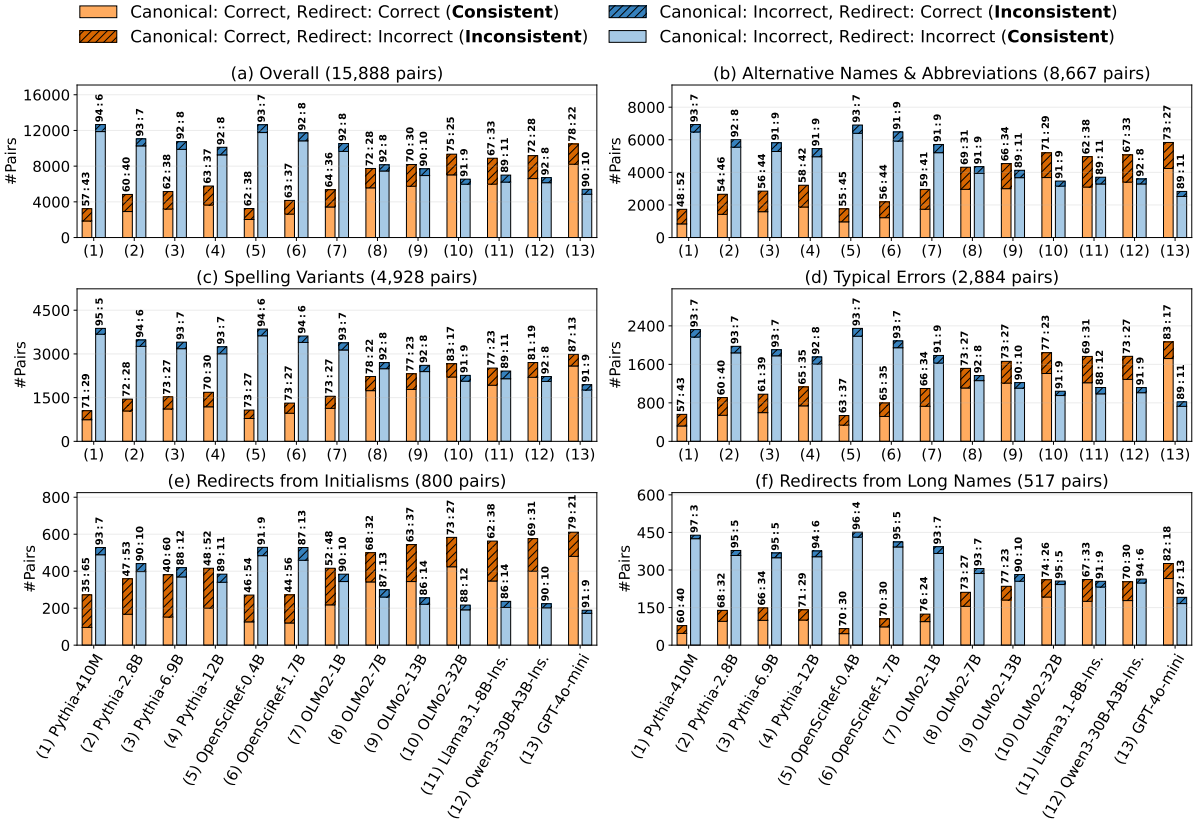



Figure 4: Prediction consistency between canonical and redirect surface forms on RedirectQA using the paraphrased question template. The plotting convention is the same as in Figure 2: light segments indicate consistent correctness outcomes, while dark hatched segments indicate correctness flips.

| | Canonical surface | |
|---|---|---|
| Redirect surface | Correct | Incorrect |
| Correct | 4,285 | 608 |
| Incorrect | 2,821 | 6,775 |

Table 4: Preliminary consistency analysis on a redirect-augmented version of PopQA using Pythia-12B. Rows indicate correctness under the redirect surface form, and columns indicate correctness under the canonical surface form. Counts are canonical–redirect comparison pairs. In 23.7% of pairs, the correctness outcome differs across the two surface forms.

frequency correlations reported in Figure 3 is statistically meaningful, we conducted Steiger's $Z$-test (Steiger, 1980; Hoerger, 2013), which compares two dependent Pearson correlations that share a common variable. For the Pythia-12B canonical-only subset, the test confirms that the correlation between entity frequency and accuracy is significantly larger than that between surface frequency and accuracy ($p < 0.01$, two-tailed).

**Correlation results for other models.** Table 5 reports Pearson correlations between log-transformed frequencies and accuracy for all training-transparent models. Across all models and subsets, both entity and surface frequencies show statistically significant positive correlations with accuracy. In the canonical-only subset, entity frequency consistently correlates more strongly with accuracy than surface frequency.

### B.4 Low-Frequency-Controlled Analysis for Redirect Surface Forms

Figure 3 shows that redirect surface forms include many extremely low-frequency cases. This raises the possibility that the redirect-only results in § 4 are driven primarily by very rare redirect surface forms, rather than by surface-form variation more generally. To address this concern, we repeat the correlation and partial-correlation analyses on a high-frequency subset of the redirect-only data. Specifically, we retain only redirect surface instances whose raw surface frequency is greater than 10 in the corresponding corpus. This filter-

| Model | Type | Overall | Canonical | Redirect |
|---|---|---|---|---|
| Pythia-410M | Entity | $0.175^{*}$ | $0.212^{*}$ | $0.138^{*}$ |
| | Surface | $0.155^{*}$ | $0.148^{*}$ | $0.126^{*}$ |
| Pythia-2.8B | Entity | $0.179^{*}$ | $0.207^{*}$ | $0.154^{*}$ |
| | Surface | $0.163^{*}$ | $0.167^{*}$ | $0.116^{*}$ |
| Pythia-6.9B | Entity | $0.199^{*}$ | $0.228^{*}$ | $0.173^{*}$ |
| | Surface | $0.172^{*}$ | $0.182^{*}$ | $0.137^{*}$ |
| Pythia-12B | Entity | $0.218^{*}$ | $0.244^{*}$ | $0.196^{*}$ |
| | Surface | $0.183^{*}$ | $0.197^{*}$ | $0.146^{*}$ |
| OpenSciRef-0.4B | Entity | $0.208^{*}$ | $0.228^{*}$ | $0.189^{*}$ |
| | Surface | $0.151^{*}$ | $0.143^{*}$ | $0.157^{*}$ |
| OpenSciRef-1.7B | Entity | $0.202^{*}$ | $0.220^{*}$ | $0.186^{*}$ |
| | Surface | $0.154^{*}$ | $0.182^{*}$ | $0.115^{*}$ |
| OLMo-2-1B | Entity | $0.205^{*}$ | $0.217^{*}$ | $0.196^{*}$ |
| | Surface | $0.166^{*}$ | $0.138^{*}$ | $0.174^{*}$ |
| OLMo-2-7B | Entity | $0.201^{*}$ | $0.203^{*}$ | $0.203^{*}$ |
| | Surface | $0.195^{*}$ | $0.160^{*}$ | $0.195^{*}$ |
| OLMo-2-13B | Entity | $0.188^{*}$ | $0.188^{*}$ | $0.191^{*}$ |
| | Surface | $0.192^{*}$ | $0.159^{*}$ | $0.193^{*}$ |
| OLMo-2-32B | Entity | $0.146^{*}$ | $0.150^{*}$ | $0.145^{*}$ |
| | Surface | $0.155^{*}$ | $0.104^{*}$ | $0.173^{*}$ |

Table 5: Pearson correlation coefficients between accuracy and log-transformed entity and surface frequencies. "Entity" uses the total frequency aggregated over all observed linked surface forms of an entity, whereas "Surface" uses the frequency of the specific surface form. Superscript $^{*}$ indicates that the correlation is significantly different from zero ($p < 0.01$).

ing is stricter than the main preprocessing, which removes only zero-frequency cases.

Table 6 compares the full redirect-only subset with this high-frequency subset. The full-subset columns reproduce the redirect-only correlations and partial correlations reported in Table 5 and Table 2, while the high-frequency columns report the same analyses after excluding extremely low-frequency redirect surface instances. Across all training-transparent models, entity frequency remains significantly associated with accuracy after controlling for surface frequency. For Pythia and OpenSciRef, the partial correlation of surface frequency becomes smaller and is not statistically significant in the high-frequency subset once entity frequency is controlled for. For OLMo 2, surface frequency retains a positive partial correlation, but the qualitative pattern remains broadly consistent with the main analysis: removing extremely low-frequency redirect surfaces does not eliminate the entity-frequency signal. These results suggest that our conclusions are not driven solely by redirect surface forms that appear only a few times in the pretraining corpus.

### B.5 Evaluating Entity Linking Between Surface Forms

As a complementary behavioral probe, we evaluate whether a model can recognize that two surface forms refer to the same entity. This experiment does not directly reveal internal representations, but provides an additional measure of surface-to-entity linking behavior that may relate to the consistency patterns observed in § 3.2.

We constructed binary questions asking whether two surface forms refer to the same entity. For positive examples, we used canonical–redirect pairs from RedirectQA and generated category-specific yes/no questions with GPT-4o. For example, for the redirect category `from initialisms`, we used the template: "Is `<redirect surface>` an initialism for `<canonical surface>`?" An example instance is "Is *NYT* an initialism for *The New York Times*?"

For each positive example, we created two negative examples: (i) surface-level negatives, obtained by randomly replacing one character in a surface form, and (ii) semantic negatives, obtained by replacing the entity with a semantically similar but distinct entity retrieved through nearest-neighbor search in the fastText embedding space (Bojanowski et al., 2017). We used the publicly available English model `cc.en.300.bin`,[8] and measured similarity using squared-L2 distance in the embedding space. Thus, the evaluation set has a 1:2 ratio of positive to negative examples.

Table 7 reports raw and balanced accuracy for Pythia-12B across redirect types and selected categories. Balanced accuracy is computed as the average of positive-example accuracy and negative-example accuracy, treating the two negative types as a single negative class. Because the evaluation set has a 1:2 positive-to-negative ratio, raw accuracy can be affected by the larger number of negative examples. Balanced accuracy therefore provides a more appropriate summary under this label imbalance.

Overall balanced accuracy is 0.522, indicating only a modest ability to recognize that two surface forms refer to the same entity. The model is substantially more accurate on negative examples than on positive examples, suggesting that it is better at rejecting mismatched surface forms than at affirming true surface-form equivalences. Across

[8] https://fasttext.cc/docs/en/crawl-vectors.html

| Family | Size | Full redirect-only subset | | | | High-frequency subset | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\rho_E$ | $\rho_S$ | $\rho_{E\|S}$ | $\rho_{S\|E}$ | $\rho_E$ | $\rho_S$ | $\rho_{E\|S}$ | $\rho_{S\|E}$ |
| Pythia | 410M | $0.138^*$ | $0.126^*$ | $0.091^*$ | $0.070^*$ | $0.147^*$ | $0.109^*$ | $0.108^*$ | 0.042 |
| | 2.8B | $0.154^*$ | $0.116^*$ | $0.114^*$ | 0.051 | $0.168^*$ | $0.108^*$ | $0.133^*$ | 0.028 |
| | 6.9B | $0.173^*$ | $0.137^*$ | $0.124^*$ | $0.066^*$ | $0.191^*$ | $0.133^*$ | $0.145^*$ | 0.045 |
| | 12B | $0.196^*$ | $0.146^*$ | $0.146^*$ | $0.064^*$ | $0.218^*$ | $0.153^*$ | $0.165^*$ | 0.053 |
| OpenSciRef | 0.4B | $0.189^*$ | $0.157^*$ | $0.133^*$ | $0.080^*$ | $0.205^*$ | $0.136^*$ | $0.161^*$ | 0.039 |
| | 1.7B | $0.186^*$ | $0.115^*$ | $0.151^*$ | 0.033 | $0.205^*$ | $0.130^*$ | $0.163^*$ | 0.033 |
| OLMo 2 | 1B | $0.196^*$ | $0.174^*$ | $0.125^*$ | $0.084^*$ | $0.209^*$ | $0.177^*$ | $0.140^*$ | $0.081^*$ |
| | 7B | $0.203^*$ | $0.195^*$ | $0.120^*$ | $0.106^*$ | $0.201^*$ | $0.181^*$ | $0.127^*$ | $0.092^*$ |
| | 13B | $0.191^*$ | $0.193^*$ | $0.106^*$ | $0.111^*$ | $0.184^*$ | $0.180^*$ | $0.108^*$ | $0.100^*$ |
| | 32B | $0.145^*$ | $0.173^*$ | $0.065^*$ | $0.114^*$ | $0.142^*$ | $0.158^*$ | $0.071^*$ | $0.100^*$ |

Table 6: Low-frequency-controlled frequency analysis for the redirect-only subset. The full redirect-only subset uses the same filtering criterion as § 4; the high-frequency subset additionally retains only redirect surface instances whose raw surface frequency satisfies $f_{\text{surf}} > 10$. $\rho_E$ and $\rho_S$ denote Pearson correlations between accuracy and log-transformed entity and surface frequencies, respectively. $\rho_{E|S}$ and $\rho_{S|E}$ denote Pearson partial correlations, corresponding to $\rho(\text{Ent}, \text{Acc} \mid \text{Surf})$ and $\rho(\text{Surf}, \text{Acc} \mid \text{Ent})$. Superscript $^*$ indicates that the correlation or partial correlation is significantly different from zero ($p < 0.01$).

| Redirect Category/Type | Raw | Bal. | Pos. | Neg. |
|---|---|---|---|---|
| Overall | 0.585 | 0.522 | 0.334 | 0.711 |
| Alt./Abbrev. | 0.587 | 0.537 | 0.386 | 0.688 |
| Spell. Var. | 0.574 | 0.507 | 0.305 | 0.708 |
| Typ. Err. | 0.604 | 0.506 | 0.212 | 0.801 |
| `from initialisms` | 0.587 | 0.564 | 0.495 | 0.633 |
| `from long names` | 0.592 | 0.559 | 0.461 | 0.657 |

Table 7: Results of the entity-linking-style binary QA task with Pythia-12B. For each positive canonical–redirect pair, we include two negative examples: one surface-level negative and one semantic negative. Raw accuracy is computed over all examples. Balanced accuracy averages positive accuracy and negative accuracy, where the two negative types are treated as a single negative class. POS. and NEG. denote accuracy on positive and negative examples, respectively.

broad redirect types, balanced accuracy is highest for *Alternative Names and Abbreviations* and close to the balanced-accuracy random baseline of 0.5 for *Spelling Variants* and *Typical Errors*. For the two selected subcategories analyzed in § 3.2, `from initialisms` and `from long names` show similar balanced accuracies, despite exhibiting different factual QA consistency patterns in Figure 2. This suggests that the binary surface-linking probe alone cannot explain the category-wise differences in factual QA consistency. A deeper analysis of the mechanisms underlying surface-dependent factual access remains an important direction for future work.

# C Data, Models, and Software

## C.1 Data

**Wikimedia Dumps** provided by the Wikimedia Foundation. License: CC BY-SA 4.0 (Wikipedia text), CC0 1.0 (Wikidata and pageviews). https://dumps.wikimedia.org/.

**PopQA** created by Mallen et al. (2023). License: MIT. https://github.com/AlexTMallen/adaptive-retrieval

**The Pile** created by Gao et al. (2020). The Pile is a composite dataset consisting of multiple component datasets; licensing and usage terms vary by component. https://pile.eleuther.ai/.

**OLMo Mix 1124** created by OLMo et al. (2024). License: ODC-By v1.0; use is also subject to Common Crawl's Terms of Use. https://huggingface.co/datasets/allenai/olmo-mix-1124.

## C.2 Models

**Pythia** created by Biderman et al. (2023). License: Apache-2.0. https://huggingface.co/collections/EleutherAI/pythia-scaling-suite.

**OLMo 2** created by OLMo et al. (2024). License: Apache-2.0. https://huggingface.co/collections/allenai/olmo-2

**open-sci-ref-0.01 Pile** created by Nezhurina et al. (2025). License: Apache-2.0. https://huggingface.co/collections/open-sci/open-sci-ref-001-pile

**Llama 3.1** created by Grattafiori et al. (2024). License: Meta Llama 3.1 Community License. https://www.llama.com/llama3_1/

**Qwen3** created by Yang et al. (2025). License: Apache 2.0. https://huggingface.co/collections/Qwen/qwen3

**GPT-4o** created by OpenAI (2024b). License: Proprietary; access governed by OpenAI's Terms of Use.

**GPT-4o-mini** created by OpenAI (2024a). License: Proprietary; access governed by OpenAI's Terms of Use.

**fastText English word vectors (cc.en.300.bin)** created by Bojanowski et al. (2017). License: CC BY-SA 3.0. https://fasttext.cc/docs/en/crawl-vectors.html.

### C.3 Software

**DBpedia Spotlight** created by Mendes et al. (2011). License: Apache-2.0. https://github.com/dbpedia-spotlight/dbpedia-spotlight.

**fastText** created by Bojanowski et al. (2017). License: MIT. https://github.com/facebookresearch/fastText.