

Geometric Separation Without Functional Specialization: Auditing Subspace Disentanglement in Dual-Encoder LLM Paraphrase Detection

Govind Arun

University of Maryland, College Park
College Park, MD, USA

Ashish Anton Abraham*

Indira Gandhi National Open University
New Delhi, India

Gadha Lekshmi P

Rutgers University, New Brunswick
New Brunswick, NJ, USA

Abstract

Explicit subspace disentanglement, enforcing orthogonality and simplex constraints on learned projections, is an increasingly common recipe for paraphrase models claimed to separate lexical, syntactic, and semantic information. A dual-encoder hierarchical model trained with these constraints on the DiPPER corpus of LLM-generated paraphrases produces near-perfect geometric separation, a pairwise Centered Kernel Alignment (CKA) of roughly 0.0007, yet that separation is not evidence of functional linguistic specialization. Three results establish this. First, near-zero CKA is largely preordained by an explicit orthogonality penalty, which converges to 4×10^{-6} . Second, each of the three subspaces, including the nominally syntactic one for which we identify no functional probe, classifies paraphrase in isolation at F_1 above 0.986, so the subspaces are not functionally distinct. Third, removing disentanglement at matched capacity costs only 0.013 F_1 , leaving most of the reported gain attributable to dual-encoder capacity rather than disentanglement. Capacity-matched ablations and controlled functional probes, not CKA alone, are required to substantiate disentanglement claims; orthogonality penalties otherwise manufacture the very metric used to certify them.

1 Introduction

A recurring design pattern in representation learning for paraphrase and semantic-matching tasks splits an encoder output into several projected subspaces and pushes those subspaces apart with explicit penalties, typically a soft orthogonality term on the projection matrices and a simplex-style term that asks the projections to jointly preserve the original representation. The subspaces are then labeled lexical, syntactic, and semantic, and low pairwise representational similarity is offered as

evidence of linguistically specialized features. Reporting a near-zero similarity metric, most often CKA (Kornblith et al., 2019), as confirmation that disentanglement worked has become routine.

The pattern rests on real evidence: probing studies show that transformer encoders distribute lexical, syntactic, and semantic information across layers and subspaces rather than token positions (Hewitt and Manning, 2019; Tenney et al., 2019), and several paraphrase systems decompose representations along these lines or combine encoders for complementary signal (Xu et al., 2020; Liu et al., 2018; Ortiz-Barajas et al., 2022; Peng et al., 2022; Yang et al., 2016). Driving inter-subspace similarity to near zero therefore appears to operationalize that picture.

This appearance does not survive scrutiny. The instance audited here is a dual-encoder hierarchical transformer (DeBERTa-v3 and RoBERTa) that projects each encoder output into three subspaces, applies orthogonality and simplex penalties, attaches level-specific auxiliary supervision under standard multi-task weighting (Dan et al., 2022; He et al., 2022), and fuses the levels through a gated bilinear meta-classifier. Trained on the DiPPER corpus (Krishna et al., 2023), it reaches $F_1 = 0.9846$ with almost perfectly separated subspaces under CKA. Three results recomputed on the trained model overturn that reading: near-zero CKA is mechanically induced by the orthogonality penalty; every subspace classifies the task in isolation at F_1 above 0.986; and removing disentanglement at matched capacity costs only 0.013 F_1 , leaving the gain dominated by dual-encoder capacity, that is, a trained ensemble. CKA-as-disentanglement-evidence can therefore mislead, and concrete safeguards follow.¹

* Corresponding author: ashishabraham22@gmail.com

¹Code: <https://github.com/GIND123/Hierarchical-Transformer>.

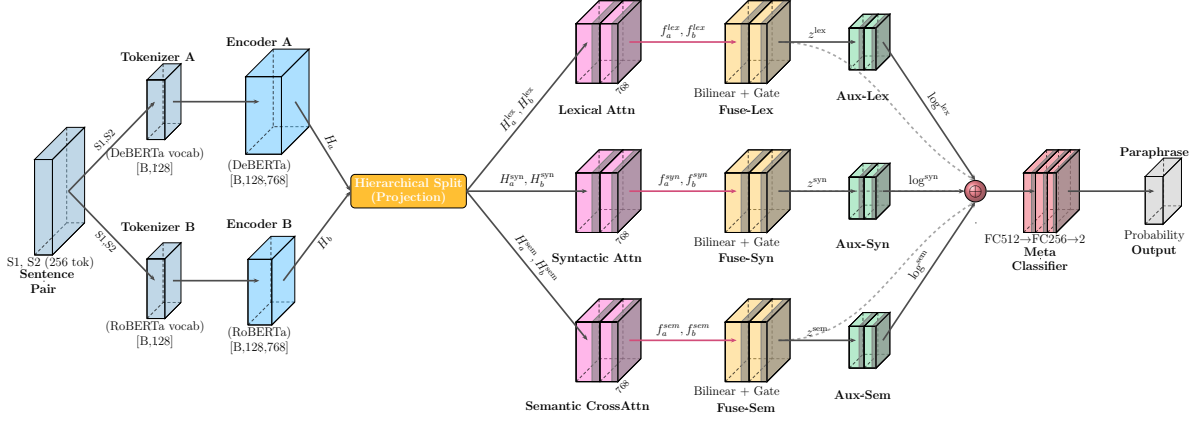


Figure 1: Hierarchical dual-encoder architecture under test: two encoders, level-wise projection into lexical, syntactic, and semantic subspaces, per-level attention and gated bilinear fusion, and a meta-classifier over the fused vectors and auxiliary logits.

2 Method under test

The audited model appears in Figure 1. Two pretrained encoders (DeBERTa-v3-base and RoBERTa-base) produce contextual states $H_a, H_b \in \mathbb{R}^{L \times d}$ for an ordered text pair. Three shared projection matrices $S_\ell \in \mathbb{R}^{d_\ell \times d}$ map each encoder output into level-specific subspaces $\ell \in \{\text{lex}, \text{syn}, \text{sem}\}$:

$$H_a^\ell = H_a S_\ell^\top, \quad H_b^\ell = H_b S_\ell^\top. \quad (1)$$

A soft orthogonality penalty $\mathcal{L}_{\text{ortho}}$ minimizes normalized Frobenius inner products $\langle S_\ell, S_{\ell'} \rangle_F$ for $\ell \neq \ell'$, and a simplex penalty $\mathcal{L}_{\text{simp}}$ constrains $\sum_\ell S_\ell S_\ell^\top \approx I_d$ (Figure 2). Additive attention pools each stream independently:

$$e_i^{(X)} = w_X^\top \tanh(W_X h_i^{(X)}), \quad (2)$$

$$\alpha_i^{(X)} = \frac{\exp(e_i^{(X)})}{\sum_{j=1}^L \exp(e_j^{(X)})}, \quad f_a^X = \sum_{i=1}^L \alpha_i^{(X)} h_i^{(X,a)}. \quad (3)$$

The semantic level adds bidirectional multi-head attention (MHA) with layer normalization (LN), $\tilde{H}_a^{\text{sem}} = \text{LN}(H_a^{\text{sem}} + \text{MHA}(H_a^{\text{sem}}, H_b^{\text{sem}}, H_b^{\text{sem}}))$ and symmetrically for b . Per level, encoder B is aligned to A and the streams are combined by a gated bilinear fusion with a residual gate:

$$b^\ell = \text{Bilinear}(f_a^\ell, W_{\text{align}} f_b^\ell), \quad (4)$$

$$g^\ell = \sigma(W_g[f_a^\ell, W_{\text{align}} f_b^\ell]), \quad (5)$$

$$z^\ell = W_p(g^\ell \odot b^\ell + (1 - g^\ell) \odot f_a^\ell). \quad (6)$$

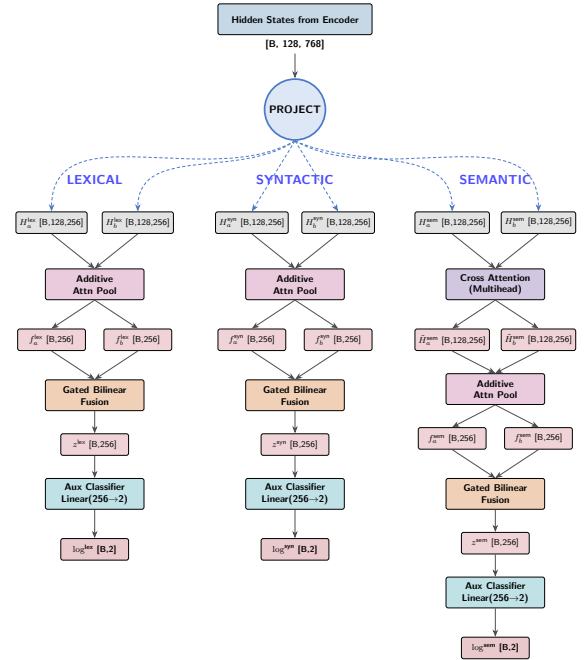


Figure 2: Level-wise splitting of features and the cross-attention applied at the semantic level.

Each level emits an auxiliary logit u^ℓ through a lightweight classifier, and a multi-layer perceptron (MLP) meta-classifier (Figure 3) consumes the three fused vectors and three auxiliary logits, $\hat{y} = \text{MLP}([z^{\text{lex}}; z^{\text{syn}}; z^{\text{sem}}; u^{\text{lex}}; u^{\text{syn}}; u^{\text{sem}}])$. Training minimizes a composite loss

$$\mathcal{L} = \alpha \mathcal{L}_{\text{main}} + (1 - \alpha) \frac{\mathcal{L}_{\text{lex}} + \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{sem}}}{3} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}} + \lambda_{\text{simp}} \mathcal{L}_{\text{simp}}, \quad (7)$$

with $\alpha = 0.6$, $\lambda_{\text{ortho}} = 0.02$, $\lambda_{\text{simp}} = 0.01$. The level-wise design is motivated by evidence of over-

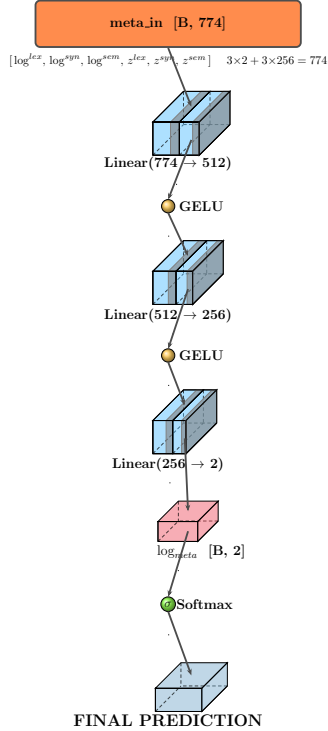


Figure 3: Meta-classifier over the concatenated fused vectors and auxiliary logits.

reliance on superficial lexical cues (Macbeth et al., 2020). What is at issue is not the architecture but the claim attached to it: that the penalties yield interpretable, linguistically specialized subspaces.

3 Setup and probing protocol

Training and evaluation use the DiPPER corpus (Krishna et al., 2023): 1,150,000 training pairs, 50,000 validation pairs, and the official test split of 324,633 pairs, with the decision threshold tuned on validation to $\tau^* = 0.55$. The probing protocol, under-reported in the original study, is specified as follows. **Pooling**: each subspace projection is pooled independently for each sentence using additive attention pooling. Both sentence-level pooled vectors are combined via a gated bilinear fusion to yield the pair representation used for probing. **CKA**: linear CKA (Kornblith et al., 2019) between the pooled projected representations of the three subspaces over the validation set, the metric the orthogonality penalty most directly influences. **Linear probing**: for each subspace independently, the pooled pair representation is used to fit an L2-regularized logistic regression on the binary paraphrase label with an L2 penalty strength of 1.0, and validation F_1 reported; a subspace that classifies

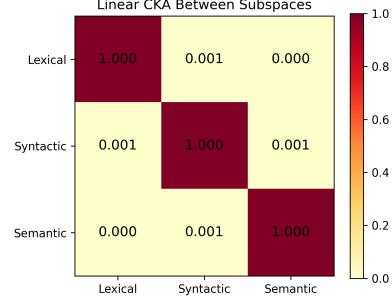


Figure 4: CKA between the learned subspaces. The near-zero off-diagonal values are driven by the orthogonality penalty ($\mathcal{L}_{\text{ortho}} = 4 \times 10^{-6}$) and restate constraint satisfaction rather than independent evidence of linguistic specialization.

well in isolation solves the task on its own. No gradients propagate into the encoder or projection layers during probing. **Probe data and splits**: each probe is trained on the first 20,000 validation pairs, split into 16,000 training and 4,000 held-out examples; this split is drawn from the head of the validation file rather than an independently reshuffled sample, so the class balance of the probing subset is not separately verified against the overall validation positive rate. **Functional grounding**: Spearman correlation between the L2 norm of a subspace representation and surface statistics, namely token Jaccard overlap and length ratio for the lexical claim and surface lexical overlap for the semantic claim. No surface-free syntactic probe was identified here, leaving standard structural probes to future work; this absence is treated below as part of the result.

4 Findings

4.1 Geometric separation is not functional disentanglement

The subspaces are almost perfectly separated: CKA is 0.0008 (lexical, syntactic), 0.0005 (lexical, semantic), and 0.0007 (syntactic, semantic), with $\mathcal{L}_{\text{ortho}}$ converging to 4×10^{-6} (Figure 4). The original study read these values as confirmation that the constraints disentangle the linguistic representation spaces. That inference is close to circular. The orthogonality penalty minimizes inter-projection alignment, which in turn drives pairwise CKA toward zero, so a near-zero value mostly restates constraint satisfaction rather than showing that the subspaces encode distinct linguistic levels. CKA under an explicit orthogonality objective is not an interpretability metric for this purpose.

Corpus	Pooled CKA			Probe F_1			
	l/s	l/m	s/m	lex	syn	sem	rand
DiPPER	0.001	0.001	0.001	> 0.986			–
PAWS	0.981	0.967	0.951	0.745	0.738	0.733	0.759
QQP	0.992	0.973	0.957	0.756	0.752	0.762	0.760

Table 1: The frozen DiPPER-trained checkpoint evaluated on two out-of-domain corpora, 2,000 evaluation pairs each, probes averaged over five seeds. Subspace pairs are lexical/syntactic (l/s), lexical/semantic (l/m), and syntactic/semantic (s/m). Pooled CKA reverses from near zero to near one while the projection matrices are unchanged. A random projection of the raw encoder states, matched in dimension, exceeds all three named subspaces on PAWS and matches them to within seed variation on QQP, so the subspace labels carry no probe-relevant information. The frozen meta-classifier attains ROC AUC 0.442 on PAWS and 0.713 on QQP, against 0.998 in domain.

Subspace	Surface correlate (ρ)	Probe F_1
Lexical	Jaccard 0.479, length 0.442	> 0.986
Syntactic	none identifiable	> 0.986
Semantic	surface overlap 0.228	> 0.986

Table 2: Functional probing of the three subspaces. Every subspace classifies paraphrase in isolation at $F_1 > 0.986$, yet the grounding is uneven: the lexical subspace has a clear surface correlate, the semantic subspace stays tied to surface overlap, and no functional correlate was identified for the syntactic subspace. Equal task performance with unequal grounding indicates the subspaces are not linguistically specialized.

The failure is sharper than circularity alone. Linear CKA between two projected representations is governed by $\|S_\ell \Sigma S_{\ell'}^\top\|_F$, where Σ is the covariance of the pooled encoder states. The penalty drives $S_\ell S_{\ell'}^\top$ toward zero, but that forces the cross term to vanish only when Σ is close to isotropic on the relevant directions. CKA is therefore a joint property of the projections and the evaluation distribution, not of the model alone. Applying the same checkpoint without retraining to PAWS (Zhang et al., 2019) and QQP (Wang et al., 2018) makes the dependence explicit (Table 1): the projections still satisfy the constraint, with pairwise Frobenius cosines below 10^{-5} , yet pooled CKA rises to 0.95 through 0.99 across all three subspace pairs. One set of weights certifies near-perfect separation on one corpus and near-total similarity on another. A metric that reverses under distribution shift while the constraint it supposedly reflects stays satisfied cannot serve as evidence of learned specialization.

4.2 All subspaces solve the task equally

The decisive evidence is functional. Linear probing on each subspace in isolation yields individual F_1 above 0.986 for all three, comparable to the full model’s 0.9846. If the lexical subspace alone predicts paraphrase as well as the semantic subspace, the subspaces are not functionally distinct on the task, contradicting the framing that each captures a different linguistic granularity. The only clean functional signal is lexical: lexical magnitude correlates with token Jaccard overlap ($\rho = 0.479$) and length ratio ($\rho = 0.442$). The semantic subspace is not cleanly abstract, retaining $\rho = 0.228$ with surface overlap, and we identify no functional probe for the syntactic subspace (Table 2). The juxtaposition is the point: a subspace with no identifiable syntactic signal classifies paraphrase as well as the others. Orthogonality has produced three redundant solvers of the same task, not a lexical, a syntactic, and a semantic specialist.

4.3 The gain is capacity, not disentanglement

The ablation in Table 3 is reread for what it controls. The full model scores $F_1 = 0.9846$. A dual-encoder variant with the same two encoders and training procedure but a single shared projection and no auxiliary losses, “Fusion Without Disentanglement,” scores 0.9715. That variant is near parameter-matched to the full model, differing only in one projection versus three, well under one percent of the 361.5M total. The three-way split plus auxiliary supervision therefore buys 0.013 F_1 . The larger increments over the single encoders (+0.018 over DeBERTa-v3, +0.023 over RoBERTa) track the addition of a second encoder, that is, capacity and fusion in the manner of a trained ensemble. The foregrounded architectural novelty contributes the smallest of the three increments, far below what its complexity implies, at a large cost: 361.5M parameters and 0.52 ms per sample against 124.6M and 0.09 ms for the RoBERTa baseline.

5 Discussion and recommendations

One mechanism unifies the three findings: an explicit penalty shaped the representation and, through a metric sensitive to that same penalty, certified the shaping as meaningful. Low CKA was induced by the objective, equal probe performance shows the resulting subspaces are interchangeable, and the capacity-matched ablation shows the elaboration adds little. The pattern is not confined to

Model	Split	F1	Accuracy	Precision	Recall	AUC ROC	AUC PR	MCC
Full hierarchical model	Test	0.9846	0.9794	0.9825	0.9867	0.9978	0.9989	0.9536
Fusion without disentanglement	Test	0.9715	0.9619	0.9681	0.9750	0.9936	0.9968	0.9140
DeBERTa-v3-base only	Test	0.9661	0.9547	0.9624	0.9699	0.9913	0.9958	0.8977
RoBERTa-base only	Test	0.9619	0.9488	0.9541	0.9699	0.9898	0.9950	0.8841
Majority baseline	Test	0.80	0.67	0.67	1.00	0.50	0.67	0.00

Table 3: Ablation on the DiPPER test split, reread for the capacity control. Disentanglement adds only 0.013 F_1 over a near parameter-matched dual encoder without disentanglement. Following the row order above, $\Delta F_1 = 0.013$, 0.018, and 0.023 against the three learned baselines. The larger increments over the single encoders track the addition of a second encoder, not subspace separation.

representation geometry. Evaluation validity has been shown to differ sharply across environments in multi-agent LLM benchmarks, where failure-heavy settings can reward robustness to opponent errors as much as strategic ability (Wang et al., 2026). In both cases the reported number is sensitive to a factor other than the construct it is taken to measure. Three safeguards follow for work reporting subspace disentanglement. First, a similarity metric directly minimized by the training objective is not evidence of specialization; CKA, cosine, or Frobenius separation between explicitly orthogonalized projections should be reported as constraint satisfaction, not interpretability. Second, a capacity-matched ablation that removes the disentanglement mechanism while holding encoders, data, and training steps fixed is mandatory, since gains from added capacity are otherwise misattributed to the proposed design. Third, each named subspace requires a controlled functional probe, ideally one a non-specialized subspace would fail; a label such as “syntactic” must be earned by a signal that distinguishes it from the other levels, not assumed from its projection.

Limitations

The audit trains a single model on a single corpus; the PAWS and QQP results evaluate that same frozen checkpoint without retraining. For lack of compute, external baselines were not trained on the DiPPER split, so the capacity argument rests on internal ablations and remains DiPPER-only. The probes are task-coupled and small-scale, and a richer battery of surface-free syntactic probes might recover a signal not found here. The conclusion is therefore scoped to this recipe, orthogonality plus simplex penalties with attached level labels, not to all representation disentanglement.

References

- Zhengjia Dan, Yue Zhao, Xiaojun Bi, Licheng Wu, and Qiang Ji. 2022. Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy*, 24(10):1429.
- Yun He, Xue Feng, Cheng Cheng, Geng Ji, Yunsong Guo, and James Caverlee. 2022. Metabalance: Improving multi-task recommendations via adapting gradient magnitudes of auxiliary tasks. In *Proceedings of the ACM Web Conference 2022*, pages 2205–2215.
- John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.
- Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500.
- Bang Liu, Ting Zhang, Fred X. Han, Di Niu, Kunfeng Lai, and Yu Xu. 2018. Matching natural language sentences with hierarchical sentence factorization. In *Proceedings of the 2018 World Wide Web Conference*, pages 1237–1246.
- J. C. Macbeth, E. Chang, J. G. Chen, and S. Grandic. 2020. A broader range for ‘meaning the same thing’: Human against machine on hard paraphrase detection tasks. In *Proceedings of The Eighth Annual Conference on Advances in Cognitive Systems (ACS-2020)*.
- Jesus-German Ortiz-Barajas, Gemma Bel-Enguix, and Helena Gómez-Adorno. 2022. Sentence-CROBI: A simple cross-bi-encoder-based neural network architecture for paraphrase identification. *Mathematics*, 10(19).

- Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-argument based bi-encoder for paraphrase identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 353–355.
- Kevin Wang, Anna Thöni, Benjamin Kempinski, Bobby Cheng, Jianzhu Yao, Benjamin Finch, Leon Guertler, Viraj Nadkarni, Yihan Jiang, Aliaksei Korshuk, and 1 others. 2026. MINDGAMES: A live arena for evaluating social and strategic reasoning in multi-agent LLMs. *NeurIPS 2025 Datasets and Benchmarks Track*.
- Sheng Xu, Xingfa Shen, Fumiyo Fukumoto, Jiyi Li, Yoshimi Suzuki, and Hiromitsu Nishizaki. 2020. Paraphrase identification with lexical, syntactic and sentential encodings. *Applied Sciences*, 10(12):4144.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.
- Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1308.