Layer-Resolved Optimal Transport for Hallucination Detection in NMT and Abstractive Summarization

Mariia Onyshchuk Maksym-Vasyl Tarnavskyi Marta Sumyk

Motivation

Transformer hallucinations are not all the same. In neural machine translation (NMT), hallucination often appears as source disengagement: the decoder stops using the source sentence. In abstractive summarization, the model may still attend to relevant evidence but misrepresent its content.

This work studies whether Optimal Transport (OT) over cross-attention distributions can reveal where hallucination signals live across decoder layers, and whether this geometric signal transfers from NMT to summarization.

Central idea: OT is most useful when hallucination is visible as a geometric failure of source routing.

Research Questions

- **RQ1:** Which decoder layers contain the strongest OT-based hallucination signal in NMT?
- **RQ2:** Are Wass-to-Unif, Wass-to-Data, and routing consistency complementary detectors?
- **RQ3:** Does cross-attention geometry transfer to abstractive summarization faithfulness detection?

Study Methodology

Input: source text + output → Attention: layer ℓ, step t → OT metrics: WTU, WTD, Step-OT, RC → Analysis: detection + dynamics

$$\pi^{(\ell,t)} = \frac{1}{H}\sum_{h=1}^{H} \alpha^{(h,\ell,t)} \qquad W_1(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int |x-y|\, d\gamma(x,y)$$

We measure how decoder attention moves across source positions. Strong OT anomalies indicate source disengagement.

Datasets and Experimental Setup

Task	Dataset	Scope
NMT	Fairseq DE–EN	3,414 translations
Summarization	AggreFact	1,116 examples
Structural analysis	T5-base	12 decoder layers

Main contributions:

- layer-resolved OT analysis across Fairseq decoder layers;
- routing consistency as an unsupervised detector;
- transfer study from NMT hallucination to summarization faithfulness;
- structural analysis of T5-base cross-attention geometry.

Transfer to Summarization

The same OT signal transfers only partially to abstractive summarization. The detector performs above chance on AggreFact, but remains below supervised faithfulness detectors.

Table 1. Faithfulness detection on AggreFact, Balanced Accuracy (%).

Model	CNN	XSum	Avg
OT Detector (T5-base)	57.2	57.6	57.4
OT Detector (Flan-T5-large)	55.6	61.4	58.5
MiniCheck-Flan-T5-L	**69.9**	**74.3**	**72.1**
Random baseline	50.0	50.0	50.0

Reason for the gap: in NMT, hallucination often means retrieval failure. In summarization, faithfulness errors can occur after retrieval: the model may attend to the right source tokens but still distort the meaning.

Key Distinction

NMT: source disengagement

Summarization: content misuse

This explains why OT transfers only partially: attention geometry detects when the model stops using the source, but not always when the model uses the source incorrectly.

Main NMT Results: Layer-Resolved Detection

Layer-resolved OT shows that hallucination detection is concentrated in early-to-middle decoder layers and depends strongly on hallucination type.

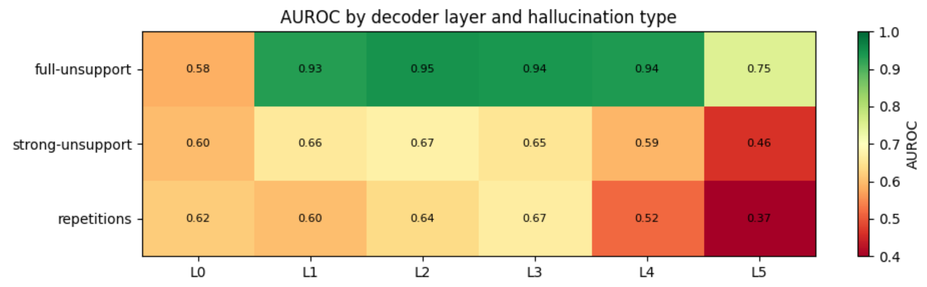


Figure 1. WTU AUROC by decoder layer and hallucination type. Full-unsupport is strongly detectable in L1–L4, while L5 becomes anti-predictive for subtler hallucination types.

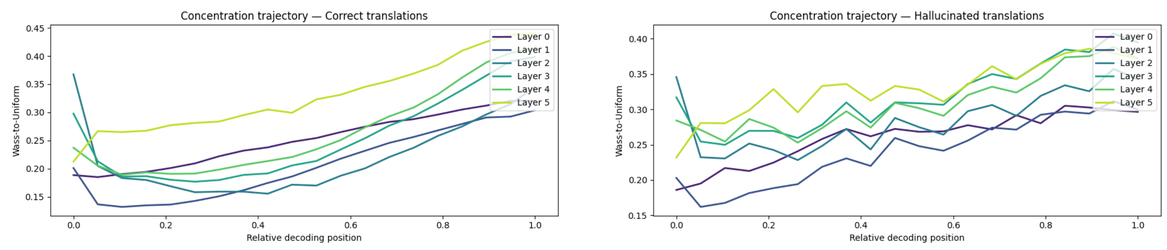
Table 2. Aggregate detector performance by hallucination type.

Hallucination type	WTD	WTU
Full-unsupport	0.801	**0.937**
Strong-unsupport	**0.770**	0.629
Repetitions	**0.790**	0.568

- **WTU** captures absolute attention concentration.
- **WTD** captures similarity to correct reference attention patterns.
- **RC** is strongest for full-unsupport hallucinations, reaching AUROC 0.957.

Generation Dynamics: Correct vs. Hallucinated

Correct translations show a two-phase trajectory: early exploration followed by commitment. Hallucinated translations skip the exploratory phase and begin in a static, concentrated state.



(a) Confirmed-correct (b) Hallucinated

Figure 2. Concentration trajectories split by quality group. Correct translations exhibit layer fan-out and an exploratory phase; hallucinated translations are compressed from the first decoding step.

Interpretation: hallucination behaves like static source routing rather than dynamic source scanning.

Structural Interpretability in T5-base

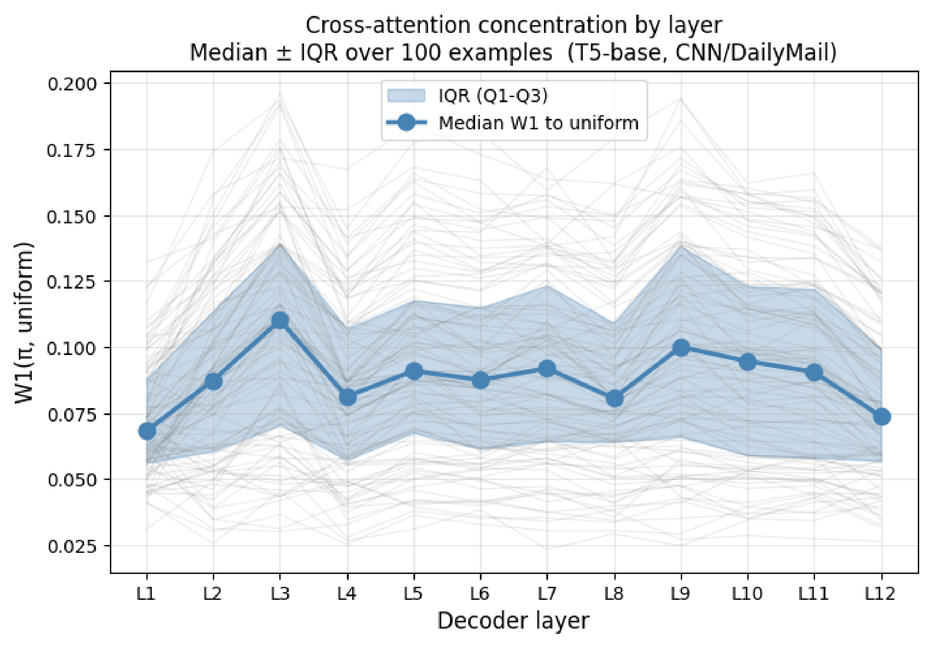


Figure 3. Cross-attention concentration by decoder layer in T5-base. Layer 3 shows the main concentration peak.

- **Layer 3** has peak concentration and strong routing distinctiveness.
- **Layers L1–L3** form an indispensable early context-encoding block.
- **Layer 12** is most critical for generation quality under ablation.

Table 3. Selected T5-base structural findings.

Layer / block	Interpretation
L3	peak concentration
L1–L3	critical early context block
L9–L11	relatively redundant cluster
L12	largest generation-quality drop when ablated

Conclusions

- OT on cross-attention is a reliable detector when hallucination reflects source disengagement.
- Layer-resolved analysis shows that the strongest NMT signal is concentrated in L1–L4.
- WTU, WTD, Step-OT, and RC are complementary, not interchangeable.
- Transfer to summarization is partial, because many faithfulness errors are content misuse rather than attention detachment.

OT is both a hallucination detector and an interpretability lens, but only for failures visible in attention geometry.