Bridging Linguistic Structure and Mechanistic Interpretability for Conceptual Interpretation in Language Models

Nura Aljaafari[1†], **Danilo S. Carvalho**[3], **André Freitas**[1,2,3]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] Idiap Research Institute, Switzerland
[3] CRUK National Biomarker Centre, University of Manchester, United Kingdom
{firstname.lastname}@[postgrad.]†manchester.ac.uk

Abstract

Understanding how language models compose meaning from linguistic input remains a central problem in interpretability research. Mechanistic studies have attributed functional roles to core transformer components; however, these findings derive largely from factual retrieval settings. Whether the same mechanisms support *conceptual interpretation*, the compositional mapping from definitional expressions to abstract meaning, remains insufficiently characterised. We introduce *DSRA* (Definitional Semantic Role Analysis), a methodology that applies causal tracing within the reverse dictionary task and augments restoration traces with definitional semantic roles (DSRs) grounded in Argument Structure Theory. This linguistic overlay identifies which compositional functions (e.g., genus, differentia quality) are associated with high-recovery states, extending activation patching beyond token-level localisation. Applied to GPT-J-6B (English) and BERTIN GPT-J-6B (Spanish), the results show that MLP layers associate content-bearing tokens with high-specificity DSR categories in early layers, MHA layers distribute integration across middle-to-upper layers with concentration at the final token, and hidden states aggregate information in upper layers. Alignment between restored states and DSR categories indicates systematic correspondence between internal activations and definitional structure, with consistent localisation patterns across both languages.

1 Introduction

Conceptual interpretation maps lexical items and syntactic structures to abstract meaning representations (Brown, 2006; Hirst, 1987). It supports sentence-level understanding and downstream tasks such as textual entailment and question answering. Transformer-based LMs achieve strong performance on these tasks, yet most evaluations focus on external accuracy and provide limited evidence about the internal computations that construct meaning. Consequently, the mechanisms by which models compose meaning from linguistic input remain insufficiently understood.

Mechanistic interpretability has identified stable functional roles for core transformer components. MLP layers have been characterised as key–value memory stores (Geva et al., 2021; Dai et al., 2022); attention heads within MHA modules often aggregate information toward the final-token position (Da et al., 2021); and upper layers concentrate representations that are critical for prediction (Li et al., 2022). Sparse autoencoders further decompose superimposed activations into approximately monosemantic features (Huben et al., 2023; Shu et al., 2025). However, these results are derived predominantly from factual retrieval prompts (e.g., "Eiffel Tower is located in [Paris]"). Such settings establish component-level behaviours, but they do not determine whether, or how, these behaviours support compositional conceptual interpretation, in which meaning depends on integrating definitional structure, thematic roles, and predicate–argument relations.

We introduce DSRA (Definitional Semantic Role Analysis), a methodology for studying definition-based inference in the reverse dictionary task. DSRA applies causal tracing (Meng et al., 2022) to localise internal states that are causally influential for predicting the *definiendum*, and it interprets restoration traces using definitional semantic roles (DSRs) (Silva et al., 2016). In the reverse dictionary task, a model receives a natural-language definition and predicts the single term it defines. For example, given "work done by one person or group that benefits another," a model should predict *service* (Figure 1). Definitions are well suited to this analysis because they exhibit systematic compositional structure: a broader category (the *genus*, e.g., *work*) is refined by distinguishing properties (the *differentia*, e.g., *done by one person*

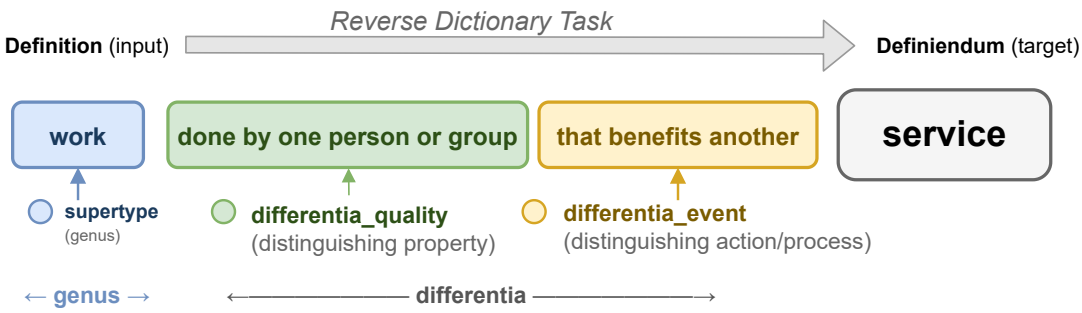


Figure 1: Definitional semantic labelling for the term *service*. Each coloured span is annotated with its Definitional Semantic Role (DSR) following the taxonomy of Silva et al. (2016). The DSR overlay provides an interpretive lens for assessing whether causally influential internal states align with definitional functions.

or group that benefits another), providing a controlled setting for studying how models integrate structured meaning into a single prediction.

The DSR taxonomy decomposes each definition into labelled spans, including *supertype* (the immediate broader category), *differentia_quality* (distinguishing intrinsic property), and *differentia_event* (distinguishing action or process), which correspond to the compositional functions of the constituents. By overlaying these labels onto causal traces, DSR annotation supports analyses that move beyond token identity and test whether restored internal states align with definitional functions.

Three research questions guide the study. **RQ1**: What consistent localisation patterns do causal traces exhibit during definition-based inference, and how do these patterns generalise across languages? **RQ2**: Is functional differentiation observable across LM components (MLP, MHA, hidden states) during definition-based inference? **RQ3**: To what extent does grouping restored states by DSR category and part of speech reveal structured alignment between internal activations and definitional structure? To clarify what would count as evidence of structure, we state the corresponding baselines. If recovery under causal tracing were not systematically localised, we would expect broadly uniform layer–token patterns and limited cross-lingual stability (RQ1). If component type were not functionally differentiated, we would expect MLP, MHA, and hidden-state restorations to exhibit similar localisation profiles (RQ2). If recovery were not aligned with definitional structure, grouping top-K restored states by DSR (and PoS) would not show systematic representation of particular roles (RQ3).

The contributions are threefold. First, DSRA provides a controlled experimental framework linking causal tracing to formal lexical semantics through the reverse dictionary task. Second, the DSR overlay reveals regularities that are not apparent from token identity alone: *differentia_quality* accounts for over 31% of the top-50 MLP restoration states, whereas *supertype* and *differentia_event* each contribute approximately 6%. Third, an integrated analysis across MLP, MHA, and hidden states in GPT-J-6B (English) and BERTIN GPT-J-6B (Spanish) shows consistent localisation patterns across both languages and clarifies how component-level behaviours jointly support definition-based inference.

2 Background

Natural language definitions. Natural language (NL) definitions provide a controlled setting for analysing conceptual interpretation in LMs. A definition consists of two parts: the *definiendum* (the term being defined) and the *definiens* (the expression that specifies its meaning). Dictionary-style definitions express the necessary and sufficient conditions and essential attributes of a term, and interpreting them requires syntactic processing, disambiguation, and semantic composition. At a high level, the goal of this work is to identify which parts of a definition contribute most to a model's prediction of the definiendum, and whether these contributions correspond to linguistically meaningful components such as category terms and distinguishing properties. Definitions provide a natural setting for this analysis because they encode structured meaning in a relatively controlled and interpretable form, enabling systematic comparison between internal model behaviour and linguistic structure.

The analysis focuses on *intensional definitions*, which specify meaning through conditions govern-

ing term usage rather than by enumeration of instances. Intensional definitions commonly exhibit *genus–differentia* structure (Silva et al., 2016): a broader category (the genus) is identified and then restricted by distinguishing properties (the differentia). For example, the term *service*, defined as "work done by one person or group that benefits another", contains *work* as genus and the remaining clause as differentia (Figure 1). This structure is systematic and linguistically principled, making it well-suited to controlled analysis of how models integrate structured meaning in internal representations.

Conceptual composition. To provide a formal basis for the DSR taxonomy used in this work, composition is examined under a Montagovian framework (Partee et al., 1984), in which the meaning of a complex expression is determined by the meanings of its parts and the way they are combined. Intuitively, a definition such as "work done by one person or group that benefits another" composes a category term (*work*) with a set of restricting properties (*done by one person...*) to yield the meaning of the definiendum (*service*). Sentence representation is formalised using Argument Structure Theory (AST) (Jackendoff, 1992; Levin, 1993; Rappaport Hovav and Levin, 2008), in which a predicate p is associated with arguments arg_i, each carrying a thematic role r_i (see App. A.6 Table 2 for the full DSR taxonomy). Following Silva et al. (2016), a definiens statement $s_{\text{definiens}}$ is modelled as a composition of predicate–argument structures. In latent space:

$$[[s_{\text{definiens}}]] = \underbrace{t_1(c_1, r_1)}_{\text{ARG0-genus}} \oplus \cdots \oplus \underbrace{t_i(c_i, r_i)}_{\text{ARGN-differentia-quality}}, \tag{1}$$

where $[[s_{\text{definiendum}}]] = [[s_{\text{definiens}}]]$ and $t_i(c_i, r_i) = c_i \otimes r_i$ follows compositional-distributional semantics notation (Smolensky and Legendre, 2006; Clark et al., 2008). The operator $\otimes$ binds lexical content to thematic role, and $\oplus$ composes lexical-semantic units into definition-level meaning.

Reverse dictionary task. A valid interpretation of a definition assigns meanings to the elements of the definiens such that the resulting representation evaluates correctly for the definiendum. For generative LMs, this corresponds to next-token prediction with the definition as input and the definiendum as the expected continuation.

Knowledge localisation. Methods including ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), and PMET (Li et al., 2024) modify factual associations by intervening on MLP representations. SAEs further reveal approximately monosemantic features (Huben et al., 2023; Shu et al., 2025). These findings motivate investigation of whether similar tools can characterise compositional conceptual interpretation beyond factual retrieval.

3 Methodology

3.1 Localisation via Causal Tracing

We adapt the causal tracing procedure of ROME (Meng et al., 2022) to the reverse dictionary task in order to localise internal states that are causally implicated in predicting the *definiendum*. Figure 2 summarises the clean, corrupted, and patched executions and the subsequent analyses. Let $\mathcal{M}$ denote a transformer with L layers that processes an input sequence X comprising a natural-language definition embedded in a fixed prompt template (e.g., "*definition* is often referred to as:"). The target is the correct definiendum token y predicted at the final position. For token position t at layer l, we denote the residual-stream representation as h_t^l. For concision, we write the residual update as:

$$h_t^l = h_t^{(l-1)} + \text{MHA}_t^l + \text{MLP}_t^l, \tag{2}$$

where MHA_t^l and MLP_t^l denote the residual contributions of the attention and feed-forward sublayers at layer l and position t, with all layer normalisation and architectural details inherited from the base model.

Causal tracing proceeds in three stages. In the *clean execution*, the model processes X and caches internal activations. In the *corrupted execution*, Gaussian noise is injected into the embedding representations of the definition span, yielding a corrupted input X^* that degrades the output distribution. In the *patched execution*, selected internal states from the corrupted run are restored to their corresponding clean-run values. Recovery toward the clean prediction indicates that the restored states are causally sufficient for predicting y under this intervention (Meng et al., 2022); that is, restoring these states is enough to recover the output. Conversely, limited recovery suggests that the restored states are not causally implicated. While this

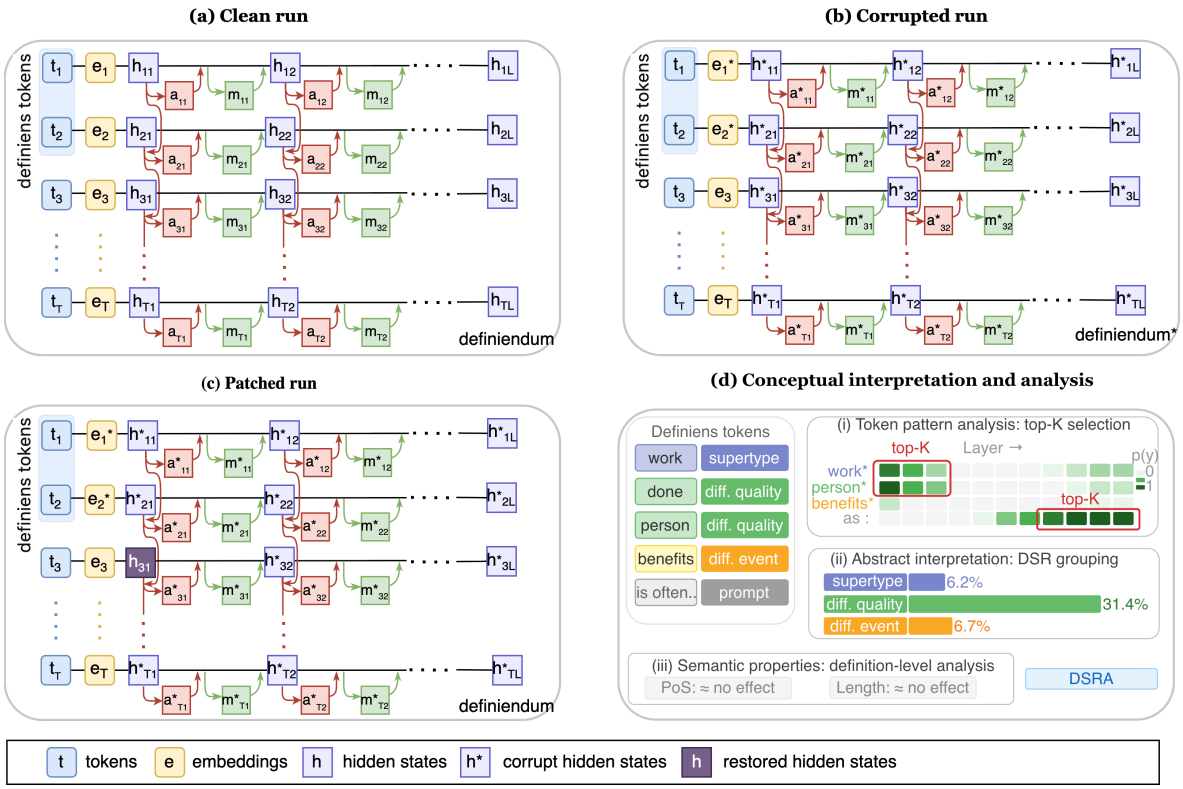


Figure 2: Overview of causal tracing and conceptual localisation in LMs. **(a)** Clean run: the model processes the original definition. **(b)** Corrupted run: Gaussian noise is added to definition-token embeddings. **(c)** Patched run: selected states are restored from clean-run values. **(d)** Conceptual interpretation analysis: restored states are grouped by token position, DSR label, and PoS. Model: GPT-J-6B (28 layers). Notation: t = tokens, e = embeddings, h = hidden states, h^* = corrupted states, $\bar{h}$ = restored states, a = attention, m = MLP.

establishes causal sufficiency at the level of individual component restorations, it does not characterise the full causal graph of interactions between components. The span-delimitation rule, noise specification, and restoration operators are provided in Appendix B.

3.2 Recovery Metric and Semantic Alignment

Let $p_y(X)$ denote the softmax probability assigned to the correct definiendum token y under the clean execution, and let $p_y(X^*)$ denote the corresponding probability under the corrupted execution. For component $k \in \{\text{MLP}, \text{MHA}, \text{hidden}\}$, restoration at token position i and layer index j yields probability $p_y(X^{*,\text{patch}(i,j,k)})$. Following Meng et al. (2022), we define the component-specific recovery score:

$$\mathbf{T}_{ij}^{(k)} = \frac{p_y(X^{*,\text{patch}(i,j,k)}) - p_y(X^*)}{p_y(X) - p_y(X^*)}. \quad (3)$$

This normalisation measures the fraction of corruption-induced degradation recovered by restoring a given internal state. Values are clipped to $[0, 1]$ for numerical stability, consistent with prior work. Each input is evaluated across 10 independent corrupted executions, and recovery scores are averaged.

Restoration granularity follows Meng et al. (2022). For MLP and MHA components, restoration is applied over a contiguous window of 10 layers centred at layer index j; accordingly, $\mathbf{T}_{ij}^{(\text{MLP})}$ and $\mathbf{T}_{ij}^{(\text{MHA})}$ summarise the effect of restoring the corresponding component across that layer window at token position i. For hidden states, restoration targets individual layer–token pairs. A formal specification of the restoration operators and normalisation is provided in App. C. Quantitative summaries use the top-K restored states per example, defined as the (i, j) pairs with the largest $\mathbf{T}_{ij}^{(k)}$ values. Unless stated otherwise, K=50 is used for reporting, and K=10 is reported for comparison. To aggregate traces across definitions of varying length, traces are length-normalised by resampling along the token dimension; details of the resampling grid and the length band are given in App. A.4. For

semantic alignment, tokens in the English setting are assigned DSR labels (Silva et al., 2016) and part-of-speech (PoS) tags. Top-K restored states are grouped by these categories, relating recovery patterns $\mathbf{T}$ to span-level semantic functions in the definition. Spanish definitions are analysed using PoS tags only because DSR labels are unavailable for SWN.

3.3 Datasets and Experimental Setup

The experimental design requires (i) a task with a verifiable target that elicits compositional semantic processing, (ii) datasets that provide structured definitional resources, and (iii) models that are sufficiently large to support per-example causal tracing. The reverse dictionary task satisfies (i) because it requires integrating multi-part definitional structure (e.g., genus, distinguishing properties, and relational content) into a single prediction and provides an unambiguous correctness criterion. This controlled formulation isolates compositional semantic processing from confounds such as multi-token generation variability, making it well-suited for per-example causal tracing at this scale. Whilst this task setting does not preclude the possibility that some high-recovery states reflect lexical co-occurrence rather than structured compositional processing, the use of DSR-based grouping enables analysis of whether recovery patterns align with functional definitional roles beyond individual token identity. Full dataset construction, filtering, and prompt selection are provided in Appendix A. The extent to which these patterns generalise to other tasks and model scales is discussed in Section 6.

Datasets. Two datasets are constructed from WordNet (Miller, 1994) and Spanish WordNet (Gonzalez-Agirre et al., 2012), referred to as EWN and SWN respectively. EWN contains 8 348 samples (80/20 train–test split), and SWN contains 7 815 samples (30/70 train–test split). Each sample in EWN is augmented with DSR labels (Silva et al., 2016).[1] In both settings, definienda are restricted to single tokens to ensure that prediction correctness is unambiguous.

Causal tracing is computed on the subset of test definitions that the fine-tuned models predict correctly under clean (unperturbed) conditions; filtering criteria and counts are provided in Appendix A. These resources were selected for their structured and consistent definitional format, which enables controlled compositional analysis; their known limitations are discussed in Section 6.

[1] DSR annotations were limited to EWN due to the unavailability of a comparable annotation resource for Spanish.

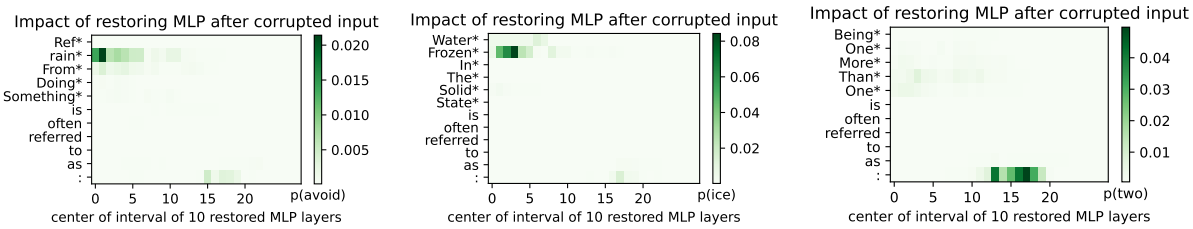
Models. The models are GPT-J-6B (Wang and Komatsuzaki, 2021) for English and BERTIN GPT-J-6B (la Rosa and Fernández, 2022) for Spanish, fine-tuned via LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2024) respectively. Both models share the same base architecture, which enables direct comparison of localisation patterns across languages while controlling for architectural variation; full specifications are provided in App. B.2.

4 Empirical Analysis

Figures 3, 4, and 5 follow the visualisation approach of Meng et al. (2022). The y-axis indexes input tokens (asterisks indicate corrupted definition tokens), and the x-axis indexes layer positions from 0 to $L-1$. Each cell reports a normalised recovery score in $[0, 1]$. Unless stated otherwise, quantitative summaries are based on the top-K restored states per example for $K \in \{10, 50\}$. To aggregate traces across definitions of varying length, global trends use length-normalised resampling over a common length band covering approximately 80% of examples (Appendix A.4). Layer-index distributions and additional grouped analyses are reported in Appendix D.

4.1 MLP: Content Association in Early and Upper Layers

Localisation pattern (RQ1). MLP traces show two recurring localisation regimes. High-recovery states appear in early layers at content-bearing tokens that are lexically or semantically related to the *definiendum* (Figure 3a and 3b). When the definition provides weaker lexical cues, recovery shifts toward upper layers and concentrates at the final token position (Figure 3c). Aggregated layer-index distributions exhibit a bimodal pattern with concentration in layers 0–5 and 17–21 (Appendix D). This structured concentration contrasts with a uniform distribution that would indicate limited localisation. The pattern is consistent with key–value retrieval behaviour reported for MLP layers (Geva et al., 2021; Dai et al., 2022). Evidence is also consistent with early tokenisation effects in which split tokens contribute to recovery when restored (Figure 3a; see also Appendix Figures 12a and 12b)



(a) (b) (c)

Figure 3: Causal traces for MLP restoration (GPT-J-6B, EWN). Each heatmap shows the normalised recovery score $\mathbf{T}^{(\mathrm{MLP})}$ when restoring a sliding window of 10 MLP layers at each input token (y-axis; asterisks denote corrupted definition tokens). (a) *avoid*: strong early-layer recovery at content-bearing tokens. (b) *ice*: a similar early-layer pattern. (c) *two*: recovery shifts toward upper layers and the final token when definition tokens provide weaker lexical cues.

DSR Label	MLP	MHA	Hidden
Top 50 states (%)			
diff. quality	31.4	34.0	32.0
diff. event	6.7	7.0	7.0
supertype	6.2	<1.0	<1.0
other DSR	6.7	4.0	10.0
prompt	49.0	54.0	51.0

Table 1: Distribution of DSR labels among top-50 causal tracing states per component type (EWN, GPT-J-6B, n=818 correctly predicted test samples). For each example and component, top-50 indices are the (i, j) pairs with the largest $\mathbf{T}^{(k)}_{ij}$ values; percentages are averaged across samples. Prompt tokens are included for completeness.

Component differentiation (RQ2). The bimodal layer concentration for MLP differs from the unimodal concentration observed for MHA (Section 4.2) and hidden states (Section 4.3), indicating differentiation across component types.

DSR alignment (RQ3). DSR-labelled MLP traces are provided in Appendix D.3. Quantitatively, *differentia_quality* accounts for over 31% of the top-50 restored states, followed by *differentia_event* and *supertype* at approximately 6.7% and 6.2% respectively (Table 1). A null model in which DSR proportions match corpus-level span frequencies would predict substantially weaker concentration; the observed dominance of *differentia_quality* therefore supports systematic correspondence between high-impact states and definitional structure.

4.2 MHA: Distributed Integration in Middle-to-Upper Layers

Localisation pattern (RQ1). MHA traces consistently concentrate at the final token, with influential states distributed across middle-to-upper layers (12–25) (Figure 4). Aggregated layer-index distributions show unimodal, negatively skewed concentration (Appendix D), contrasting with the bimodal MLP pattern.

Component differentiation (RQ2). Prompt tokens dominate among top-10 states, whereas content-bearing roles enter the high-impact set when considering top-50 states (Table 1). This pattern is consistent with gradual information aggregation during definition-based inference, with definition content contributing more strongly at broader restoration thresholds.

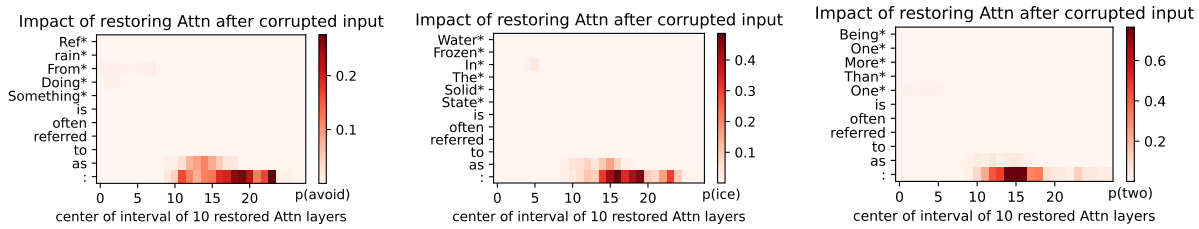
DSR alignment (RQ3). DSR-labelled MHA traces are provided in Appendix D.3. The shift from prompt-dominated top-10 states to DSR-represented top-50 states indicates progressive integration of definition content.

4.3 Hidden States: Upper-Layer Concentration

Localisation pattern (RQ1). Hidden-state traces concentrate in upper layers (20–27) and at the final token (Figure 5). Aggregated layer-index distributions show unimodal concentration in the top layers (Appendix D). Tokens outside the prompt generally show low impact unless they are lexically related to the *definiendum*.

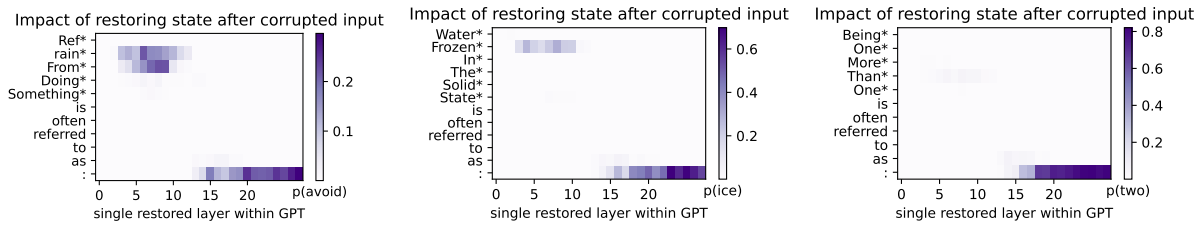
Component differentiation (RQ2). The layer concentration for hidden states is more strongly upper-layer than the concentration for MHA, consistent with hierarchical aggregation into the final representation.

DSR alignment (RQ3). *differentia_quality* (32%) and *differentia_event* (7%) remain present among top-50 restored states (Table 1), indicating persistence of definitional information into the



(a) (b) (c)

Figure 4: Causal traces for MHA restoration (GPT-J-6B, EWN). Each heatmap shows the normalised recovery score $\mathbf{T}^{(\mathrm{MHA})}$ when restoring a sliding window of 10 MHA layers (x-axis: window centre, layers 0–27) at each input token (y-axis; asterisks denote corrupted definition tokens). Influential states concentrate at the final token across middle-to-upper layers.



(a) (b) (c)

Figure 5: Causal traces for hidden-state restoration (GPT-J-6B, EWN). Each heatmap shows the normalised recovery score $\mathbf{T}^{(\mathrm{hidden})}$ (colour scale, clipped to $[0, 1]$) when restoring a single hidden state (x-axis: layer index 0–27) at each input token (y-axis; asterisks denote corrupted definition tokens). Influential states concentrate in upper layers and at the final token.

final representation. DSR-labelled MLP, MHA and hidden-state traces are provided in Appendix D.3.

4.4 Validation and Cross-Lingual Stability

Alternative definitions. Alternative definitions are generated as described in Appendix A.6, with five alternatives per sample. Traces computed on correctly predicted alternatives preserve the broad localisation patterns while shifting specific high-impact positions across paraphrases (see Figures 12–14 in Appendix D).

Cross-lingual stability (RQ1). Experiments replicated on SWN with BERTIN GPT-J-6B yield localisation patterns consistent with the English analysis (see Figure 17 in Appendix D). Differences in the distribution of high-impact MLP positions are consistent with tokenisation variation while preserving component-level behaviour. Cross-lingual comparison is therefore limited to localisation patterns, as DSR-based semantic alignment is evaluated only in the English setting.

4.5 Synthesis: A Coarse-Grained Account

Figure 6 summarises a coarse-grained account of coordinated behaviour across MLP, MHA, and hidden states. Overall, the results are inconsistent with simple baselines. MLP recovery concentrates in two regimes rather than exhibiting approximately uniform layer occupancy. In addition, localisation profiles differ across components: MLP bimodality contrasts with unimodal MHA and hidden-state concentration, indicating component differentiation. Under DSR grouping, *differentia_quality* is highly represented among high-impact states relative to frequency-matched expectations, suggesting a systematic correspondence between internal states and definitional structure rather than span prevalence alone. This correspondence is indicative of structured semantic alignment, though it does not by itself establish a complete account of compositional computation. These trends persist under paraphrase and replicate in Spanish, arguing against artefacts of surface form or language. MLP traces emphasise content association in early layers and shift toward upper-layer concentration at the final token when lexical cues are weak. MHA traces implement distributed aggregation centred on the final token across middle-to-upper layers. Hidden-state traces concentrate recovery in upper layers while retaining definitional structure under DSR

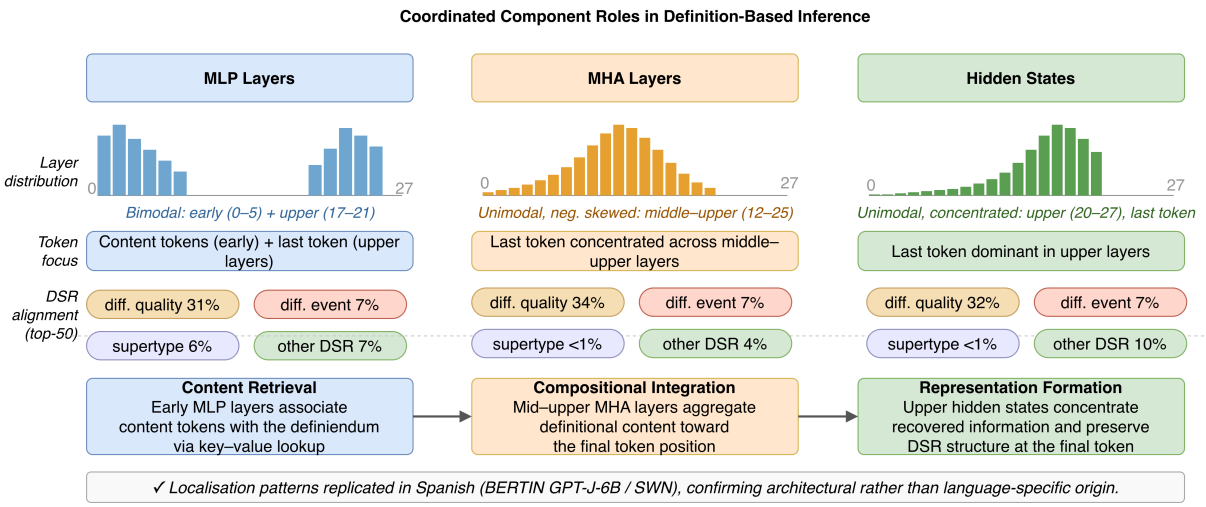


Figure 6: Coarse-grained synthesis of localisation results across MLP, MHA, and hidden-state components (GPT-J-6B, EWN). Bars schematically summarise layer-wise concentration patterns, and DSR percentages are taken from the top-50 analysis (Table 1). Arrows indicate a proposed directional account of information aggregation consistent with the observed traces.

grouping. Testing this account with finer-grained causal interventions is left to future work.

5 Related Work

Mechanistic interpretability (MI). MI decomposes neural networks into human-understandable computational structures by identifying functional roles for internal components (Conmy et al., 2023). Probing studies have shown that transformer layers encode linguistic knowledge in a hierarchical progression (Tenney et al., 2019), establishing that internal representations align with structured linguistic abstractions. SAEs decompose superimposed activations into approximately monosemantic features (Huben et al., 2023; Shu et al., 2025) and have been applied to code correctness (Tahimic and Cheng, 2025), hallucination detection (Xiong et al., 2026), and causal representation learning (Song et al., 2025). While probing identifies the presence of linguistic information and SAEs isolate interpretable features, neither approach directly characterises how such information is composed and integrated across components during meaning construction. The present work addresses this gap by applying intervention-based analysis to compositional definitional processing.

Activation patching and causal tracing. Causal mediation analysis was first applied to transformer internals by Vig et al. (2020), who treated attention heads as causal mediators of gender bias. Building on this framework, activation patching modifies and restores internal activations to identify causally relevant states (Heimersheim and Nanda, 2024). ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023) apply this approach to knowledge localisation in MLP layers; Causal Scrubbing (Chan et al., 2022) and Attribution Patching (Nanda, 2023) refine causal analysis at scale. Causal abstraction methods further align neural computations with formal semantic structures (Geiger et al., 2021), but have not been applied to definitional interpretation or lexical-semantic role taxonomies. Among available localisation methods, ROME-style causal tracing is adopted because it enables intervention-based localisation of internal representations at the component level, providing a direct and well-established approach for analysing decoder-only transformers at this scale (Meng et al., 2022), in contrast to gradient-based attribution methods, which estimate rather than directly intervene on internal states. DSRA extends this framework to conceptual interpretation by integrating DSR labels into restoration traces, enabling attribution at the level of semantic roles rather than token positions alone.

Logit attribution and circuit analysis. Logit attribution links predictions to input tokens via logit contributions (Elhage et al., 2021) and has been applied to feed-forward layers (Geva et al., 2022) and MHA heads (Ferrando et al., 2023). More recent attribution-based approaches, such as information flow routes (Ferrando and Voita, 2024), scale circuit discovery via single forward-pass attribution but, like logit attribution, rely on gradient-based

estimates rather than direct causal intervention. Circuit analysis identifies functional subgraphs for specific behaviours (Olsson et al., 2022; Wang et al., 2022; Conmy et al., 2023), typically focusing on individual attention heads and edge-level structures rather than the joint contribution of multiple component types. DSRA differs in examining MLP, MHA, and hidden-state components jointly under direct intervention, and relates recovery patterns to span-level definitional semantic structure rather than token-level logit contributions.

Knowledge localisation in MLP layers. MLP layers have been characterised as key–value memory stores (Geva et al., 2021), and Dai et al. (2022) demonstrated that specific MLP neurons directly affect factual recall, providing causal evidence for content-specific storage. These findings raise the question of whether MLP layers play analogous content-association roles during compositional conceptual interpretation, where meaning is derived from structured definitional input rather than a single factual association.

Definition modelling and reverse dictionary. Definition modelling, introduced by Noraset et al. (2017), established the task of generating dictionary definitions from word embeddings. Bevilacqua et al. (2020) extended this to pretrained transformers, demonstrating that fine-tuned models capture contextually appropriate definitional knowledge. More recently, Xu et al. (2024) used the reverse dictionary task as a probe for LLM conceptual inference, showing that model representations encode object categories and fine-grained features; however, this work relies on probing and in-context learning rather than causal intervention, and does not examine internal component-level behaviour. DSRs (Silva et al., 2016) offer a structured framework for analysing definitions through genus–differentia structure and predicate–argument relations. This work addresses the gap at the intersection of these lines: to our knowledge, intervention-based causal analysis has not yet been applied to the reverse dictionary task to explore how definitional semantic structure may be reflected in transformer component behaviour. DSRA fills this gap by integrating the DSR lexical-semantic framework with mechanistic interpretability, moving beyond probing to causal localisation of definitional structure within transformer components.

6 Conclusion

This work demonstrates the value of integrating formal linguistic structure into mechanistic interpretability through DSRA, which augments causal tracing with definitional semantic roles within the reverse dictionary task. The focus of the present work is localisation and semantic alignment; causal intervention for model editing is left to future work.

With respect to **RQ1**, causal traces exhibit consistent localisation patterns: MLP states cluster bimodally in early and upper layers, MHA states concentrate in middle-to-upper layers with a negatively skewed distribution, and hidden states concentrate in upper layers at the final token. Similar localisation patterns are observed in both English and Spanish. With respect to **RQ2**, the three component types exhibit distinct distributional signatures, bimodal (MLP), unimodal upper-skewed (MHA), and concentrated upper-layer (hidden), supporting functional differentiation in definition-based inference. With respect to **RQ3**, DSR grouping reveals that content-bearing semantic roles, particularly *differentia_quality*, account for a substantial fraction of high-impact states, suggesting that semantic role structure provides additional explanatory structure beyond token identity.

The principal methodological contribution is the DSR overlay for causal tracing, which functions as an interpretive lens bridging activation patching and formal lexical semantics. Future work may pursue finer-grained tracing at the level of individual attention heads, integrate SAEs for feature-level DSR alignment, and extend the framework beyond definitional interpretation to assess generality.

Limitations

This study is limited to a single scale of autoregressive transformer model and focuses on sentence-level meaning with single-token prediction, which does not fully reflect the complexities of conceptual processing across other tasks. The analysis examines each component type individually rather than investigating causal relationships between components, and MHA is studied as a whole rather than at the level of individual heads. DSR annotations are available only for the English dataset. Selecting causal tracing as the primary method may have precluded other valuable analytical perspectives. The experimental design relies on WordNet and Spanish WordNet, which provide a controlled setting but carry known limitations, such as sense distinc-

tions that could be overly fine-grained, and hierarchical structures that do not always correspond to cross-linguistically stable semantic categories. These factors may affect DSR annotation quality and the generalisability of the observed localisation patterns. As WordNet definitions are more structured and templated than naturally occurring text, it remains an open question whether the observed alignment between internal activations and definitional structure would persist with explanatory dictionaries or corpus-derived definitions exhibiting freer linguistic variation.

Ethical Considerations

This study applies DSRA using WordNet, abstract labelling, and causal tracing to analyse conceptual interpretation mechanisms in transformer-based models. The use of WordNet carries potential risks, including biases inherent in the dataset. Abstract labelling may introduce distorted representations. The identified mechanisms could potentially be employed to alter model behaviour in unintended ways. Code and datasets are released to enable replication and verification.

Acknowledgements

This work was partially funded by the SNSF project RATIONAL (200021E_229196), the CRUK National Biomarker Centre, and supported by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.

References

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "How we went beyond word sense inventories and learned to gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7207–7221. Association for Computational Linguistics.

Keith Brown. 2006. *Encyclopedia of language and linguistics (2nd Edition)*, volume 1. Elsevier.

Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In *AI Alignment Forum*, pages 1828–1843.

Stephen Clark, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 133–140. Oxford.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Jeff Da, Ronan Le Bras, Ximing Lu, Yejin Choi, and Antoine Bosselut. 2021. Analyzing commonsense emergence in few-shot knowledge models. In *Conference on Automated Knowledge Base Construction*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Javier De la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. BERTIN: Efficient pretraining of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1:1.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2101.00027*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build

predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Graeme Hirst. 1987. *Semantic interpretation and the resolution of ambiguity*. Cambridge University Press.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Ray S Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

Javier De la Rosa and Andres Fernández. 2022. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18564–18572.

Zhen Li, Xiting Wang, Weikai Yang, Jing Wu, Zhengyan Zhang, Zhiyuan Liu, Maosong Sun, Hui Zhang, and Shixia Liu. 2022. A unified understanding of deep nlp models for text classification. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4980–4994.

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.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass editing memory in a transformer. *The Eleventh International Conference on Learning Representations (ICLR)*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. *URL: https://www. neelnanda. io/mechanistic-interpretability/attribution-patching*.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 3259–3266.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

OpenAI. 2023. ChatGPT: Optimizing language models for dialogue.

Barbara Partee et al. 1984. Compositionality. *Varieties of formal semantics*, 3:281–311.

Malka Rappaport Hovav and Beth Levin. 2008. The english dative alternation: The case for verb sensitivityl. *Journal of linguistics*, 44(1):129–167.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1690–1712, Suzhou, China. Association for Computational Linguistics.

Vivian Silva, Siegfried Handschuh, and André Freitas. 2016. Categorization of semantic roles for dictionary definitions. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 176–184, Osaka, Japan. The COLING 2016 Organizing Committee.

Paul Smolensky and Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT press.

Xiangchen Song, Jiaqi Sun, Zijian Li, Yujia Zheng, and Kun Zhang. 2025. LLM interpretability with identifiable temporal-instantaneous representation. In *The*

Thirty-ninth Annual Conference on Neural Information Processing Systems.

Kriz Tahimic and Charibeth Cheng. 2025. Mechanistic interpretability of code correctness in llms via sparse autoencoders. *arXiv preprint arXiv:2510.02917*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Guangzhi Xiong, Zhenghao He, Bohan Liu, Sanchit Sinha, and Aidong Zhang. 2026. Toward faithful retrieval-augmented generation with sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*.

Ningyu Xu, Qi Zhang, Menghan Zhang, Peng Qian, and Xuanjing Huang. 2024. On the tip of the tongue: Analyzing conceptual representation in large language models with reverse-dictionary probe. *arXiv preprint arXiv:2402.14404*.

A Dataset Creation, Filtering, and Prompt Selection

Three eligibility conditions governed dataset and model choices. First, the task must elicit compositional semantic processing with a verifiable target. Second, the datasets must support structured definitional analysis, including span-level labels when available. Third, the models must be compatible with per-example causal tracing under the available compute budget. This section documents dataset construction, filtering, prompt selection, and trace aggregation.

A.1 English WordNet (EWN)

The EWN dataset was constructed from Princeton WordNet (Miller, 1994) by selecting definienda consisting only of alphabetical characters and restricting definitions to a maximum of 25 words for computational efficiency. Samples were retained only if the definiendum matched a single token in the model vocabulary, confining predictions to single-token inferences. This yielded 8,348 examples, split 80/20 for training and testing.

A.2 Spanish WordNet (SWN)

A similar procedure was followed for SWN using the Open Multilingual Wordnet (Gonzalez-Agirre et al., 2012). Alphabetic definienda were selected and their definitions were queried in Spanish WordNet. When a Spanish definition was available, it was selected, cleaned, and used. When no Spanish definition was found, the Spanish definiendum was translated into English, and an English WordNet entry was retrieved with a matching part of speech. The retrieved English definition was then translated into Spanish using the Googletrans library,[2] and back translation was used as a validation check for translation correctness. This procedure yielded 7,815 examples. Definitions were restricted to a maximum of 25 words, covering 99% of samples. The 30/70 train–test split was chosen to maximise the number of held-out definitions available for causal tracing while retaining sufficient training data for stable task performance under parameter-efficient fine-tuning.

A.3 Preprocessing and Prompt Selection

Inputs consist of the definition text with a prompt template appended. Three prompt formats were evaluated: a direct instruction template (“`Identify the term defined as: \{definition\}.`”), a question template (“`What word is described by this definition: \{definition\}?`”), and a completion-style template (“`\{definition\} is often referred to as:`”). Additional metadata (e.g., the part of speech of the definiendum) was also tested. Simple prompts without additional metadata yielded the highest accuracy. For causal tracing analysis,

[2] https://github.com/ssut/py-googletrans

only examples correctly predicted by the fine-tuned models under clean (unperturbed) conditions are included. Additional preprocessing details are provided in the released code repository.[3]

A.4 Length Normalisation

To aggregate traces across examples of varying definition length, the token dimension of each trace tensor is resampled to a fixed grid within a common length band covering approximately 80% of examples. The band is defined by the empirical 10th and 90th percentiles of tokenised definition length under the analysis prompt. Traces outside the band are excluded from length-normalised aggregations. Resampling uses linear interpolation over token index to preserve coarse positional structure while enabling global layer-wise summaries.

A.5 Definitional Semantic Role Annotation

DSR labels follow the taxonomy of Silva et al. (2016). Annotations are available for English WordNet definitions only; Spanish WordNet definitions are analysed without DSR alignment. Table 2 describes each DSR used in this work along with its definition.

Role	Definition
supertype	The superclass of the immediate entity or an ancestor of it
differentia quality	A fundamental, intrinsic attribute that distinguishes the entity from others within the same supertype
differentia event	An action, state, or process in which the entity engages, necessary for differentiation from others within the same supertype
event time	The time at which a differentia event occurs
event location	The specific location of a differentia event
quality modifier	A modifier indicating degree, frequency, or manner that refines a differentia quality
origin location	The place of origin of the entity
purpose	The primary objective behind the existence or occurrence of the entity
associated fact	A fact connected to the existence or occurrence of the entity, serving as an incidental attribute
accessory determiner	A determiner expression that does not restrict the scope of supertype–differentia
accessory quality	A non-essential attribute for characterising the entity
[role] particle	A particle not contiguous with other role components

Table 2: Definitional semantic roles and their definitions following the taxonomy of Silva et al. (2016).

A.6 Alternative Definitions

Five alternative definitions per sample are generated using GPT-3.5 (OpenAI, 2023) with the following prompt instruction: *Given the word "{definiendum}" and its dictionary definition "{definition}", generate five alternative definitions that: (1) preserve the core meaning of the original definition, (2) do not include the word itself or direct morphological variants and (3) remain plausible as dictionary-style entries. Return only the five definitions, numbered 1–5.*
Alternatives are manually filtered to ensure semantic similarity to the original definition, plausibility as standalone dictionary entries, and absence of the definiendum in the generated text. Only alternatives predicted correctly by the fine-tuned models are included in causal tracing to match the clean-evaluation filter described in Appendix A.3.

B Implementation Details

B.1 Causal Tracing Specification

This subsection specifies the corruption and restoration operators used in causal tracing. The definition span is delimited as the token subsequence corresponding to `\{definition\}` in the prompt template (Appendix A.3). Corruption injects independent Gaussian noise into the token embedding vectors of the delimited span in the corrupted execution. The noise scale and any exclusions (e.g., prompt tokens and the final prediction position) are fixed across experiments and reported with the released code.

For restoration, we follow Meng et al. (2022). For component $k \in \{\text{MLP}, \text{MHA}\}$, restoration at (i, j, k) replaces the corresponding component activations at token position i across a contiguous 10-layer window centred at j with their clean-run values. For hidden-state restoration, patching replaces the residual-stream representation at a single layer-token pair. Boundary handling for the layer window and the exact operator definitions are provided in App. C.

B.2 Models and Selection Rationale

Table 3 summarises the architecture hyperparameters of the models used. Model selection is constrained by the requirements of intervention-based

[3] anonymised

causal tracing. First, the model must permit extraction and restoration of internal activations at the level of MLP, MHA, and residual-stream states, since the tracing protocol quantifies the causal contribution of restored component activations. Second, the model must operate at a scale for which component-level localisation patterns under causal interventions have been shown to be stable and interpretable in prior mechanistic work (Geva et al., 2021; Meng et al., 2022). Third, to support cross-lingual comparison, the models must share the same base architecture so that observed differences in localisation patterns are not attributable to architectural variation.

GPT-J-6B is used for English because it is an open-weight, decoder-only transformer that satisfies these constraints while remaining computationally feasible for per-example causal tracing. GPT-J-6B is trained using Mesh Transformer JAX (Wang, 2021) on the Pile dataset (Gao et al., 2020). BERTIN GPT-J-6B is used for Spanish because it instantiates the same GPT-J architecture and is trained on Spanish data (mC4-es-sampled (Gaussian)) (De la Rosa et al., 2022). This pairing provides architectural control over layer depth, hidden dimensionality, and attention configuration (Table 3), enabling cross-lingual analyses in which language and tokenisation constitute the primary sources of variation.

Models are loaded via the HuggingFace Transformers library in evaluation mode with gradients disabled; half-precision is used when required to satisfy memory constraints. Both models are fine-tuned for the reverse dictionary task using parameter-efficient adaptation via LoRA (Hu et al., 2021), with settings reported in Table 4. This selection therefore prioritises methodological validity for causal tracing, architectural comparability for cross-lingual analysis, and practical feasibility for exhaustive tracing at the level of individual examples.

B.3 Experimental Setup

Experiments were conducted on an NVIDIA RTX A6000 GPU using Python 3.11.7, PyTorch 2.2.2, Hugging Face Transformers 4.39.0, and PEFT 0.8.2. Data processing relied on Datasets 2.16.1, NumPy 1.26.3, Pandas 2.2.0, and scikit-learn 1.4.0. Text processing used NLTK 3.8.1 and spaCy 3.7.2.

Hyperparameter	Value
$n_{\text{parameters}}$	6,053,381,344
n_{layers}	28*
d_{model}	4096
d_{ff}	16384
n_{heads}	16
d_{head}	256
n_{ctx}	2048
n_{vocab}	50257 / 50400†
Positional Encoding	RoPE
RoPE Dimensions	64

Table 3: Architecture hyperparameters for GPT-J-6B and BERTIN-GPT-J-6B. Both models share the same base architecture. *Includes one embedding layer and 27 transformer blocks. †50,257 used tokens / 50,400 vocabulary size including padding.

Setting	GPT-J-6B	BERTIN-GPT-J-6B
Epochs	10	20
Batch size	16	16
LoRA rank (r)	64	16
LoRA α	32	32
Dropout	0.1	0.3

Table 4: LoRA fine-tuning hyperparameters for the reverse dictionary task.

C Formal Background and Tracing Protocol

C.1 Transformer Formalisation

We summarise the standard transformer architecture (Vaswani et al., 2017) to establish notation used throughout the paper. A decoder-only transformer with L layers processes an input sequence $X = (x_1, \ldots, x_T)$ by computing a sequence of hidden representations. At each layer $l \in \{1, \ldots, L\}$ and token position t, the hidden state is updated via the residual stream:

$$h_t^{(l)} = h_t^{(l-1)} + \text{MHA}_t^l + \text{MLP}_t^l, \tag{4}$$

where $h_t^{(0)}$ is the sum of the token embedding and positional encoding for x_t. The multi-head attention (MHA) term concatenates the outputs of H independent attention heads:

$$\text{MHA}_t^l = W_O^l \, \text{Concat}_{h=1}^{H} \Big(\text{Attn}^{l,h}(h_{\leq t}^{(l-1)}) \Big), \tag{5}$$

where each head computes scaled dot-product attention over all positions $\leq t$ (causal masking) and $W_O^l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ is the output projection. The MLP term operates on the mid-layer residual $h_t^{\text{mid},l} = \text{LayerNorm}(h_t^{(l-1)} + \text{MHA}_t^l)$:

$$\text{MLP}_t^l = W_{\text{proj}}^l \, \sigma\Big(W_{\text{fc}}^l \, h_t^{\text{mid},l} \Big), \tag{6}$$

where σ denotes a non-linear activation function, $W_{\text{fc}}^l \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ is the up-projection, and $W_{\text{proj}}^l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ is the down-projection. The final prediction is obtained by applying a language modelling head to the last-position hidden state $h_T^{(L)}$.

C.2 ROME Formalisation

Rank-One Model Editing (ROME) (Meng et al., 2022) provides the causal tracing framework adapted in this work. ROME interprets MLP layers as implicit key–value stores. At layer l, the MLP output for token t can be written as:

$$\text{MLP}_t^l = W_{\text{proj}}^l \, \sigma\left(W_{\text{fc}}^l \, h_t^{\text{mid},l}\right) = W_{\text{proj}}^l \, k_t^l, \quad (7)$$

where $k_t^l = \sigma(W_{\text{fc}}^l \; h_t^{\text{mid},l}) \in \mathbb{R}^{d_{\text{ff}}}$ serves as the key vector and each column of W_{proj}^l serves as a value vector. Under this interpretation, the MLP computes a weighted combination of stored values indexed by the input-dependent key. ROME exploits this structure to locate factual associations by identifying the layer l^* and subject token t^* at which restoring clean-run activations maximally recovers the correct prediction. The present work adapts the localisation procedure to the reverse dictionary setting; model editing is not performed.

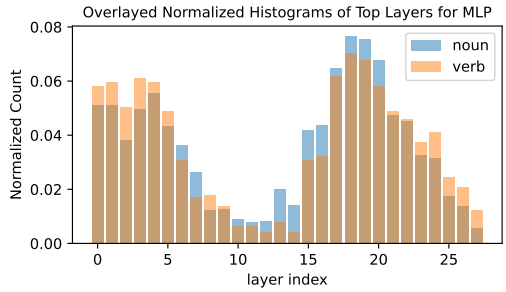
C.3 Causal Tracing Protocol

The causal tracing procedure follows Meng et al. (2022) and proceeds in three stages.

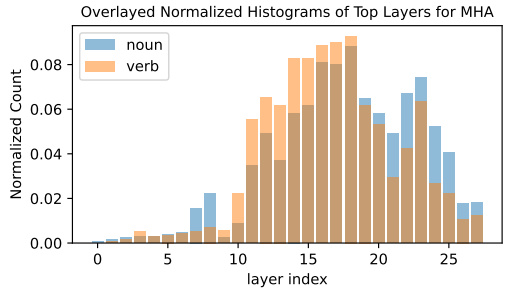
Stage 1: Clean execution. The model processes the original input X and caches internal states $\{h_t^{(l)}, \text{MHA}_t^l, \text{MLP}_t^l\}$ for all layers l and token positions t. The clean-run probability assigned to the correct definiendum token y is recorded as $p_y(X)$.

Stage 2: Corrupted execution. Gaussian noise $\mathcal{N}(0, (3\sigma_e)^2)$ is added to the token embeddings of the definition span only, where σ_e is the empirical standard deviation of embedding activations computed over the dataset. This produces a corrupted input X^* and a degraded prediction probability $p_y(X^*)$. Each input is evaluated using one clean run and 10 corrupted runs with independent noise samples.

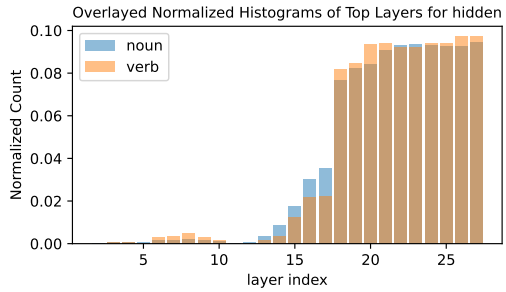
Stage 3: Restoration (patching). For each component type $k \in \{\text{MLP}, \text{MHA}, \text{hidden}\}$, the corrupted activation at a target position (i, j) is replaced with the corresponding clean-run value, yielding $p_y(X^{*,\text{patch}(i,j,k)})$. This patched probability is used to compute the normalised recovery score in Eq. 3. For MLP and MHA components, restoration is applied over a sliding window of 10 consecutive layers centred at layer index j, matching the protocol of Meng et al. (2022). For hidden states, restoration targets individual layer–token pairs.



(a)



(b)



(c)

Figure 7: Top-10 restored states from (a) MLP, (b) MHA, and (c) hidden components, grouped by the part of speech of the definiendum (GPT-J-6B, EWN, $K = 10$). Normalised histograms are overlaid. X-axis: layer index; y-axis: normalised count.

Recovery scores are averaged across the 10 corrupted runs to obtain stable estimates. Quantitative summaries are based on the top-K restored states per example, defined as the (i, j) pairs with the largest recovery scores $\mathbf{T}_{ij}^{(k)}$ (Eq. 3). Unless stated otherwise, $K = 50$ is used for reporting, with $K = 10$ provided for comparison.

D Supplementary Results

D.1 PoS Grouping

Figure 7 presents overlaid normalised histograms of top-10 restored states grouped by the part of speech of the definiendum (nouns vs. verbs). The distributions are broadly similar across all three component types, indicating that the PoS of the definiendum has minimal effect on localisation patterns. Minor differences are observable: in MLP layers, noun-related states are slightly more concentrated in upper layers compared to verbs; in MHA layers, the distribution for verbs approximates a Gaussian shape while that for nouns exhibits slight negative skew; no clear differences are discernible in hidden layers. These observations suggest that the localisation patterns reported in Section 4 are robust to variation in definiendum PoS and are therefore driven primarily by definitional content rather than the grammatical category of the target word.

D.2 Correlation with Definition Length

Figure 8 plots the layer indices of top-10 restored states against input length for each component type. No systematic association between definition length and the layer distribution of influential states is observed across MLP, MHA, or hidden components. This result indicates that the localisation patterns reported in Section 4 are not confounded by input length and that the models apply consistent processing strategies across definitions of varying complexity.

D.3 DSR-Labelled MHA and Hidden-State Traces

Figures 9, 10 and 11 show the DSR-grouped recovery patterns for MLP, MHA and hidden-state components respectively. For MHA restoration (Figure 10), recovery is concentrated at prompt tokens across middle-to-upper layers, consistent with the aggregation pattern reported in Section 4. DSR-labelled definition tokens contribute comparatively low recovery at the top-10 threshold, but their contribution increases at broader aggregation levels (Table 1), indicating that MHA integrates definitional content progressively rather than at discrete layer–token positions. For hidden-state restoration (Figure 11), recovery is dominated by upper-layer states at the final token, with *supertype* and differentia tokens exhibiting localised contributions in layers 20–27. These patterns are consistent with the upper-layer concentration reported in Section 4 and support the observation (RQ3) that definitional semantic structure persists through aggregation into the final representation.

D.4 Alternative Definitions

Figures 12–14 present causal traces for three alternative definitions of the definiendum *alone*, generated as described in Appendix A.6. These traces serve as a robustness check: if localisation patterns are artefacts of specific surface forms, they should vary substantially across semantically equivalent paraphrases.

For MLP restoration (Figure 12), the broad bimodal pattern (early and upper layers) is preserved across all three alternatives, although the specific token positions associated with high recovery shift in accordance with lexical variation. This is consistent with MLP layers encoding content-specific associations that depend on token identity while maintaining a stable layer-level distribution. For MHA restoration (Figure 13), patterns are more stable across paraphrases, with recovery consistently concentrated at the last token in middle-to-upper layers regardless of surface form. For hidden-state restoration (Figure 14), upper-layer concentration at the final token is consistent across all three alternatives, confirming that this pattern reflects a structural property of the aggregation process rather than a lexical artefact.

D.5 Supplementary English Results

Figure 15 presents additional causal tracing results that illustrate two phenomena: homonym handling and numerical concept processing.

Figures 15a and 15b show traces for two distinct definitions of the definiendum *baby*. For the second definition, which contains words strongly linked to the definiendum (e.g., *child*, *fetus*), hidden-state recovery is concentrated entirely at the last token (Figure 15b). In contrast, for the first definition, which lacks such direct lexical cues, additional definition tokens participate in high-recovery states (Figure 15a). MLP layers exhibit greater sensitivity to semantically related words: when the definition contains tokens with strong lexical association to the definiendum, multiple early-layer token positions contribute to recovery (Figure 15e). MHA layers maintain consistent last-token concentration across both definitions (Figures 15g and 15h), further supporting the component differentiation reported in Section 4 (RQ2).

Number handling provides an additional test

case. For the definiendum *billion* (Figures 15c and 15f), recovery is not associated with tokens containing other numerical values (e.g., *million*, *one*). Instead, the model treats spelled-out numbers as conceptual units rather than attending to their arithmetic content, consistent with the compositional interpretation account in which meaning is derived from definitional structure rather than token-level numerical matching.

D.6 Supplementary Spanish Results

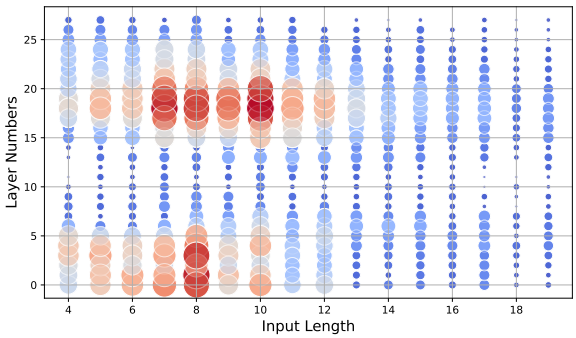
Figure 17 presents causal traces for three definienda (*lying*, *leave*, *hospital*) from SWN using BERTIN GPT-J-6B. These results replicate the English analysis and provide evidence for the cross-lingual stability of localisation patterns reported in Section 4.4 (RQ1).

MLP traces (Figure 17, a–c) exhibit early-layer recovery at content-bearing tokens, consistent with the bimodal pattern observed for GPT-J-6B on EWN. The distribution of high-impact positions is slightly broader than in the English setting, which is consistent with increased subword fragmentation in the Spanish tokenisation (e.g., *Deliberado* → *Del**, *iber**, *ado**). MHA traces (Figure 17, d–f) show last-token concentration in middle-to-upper layers, mirroring the English pattern with no qualitative differences. Hidden-state traces (Figure 17, g–i) confirm upper-layer concentration at the final token across all three examples.

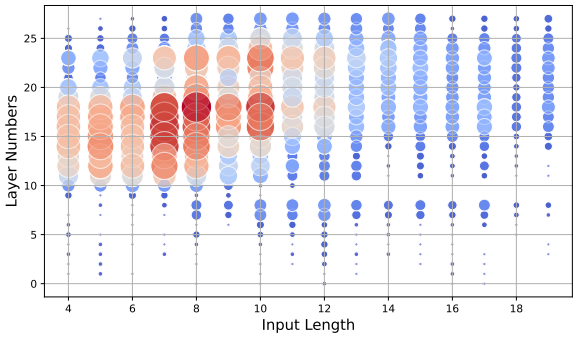
The consistency of these patterns across languages and tokenisation schemes supports the interpretation that the traced behaviours reflect architectural properties of the transformer components rather than artefacts of English-specific lexical structure.

D.7 Layer-Index Distributions

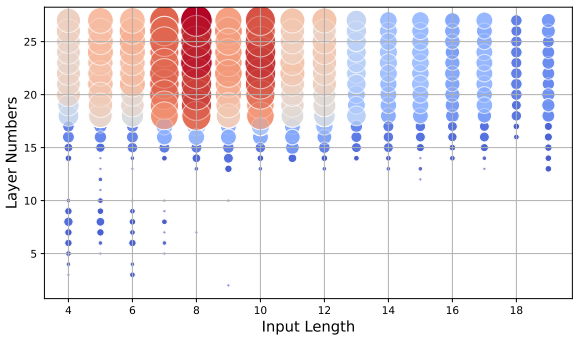
This subsection reports layer-index distributions of top-K restored states for each component. The plots complement the example-level causal traces shown in Section 4 by providing aggregated evidence for the bimodal MLP pattern and the unimodal MHA and hidden-state patterns.



(a) MLP.

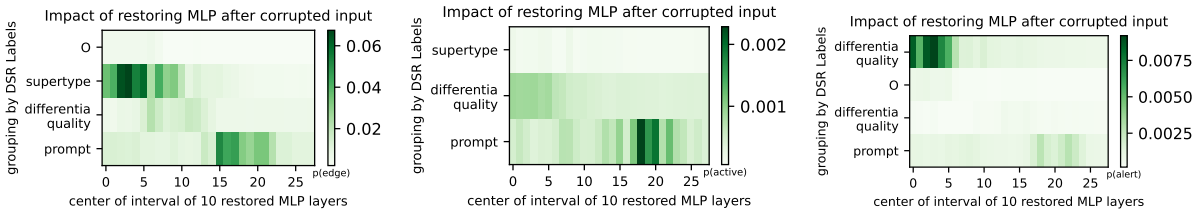


(b) MHA.



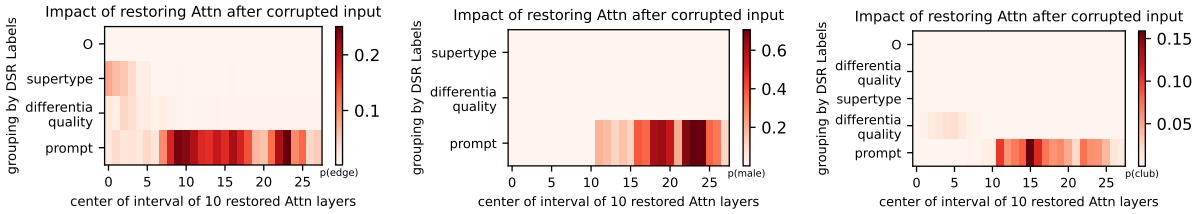
(c) Hidden states.

Figure 8: Top-10 restored states from (a) MLP, (b) MHA, and (c) hidden components plotted against input length (GPT-J-6B, EWN, $K = 10$). X-axis: input length (tokens); y-axis: layer index of top states. No systematic association between definition length and the layer distribution of restored states is observed.



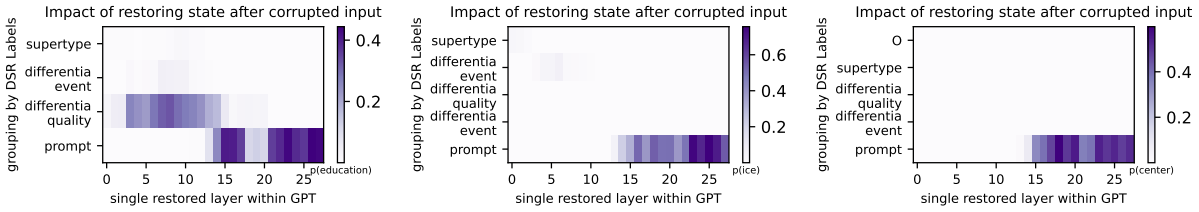
(a) (b) (c)

Figure 9: Sample of causal tracing with DSR labelling when restoring a window of 10 MLP layers. The representation highlights the distribution of important states over several layers and the importance of content words, mainly captured in supertype (a) and differentia quality (b) and (c).



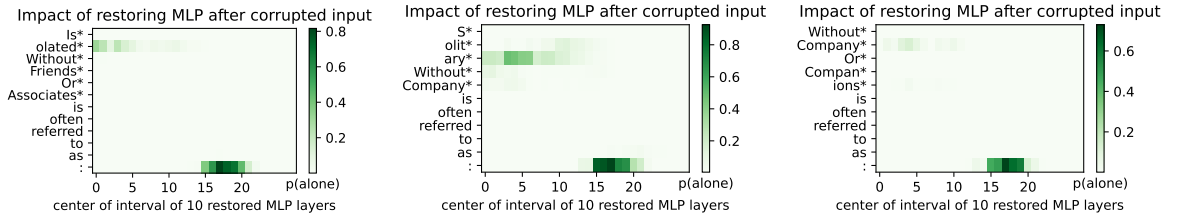
(a) (b) (c)

Figure 10: Causal traces with DSR labelling for MHA restoration (GPT-J-6B, EWN). Each heatmap shows recovery probability p_y (colour scale) when restoring a window of 10 MHA layers (x-axis: window centre) at tokens grouped by DSR label (y-axis). (a) *edge*, (b) *male*, (c) *club*.



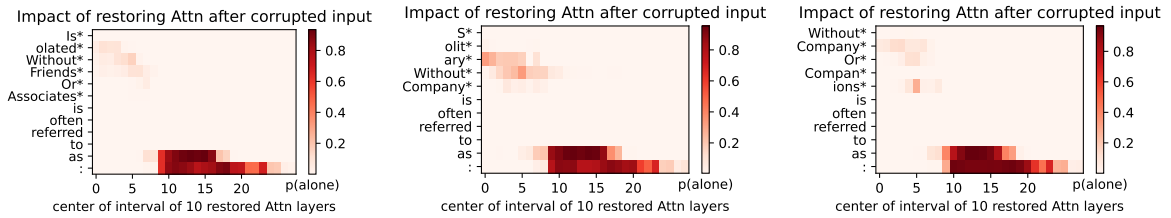
(a) (b) (c)

Figure 11: Causal traces with DSR labelling for hidden-state restoration (GPT-J-6B, EWN). Each heatmap shows recovery probability p_y (colour scale) when restoring a single hidden state (x-axis: layer index) at tokens grouped by DSR label (y-axis). (a) *education*, (b) *ice*, (c) *center*.



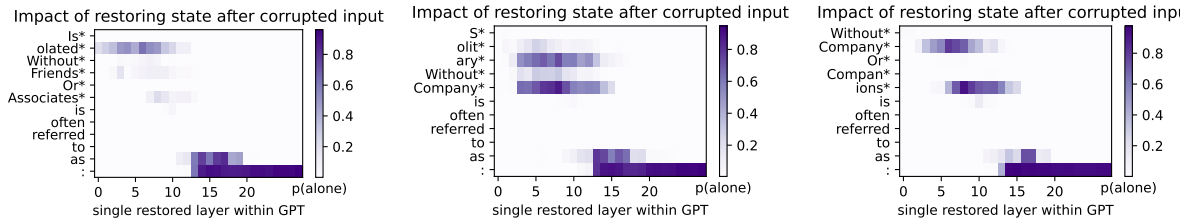
(a) (b) (c)

Figure 12: MLP traces for alternative definitions of *alone* (GPT-J-6B, EWN). Recovery probability p_y (colour scale) when restoring a window of 10 MLP layers (x-axis: window centre) at each token (y-axis). Broad localisation patterns are preserved across paraphrases; specific high-impact positions shift.



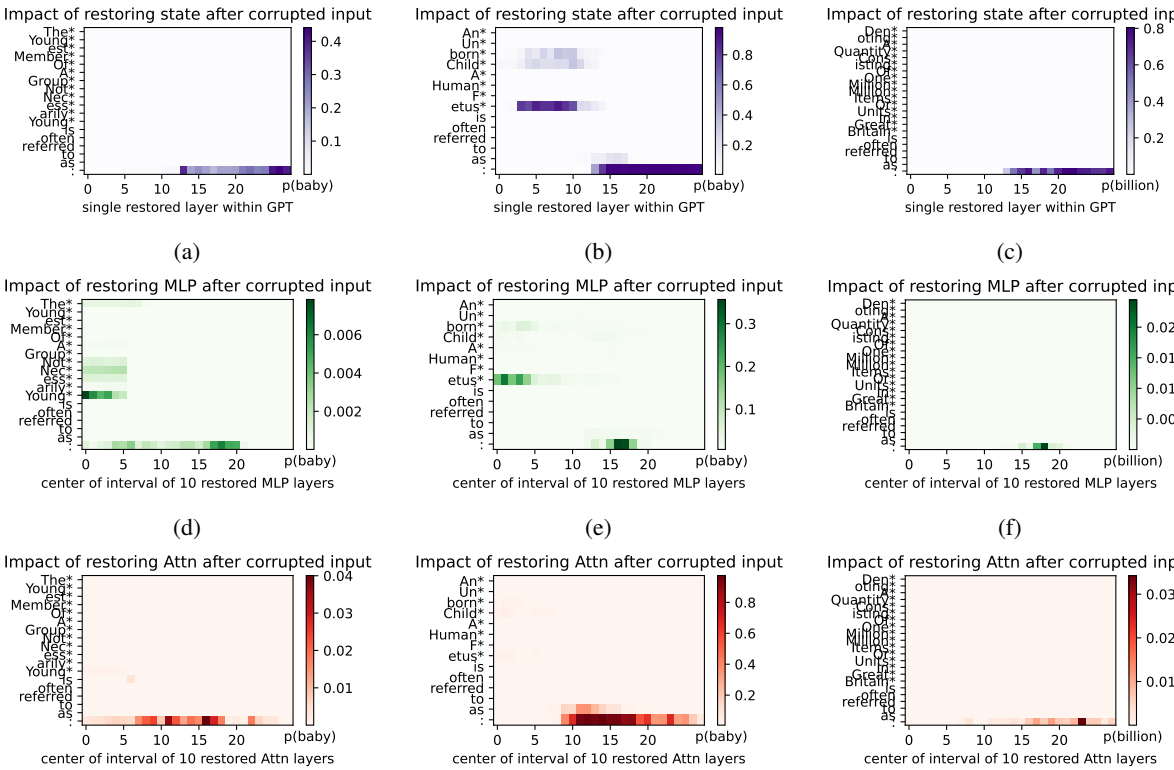
(a) (b) (c)

Figure 13: MHA traces for alternative definitions of *alone* (GPT-J-6B, EWN). Recovery probability p_y (colour scale) when restoring a window of 10 MHA layers (x-axis: window centre) at each token (y-axis). MHA patterns are more stable across paraphrases than MLP patterns.



(a) (b) (c)

Figure 14: Hidden-state traces for alternative definitions of *alone* (GPT-J-6B, EWN). Recovery probability p_y (colour scale) when restoring a single hidden state (x-axis: layer index) at each token (y-axis). Upper-layer concentration at the last token is consistent across paraphrases.

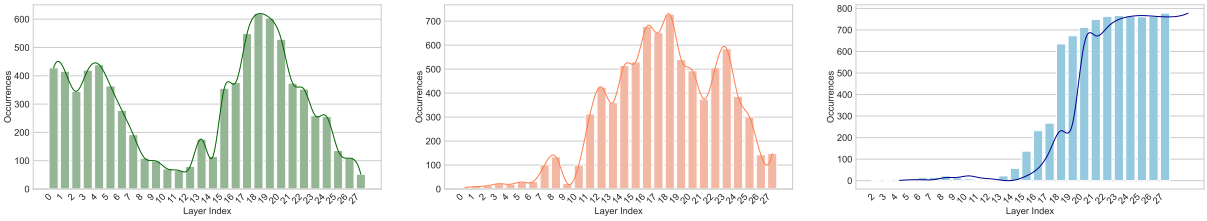


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 15: Supplementary causal traces for GPT-J-6B (EWN). Top row: hidden-state restoration (single layer); middle row: MLP restoration (window of 10 layers); bottom row: MHA restoration (window of 10 layers). Columns show different definienda: (a,d,g) *baby* (definition 1); (b,e,h) *baby* (definition 2); (c,f,i) *billion*. Colour scale: recovery probability p_y, clipped to $[0, 1]$.



(a) MLP (b) MHA (c) Hidden states

Figure 16: Distribution of layer indices among the top-10 restored states across correctly predicted GPT-J-6B EWN test samples. MLP restoration shows a bimodal pattern with clusters in early layers 0–5 and upper layers 17–21; MHA restoration concentrates in middle-to-upper layers 12–25; hidden-state restoration concentrates in upper layers 20–27.

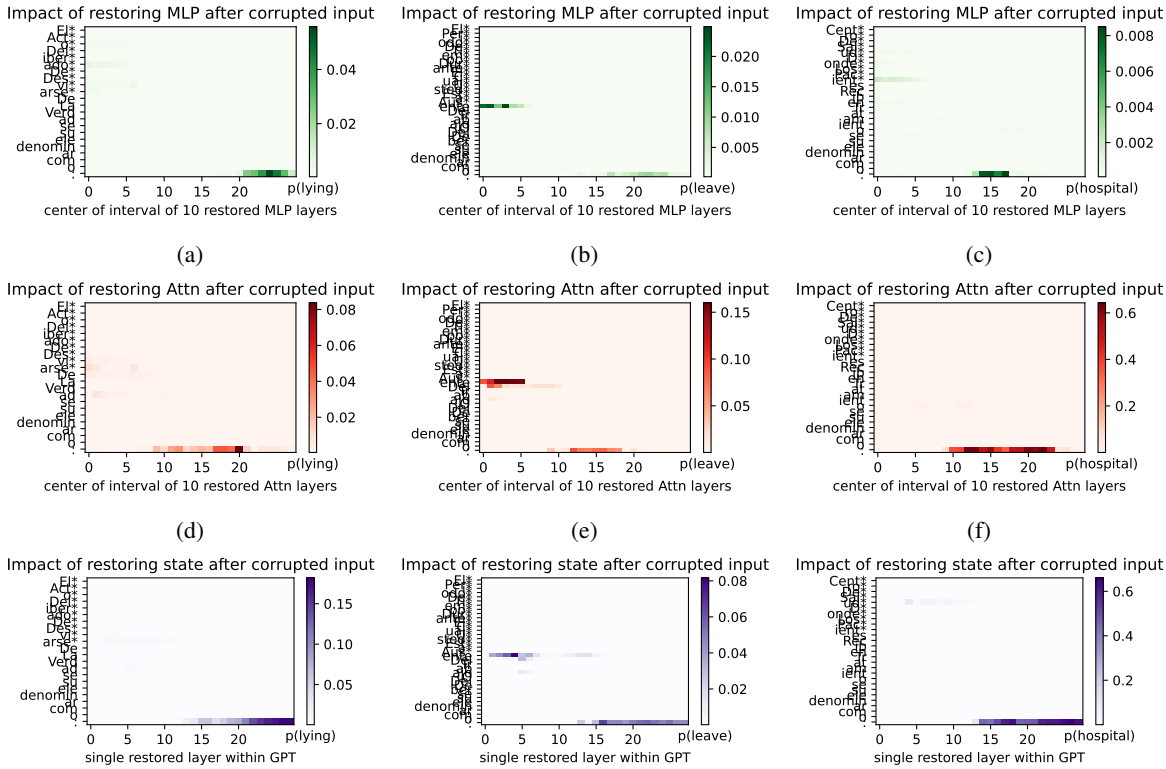


Figure 17: Cross-lingual causal traces for BERTIN GPT-J-6B (SWN). Top row: MLP restoration (window of 10 layers); middle row: MHA restoration (window of 10 layers); bottom row: hidden-state restoration (single layer). Columns show different definienda: (a,d,g) *lying*; (b,e,h) *leave*; (c,f,i) *hospital*. Colour scale: recovery probability p_y, clipped to $[0, 1]$. Y-axis: input tokens (asterisks denote corrupted definition tokens). X-axis: window centre (MLP, MHA) or layer index (hidden states).