Where Is Politeness in Japanese BERT? A Layerwise Probing and CLS Activation Patching Study

Shusuke Hashimoto, Wenchen Shi
Department of Linguistics, Indiana University Bloomington
{shuhashi, wencshi}@iu.edu

Abstract

Politeness is a key aspect of language use and pragmatics, and Japanese honorifics provide a useful testbed for analyzing how pretrained encoders represent socially meaningful distinctions. We study whether LineDistilBERT encodes Japanese honorific level in its internal representations using the KeiCO corpus, which annotates sentences with four honorific levels. To isolate pretrained representations while enabling task evaluation, we freeze the encoder and train only a lightweight classification head as a minimal readout. We then conduct layerwise linear probing by training multinomial L2-regularized logistic-regression probes on `[CLS]` representations from each layer, identifying an intermediate layer where honorific level is most linearly decodable. Finally, we test causal leverage via `[CLS]` activation patching and measure how predictions shift when donor activations are transplanted into receiver sentences. Overall, we find that honorific level is broadly decodable across layers and that `[CLS]` interventions can systematically steer the classifier's predictions, with strong depth dependence. These results show how combining linear decodability and causal interventions can yield complementary evidence about how a model encodes socially meaningful distinctions.

Keywords: BERT interpretability, linear probing, activation patching, Japanese honorifics

1. Introduction

Japanese honorifics provide a compact, well-studied testbed for pragmatic representation learning. Speakers choose honorific forms to reflect interpersonal relationships, social hierarchy, and situational formality, and these choices interact with morphosyntax (e.g., honorific verb forms) and conventionalized expressions. This makes honorific level a natural target for asking whether pretrained encoders represent pragmatically relevant distinctions beyond shallow lexical cues.

Several foundational studies have used BERT-based models to investigate pragmatic phenomena (Cho and Kim, 2024; Wise et al.). Prior work has also examined Japanese politeness from a pragmatics perspective (Hill et al., 1986; Liu and Allen, 2014; Pizziconi, 2003). In addition, BERT-based approaches have been applied to Japanese pragmatics (Chia et al., 2024). To our knowledge, however, no prior study has specifically examined how Japanese honorific-level politeness is represented in Japanese BERT-style encoders by combining layerwise probing with `[CLS]` activation patching. Accordingly, we investigate where honorific level is encoded in a Japanese BERT-style encoder, LineDistilBERT, and whether internal representations can be causally leveraged to steer predictions using the four-level KeiCO honorific corpus. To isolate what is already present in the pretrained encoder representations, we freeze all LineDistilBERT encoder parameters and train only a lightweight classification head on top of `[CLS]` as a minimal supervised readout for probing and patching.

Our analysis combines layerwise linear probing and targeted causal intervention. We first train multinomial L2-regularized logistic-regression probes on `[CLS]` representations from each layer to quantify how linearly decodable honorific level is across depth and to identify a dev-selected best layer. We then perform `[CLS]` activation patching at selected layers by overwriting a receiver sentence's `[CLS]` activation with a donor activation and measuring the resulting prediction shifts, together with standard controls.

This paper addresses three research questions. **RQ1** asks to what extent pretrained LineDistilBERT encodes honorific-level distinctions in its internal representations. **RQ2** asks which layer provides the strongest linear decodability of honorific level and how sharply this information is localized around that layer. **RQ3** asks whether the probe-selected best layer plays a causal role in honorific prediction and whether targeted activation patching at that layer can reliably steer honorific predictions.

In summary, we contribute (i) a layerwise decodability profile for Japanese honorific level in a pretrained encoder, (ii) a best-layer evaluation against a frozen-encoder head-only baseline, and (iii) a causal patching analysis with self-, random-, and wrong-layer controls that probes depth-dependent sensitivity.

2. Related Work

We build on two interpretability paradigms for transformer encoders. Probing studies train lightweight

Proceedings of the 1st SoCon Workshop & the 2nd NLPSI Workshop @ LREC 2026, pages 67–75
12 May 2026.

classifiers on frozen representations to quantify linear decodability and its variation across layers (Alain and Bengio, 2018; Belinkov, 2022), and layerwise probe curves have been used to track how linguistic properties become accessible with depth in BERT-style models (Jawahar et al., 2019; Rogers et al., 2020). Activation patching provides a causal counterpart by overwriting internal activations at a chosen site and measuring output changes (Heimersheim and Nanda, 2024; Dumas et al., 2025), with recent work emphasizing careful controls and interpretation (Zhang and Nanda, 2024).

3. Dataset

We use KeiCO corpus (Liu and Kobayashi, 2022), which annotates each sentence with one of four honorific (politeness) levels defined under a Systemic Functional Linguistics (SFL) framework (Liu and Kobayashi, 2022). Detailed information about the dataset is provided in (Appendix 1). We follow the dataset's label descriptions, ranging from highly formal honorific constructions (Levels 1–2) to polite language with limited honorific morphology (Level 3) and informal speech with no honorifics (Level 4), and we split the data into stratified 80%/10%/10% train/dev/test partitions.

Our model is LineDistilBERT with a frozen encoder and a lightweight sequence-classification head trained as a minimal readout. For probing, we select the best layer by dev macro-F1 and report final results on the test set.

4. Methods

We analyze pretrained LineDistilBERT with a frozen encoder and a lightweight classification head trained as a minimal readout (Koga et al., 2023). We chose LineDistilBERT because its simplified layer structure makes it easier to interpret in layerwise probing. Our methods combine layerwise linear probing to locate decodable honorific information and `[CLS]` activation patching to test causal steering. We treat the head-only classifier as the fixed readout function used to compute both baseline and patched predictions, so interventions are evaluated without updating encoder parameters.

4.1. Layerwise probing

For each input sentence, we run the frozen encoder and extract the `[CLS]` vector from each hidden-state index as a layer-specific sentence representation. Here, hidden-state index 0 corresponds to the embedding output (pre-contextualization), and indices 1–6 correspond to the outputs of successive transformer layers. We adopt the working hypothesis that honorific-related information is substantially aggregated into `[CLS]` and is therefore accessible via a single sequence-level embedding, while noting that relevant cues may also be token-local (e.g., honorific-bearing verbs or function morphemes), which we leave to future work.

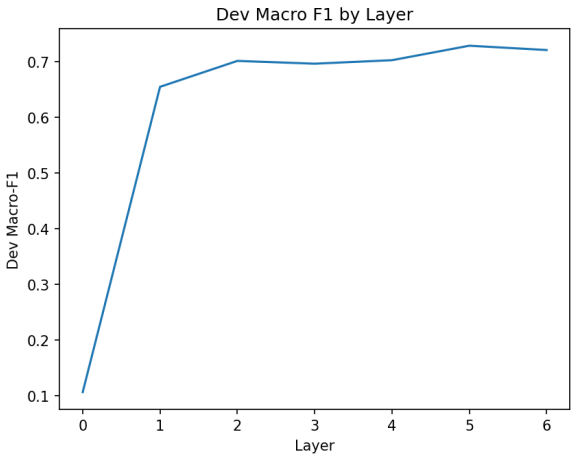

Figure 1: Dev macro-F1 of layerwise `[CLS]` probes in LineDistilBERT.

For each layer l, we train a multinomial L2-regularized logistic-regression probe on $X^{(l)}_{\text{train}} \in R^{N\times H}$ and evaluate on $X^{(l)}_{\text{dev}}$. We select the best layer by dev macro-F1,

$$best\_layer = \arg\max_{l} \mathrm{MacroF1}_{dev}(l),$$

and report a layerwise decodability curve together with the dev-selected $best\_layer$.

4.2. Activation patching

To test causal leverage, we perform `[CLS]` activation patching at selected layers. Let $\mathrm{CLS}^{(l)}_{x} \in R^{H}$ denote the `[CLS]` hidden state at layer l. For a donor input x_d and receiver input x_r, we replace

$$\mathrm{CLS}^{(l)}_{x_r} \leftarrow \mathrm{CLS}^{(l)}_{x_d}$$

during the receiver forward pass and then measure changes in logits and predictions under the same readout. Donors are drawn from Level 1 and receivers from Level 4. Operationally, we implement patching by overwriting the CLS vector in the layer output tensor during the receiver pass. We summarize effects using (i) the transition matrix from baseline to patched predictions, (ii) Δ target-class logit, and (iii) the flip-to-target rate.

We include three standard controls. Self-patch uses $x_d = x_r$ and should yield near-identity behavior. Random-donor patch permutes donor `[CLS]` vectors within a batch to test whether effects depend on specific donor–receiver pairings. Wrong-

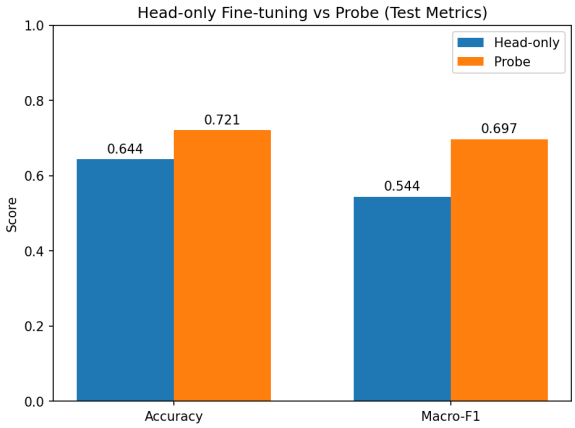

Figure 2: Head-only baseline vs. linear probe performance on the test set

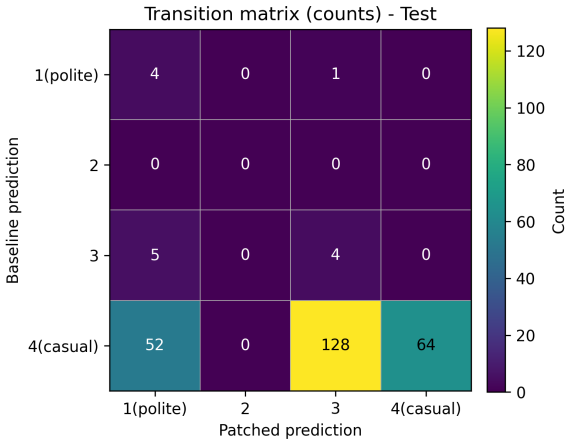

Figure 3: Test-set transition matrix (counts) for CLS activation patching at Layer 5.

layer patch sweeps non-target layers to assess depth localization of causal sensitivity.

5. Results

5.1. Decodability and best layer

Figure 1 reports dev macro-F1 scores of multinomial logistic-regression probes trained on `[CLS]` representations extracted from each hidden-state index. Scores rise sharply from Layer 0 to Layer 1 and remain consistently high across Layers 1–6 within a narrow range. This addresses **RQ1** and suggests that honorific-level distinctions are already encoded in pretrained LineDistilBERT representations in a form that is broadly linearly decodable across the encoder stack, rather than emerging only in a single late layer. The best-performing layer on the dev set is Layer 5 (macro-F1 ≈ 0.73), which we designate as the *best layer*. This addresses **RQ2** and indicates that the peak is shallow rather than sharply localized, since neighboring layers, especially Layers 4–6, achieve nearly identical dev macro-F1.

5.2. Best-layer probe vs. head-only baseline

After selecting the best layer on the development set (Layer 5), we trained a multinomial L2-regularized logistic-regression probe on the training `[CLS]` features from that layer and evaluated it once on the test set. Figure 2 compares test performance between (i) the head-only baseline, where the pretrained encoder is frozen and only the task head is trained, and (ii) the linear probe, which reads out honorific labels from fixed Layer 5 representations. The probe outperforms the head-only baseline on the test set. Together with the shallow best-layer peak in Figure 1, this further supports **RQ2** by showing that the dev-selected best layer yields a strong readout on test data, while nearby layers remain competitive.

5.3. Causal steering at the best layer

To test whether the probed representation is not only linearly decodable but also causally influential for predictions, we intervene on the forward pass via `[CLS]` activation patching at the selected layer (hidden-state index 5). We overwrite a receiver sentence's `[CLS]` vector with a donor `[CLS]` vector (donors drawn from Level 1; receivers drawn from Level 4) and re-run the head-only classifier. Figure 3 reports the resulting transition matrix on the test set. The dominant effect is a marked reduction in Level 4 predictions after patching. Among baseline Level 4 predictions ($n = 244$), 64 remain Level 4 under patching, while 180 transition to other labels (52 to Level 1 and 128 to Level 3). Thus, patching at this layer changes the predicted label for approximately $180/244 \approx 74\%$ of baseline-4 cases. These systematic shifts directly speak to **RQ3** by showing that targeted patching at the probe-selected best layer can steer honorific predictions for a large fraction of cases.

5.4. Controls and depth dependence

We next evaluate standard controls to clarify how to interpret the steering effect. Self-patch yields near-identity behavior with no off-diagonal movement (Appendix 2), showing that patching itself does not alter the predicted label. Random patch, which permutes donors within a batch, produces a transition pattern highly similar to aligned patching, suggesting that the main effect is not driven by specific donor–receiver pairings (Appendix 2).

Finally, wrong-layer sweeps show strong depth dependence. Early-layer patching yields little to no change, whereas later-layer patching increasingly steers predictions, with the strongest shifts at the final layer (Layer 6) where the classification head consumes `[CLS]` most directly (see Appendix3, Figure10). This pattern suggests that patching effects are amplified by proximity to the classifier input and do not isolate a single honorific-specific causal layer.

6. Discussion

Our results show a nuanced view of where honorific information is accessible and where it influences prediction in a frozen-encoder classifier. First, `[CLS]` transplantation is a broad intervention. The close similarity between aligned-donor and random-donor patching suggests that overwriting `[CLS]` can induce large prediction shifts even when donor identity is randomized. This pattern is consistent with at least two hypotheses. One is that many donor `[CLS]` activations share a common component that tends to push predictions away from the most casual region, making the effect relatively robust to which donor is used. Another is that overwriting `[CLS]` at this layer acts as a generic but structured perturbation that biases decisions toward non-casual labels, largely independent of donor label.

These hypotheses make different predictions that can be tested without changing the overall intervention framework. If a shared "politeness" component dominates, then the effect should vary systematically with donor label or donor strength (e.g., donors from more polite levels, or donors that the baseline classifier assigns high confidence, should induce larger shifts). If the effect is largely a generic perturbation, then shifts should be comparatively insensitive to donor label and instead correlate more with receiver-side factors such as baseline confidence or juxtaposition to the decision boundary (e.g., low-margin cases should flip more easily). In both cases, label-conditioned randomization provides a direct diagnostic by stratifying donors and receivers and comparing effect sizes across strata.

Two broader implications follow. The high and relatively flat decodability curve across Layers 1–6 suggests that honorific cues are distributed across depth rather than isolated to a single late layer. At the same time, the wrong-layer sweep indicates that causal steering is strongest near the final layer. This mismatch is interpretable because probing and patching capture different notions. Layerwise probing identifies where honorific level is most *linearly decodable* under an external diagnostic classifier, whereas patching measures where overwriting the representation most strongly controls the model's own downstream decision. These need not coincide. In an encoder–classification architecture, the classification head directly consumes the final-layer `[CLS]` representation, so patching at Layer 6 overwrites the classifier's immediate input and can therefore exert the strongest causal control even if linear decodability peaks slightly earlier.

Taken together, these results suggest that pretrained LineDistilBERT carries a strong and linearly accessible signal for honorific level across most of the encoder stack, and that a minimal readout can be systematically steered by intervening on internal `[CLS]` activations. At the same time, the similarity between aligned and randomized patching highlights that full-vector `[CLS]` transplantation is not cleanly honorific-specific, and that causal steering in this setup may mix honorific-related content with other factors aggregated in `[CLS]`.

7. Conclusion

We studied where honorific-level distinctions are represented in pretrained LineDistilBERT and whether those representations can steer a frozen-encoder classifier. For **RQ1**, layerwise probing showed that honorific level is broadly linearly decodable across most encoder layers. For **RQ2**, dev selection identified an intermediate best layer (Layer 5), but the peak was shallow, with neighboring layers performing similarly, suggesting limited localization. For **RQ3**, `[CLS]` activation patching at the best layer induced systematic shifts away from the most casual class, demonstrating causal leverage of the probed representation under our readout. Controls clarify the interpretation. Self-patch confirms that hooking does not induce drift, random patch indicates the effect is not driven by donor–receiver pairing, and wrong-layer sweeps show strong depth dependence with the largest shifts near the final layer. These results suggest that honorific-related information is present in hidden states and that `[CLS]` activations can serve as an effective control point for steering honorific predictions, motivating token-localized sites and label-conditioned causal tests.

By grounding interpretability analysis in a well-studied sociolinguistic system, we illustrate how combining linear decodability and causal interventions yields complementary evidence about what it means for a model to "encode" social meaning. In particular, the dissociation between the layer of maximal linear decodability and the layer of maximal causal influence suggests that representational accessibility and decision control can occupy different loci within the model stack. This distinction matters for interpreting probing results and for designing interventions that target socially meaningful information.

8. Limitation

(1) The nature of Japanese politeness While Japanese politeness is often overtly marked in the surface string, it is also shaped by social relations and situational norms. As a result, some cases remain underspecified without context, and the same surface form can support multiple pragmatic readings. Because our modeling setting provides only sentence-level text as input, sentence-level classification may consequently over-rely on token-level markers that correlate with politeness in the corpus, rather than the underlying situational variables. A stronger test would use context-rich instances (multiple sentences per example) and/or controlled subsets where overt honorific markers are minimized, forcing models to exploit discourse and situational cues rather than recognizing surface forms.

(2) Dataset context and intervention interpretability Our intervention analyses inherit limitations from the dataset and input representation. KeiCO provides topical field labels, which can be useful for controlling topic confounds, but these tags do not explicitly encode interactional context such as speaker–addressee relations, social hierarchy, or situational formality. This matters for `[CLS]` patching. Because `[CLS]` aggregates many factors beyond honorific level (topic, semantics, discourse structure), transplanting an entire `[CLS]` vector from an unrelated sentence can introduce broad distributional changes that are not specific to politeness. The observation that random-donor patching yields effects comparable to aligned-donor patching is consistent with this concern, suggesting that part of the steering may reflect generic perturbation rather than label-specific transfer.

A stronger approach would therefore increase contextual control at the instance level and narrow the intervention target. For example, one can construct minimal pairs that differ only in situational metadata (e.g., explicit formality/relationship tags) and patch across these controlled contexts, or restrict donor/receiver sampling within the same topical field to reduce topic-driven shifts. Another direction is to move from full-vector transplantation toward more targeted interventions such as editing representations along a learned politeness direction, which could better isolate the causal contribution of honorific-related features.

(3) High performance does not necessarily imply understanding Our analyses are motivated by the hypothesis that politeness-related features are important for solving honorific-level classification. However, strong accuracy or decodability does not by itself establish that such features are *necessary* for task success. A direct next step is an ablation-based causality test, which removes or suppresses candidate politeness-related components and measures the performance change relative to the intact model. This follows a common interpretability logic that quantifies component importance by comparing a full model to an ablated model and inspecting the resulting performance difference (Li and Janson, 2024). Related causal-concept work similarly uses concept ablation (with random-concept controls) to test whether a hypothesized concept actually plays a causal role in predictions (Singla et al., 2021). If performance remains largely unchanged after ablating the "purported" politeness mechanism, then the model may be exploiting alternative correlates in the dataset, and understanding politeness (as operationalized by our identified features) may not be required for this benchmark.

(4) Heterogeneity in patchability across instances Our patching analyses aggregate effects across many donor–receiver pairs, implicitly treating instances as equally patchable. In practice, patching sensitivity may vary. Some donor `[CLS]` vectors may more easily steer receiver predictions than others, and some receiver sentences may be relatively resistant to donor signals. Future work should characterize this heterogeneity explicitly by, for example, stratifying patch effects by baseline confidence (e.g., max softmax probability / logit margin), sentence length, presence of overt honorific markers, or semantic/domain similarity between donor and receiver, and by reporting instance-level distributions rather than only aggregate transition counts.

9. Acknowledgments

We are grateful to Luke Gessler and Phakphum Artkaew for their valuable feedback and support throughout this work.

10. References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Zheng Lin Chia, Michal Ptaszynski, Marzena Karpinska, Juuso Eronen, and Fumito Masui. 2024. Initial exploration into sarcasm and irony through machine translation. *Natural Language Processing Journal*, 9:100106.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20, Bangkok, Thailand. Association for Computational Linguistics.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching.

Beverly Hill, Sachiko Ide, Shoko Ikuta, Akiko Kawasaki, and Tsunao Ogino. 1986. Universals of linguistic politeness: Quantitative evidence from japanese and american english. *Journal of Pragmatics*, 10(3):347–371.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Kobayashi Koga, Shengzhe Li, Akifumi Nakamachi, and Toshinori Sato. 2023. Line distilbert japanese. GitHub repository.

Maximilian Li and Lucas Janson. 2024. Optimal ablation for interpretability.

Muxuan Liu and Ichiro Kobayashi. 2022. Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

Xiangdong Liu and Todd James Allen. 2014. A study of linguistic politeness in japanese. *Open Journal of Modern Linguistics*, 4(5):651–663.

Barbara Pizziconi. 2003. Re-examining politeness, face and the japanese language. *Journal of Pragmatics*, 35:1471–1506.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. 2021. Using causal analysis for conceptual deep learning explanation.

Matt Wise, Houda Nait, El Barj, and Anna Goldie. No. Pragmabert: Analyzing pragmatic markers in political speech.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods.

Appendix

1. Dataset Overview

Table 1: Dataset statistics (from Liu & Kobayashi, 2022; Table 3).

Polite	Sent.	Avg.len	Wordtokens	Wordtypes	Sentence Example
Level 1	2,584	18.2	47,111	4,744	担当の者をお呼びしました。 *"I called the person in charge."*
Level 2	2,046	16.4	33,476	3,897	一度、ゆっくりお礼にあがります。 *"I'll come by sometime and thank you properly."*
Level 3	2,694	15.2	40,980	4,448	あの人のどこが嫌いなんですか? *"What do you dislike about that person?"*
Level 4	2,683	13.5	36,233	4,315	これ、うちのオススメ。 *"This is our recommendation."*
Total	**10007**	**15.8**	**157806**	**6465**	

2. Self- and Random-Patch Transition Matrices

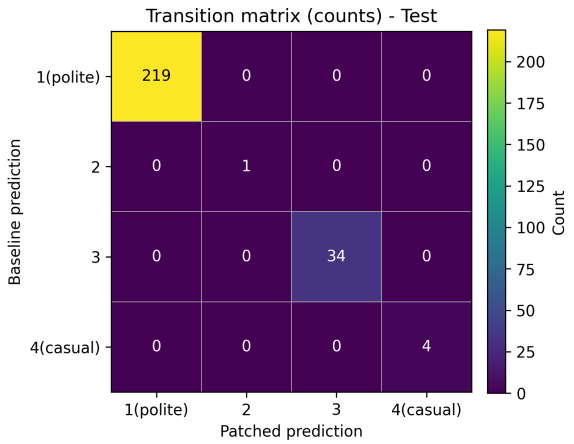

Figure 4: Self-patch transition matrix (counts) for CLS activation patching at Layer 5.

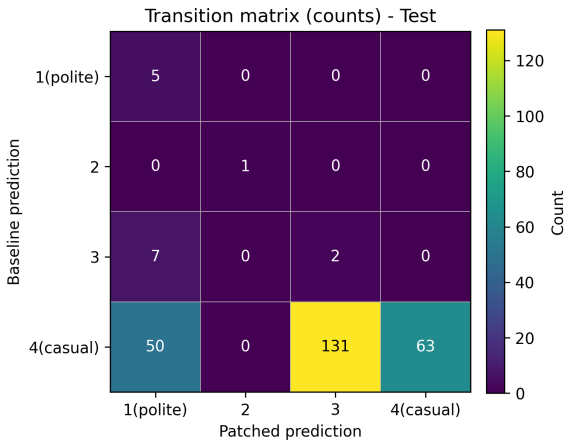

Figure 5: Random-patch transition matrix (counts) for CLS activation patching at Layer 5.

3. Wrong-Layer Patch Transition Matrices

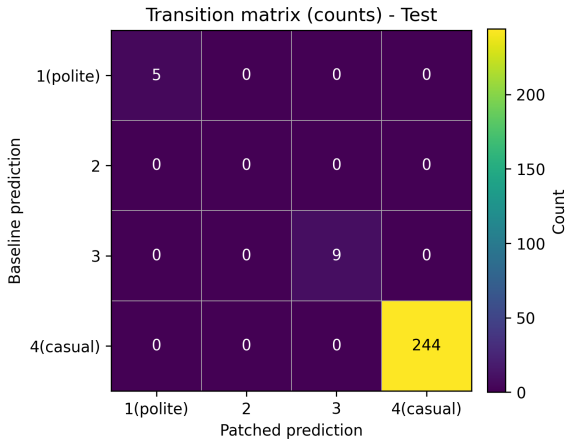

Figure 6: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 1.

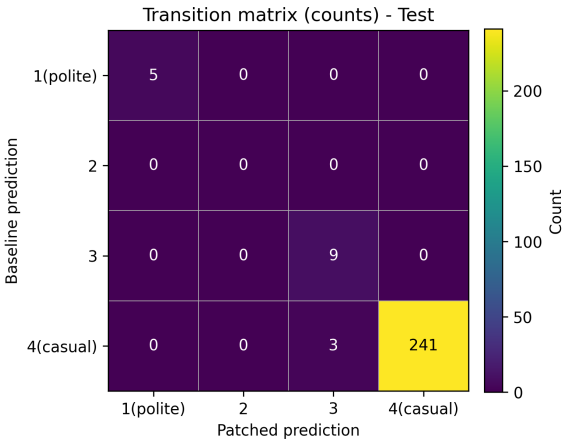

Figure 7: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 2.

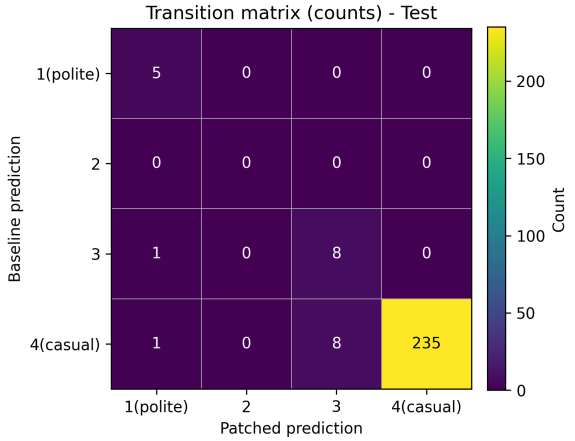

Figure 8: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 3.

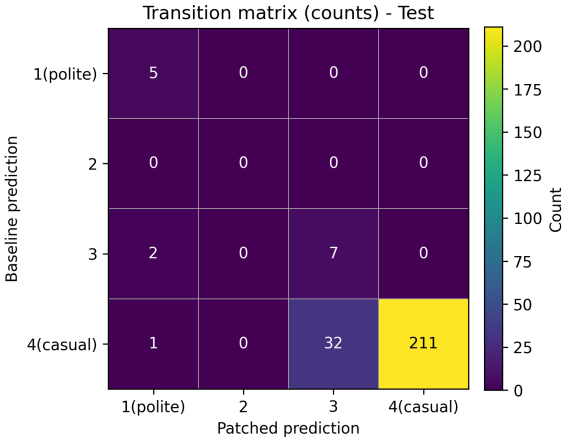

Figure 9: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 4.

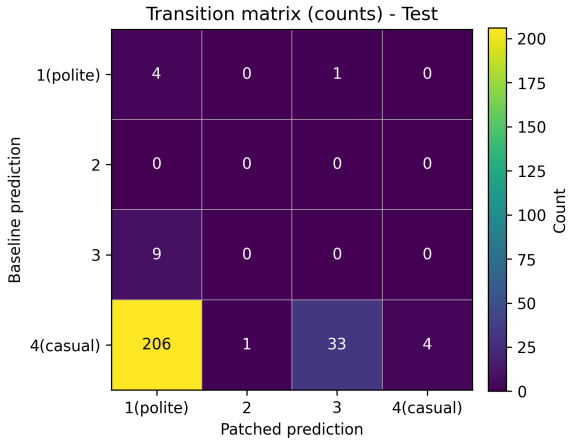

Figure 10: Wrong-layer patch transition matrix (counts) for CLS activation patching at Layer 6.

4. Hyperparameters for Baseline Training (frozen-encoder readout)

Component	Details
Optimizer	AdamW
Learning Rate	2.0×10^{-5}
Weight Decay	0.01
Scheduler	Linear
Batch Size	16
Epochs	5
Tokenizer	`line-corporation/line-distilbert-base-japanese`
Padding	95th percentile of sentence lengths

5. Test Metrics

Setting	Test metrics
Best-layer patch (hs index = 5)	$N = 258$; Δ target logit $= 0.872$; flip-to-target $= 0.221$
Control: Self-patch	$N = 258$; Δ target logit $= 0.000$; flip-to-target $= 0.000$
Control: Random-patch (permute within batch)	$N = 258$; Δ target logit $= 0.873$; flip-to-target $= 0.221$
Control: Wrong-layer patch (hs index = 1; embedding output)	$N = 258$; Δ target logit $= -0.000$; flip-to-target $= 0.000$
Control: Wrong-layer patch (hs index = 2)	$N = 258$; Δ target logit $= 0.065$; flip-to-target $= 0.000$
Control: Wrong-layer patch (hs index = 3)	$N = 258$; Δ target logit $= 0.135$; flip-to-target $= 0.008$
Control: Wrong-layer patch (hs index = 4)	$N = 258$; Δ target logit $= 0.375$; flip-to-target $= 0.012$
Control: Wrong-layer patch (hs index = 6; final layer)	$N = 258$; Δ target logit $= 1.694$; flip-to-target $= 0.833$