Enhanced Hallucination Detection in Neural Machine Translation through Simple Detector Aggregation

Anonymous ACL submission

Abstract

Hallucinated translations pose significant threats and safety concerns when it comes to practical deployment of machine translation systems. Previous research works have identified that detectors exhibit complementary performance — different detectors excel at detecting different types of hallucinations. In this paper, we propose to address the limitations of individual detectors by combining them and introducing a straightforward method for aggregating multiple detectors. Our results demonstrate the efficacy of our aggregated detector, providing a promising step towards evermore reliable machine translation systems.

1 Introduction

Neural Machine Translation (NMT) has become the dominant methodology for real-world machine translation applications and production systems. As these systems are deployed *in-the-wild* for real-world usage, it is ever more important to ensure that they are highly reliable. While NMT systems are known to suffer from various pathologies (Koehn and Knowles, 2017), the most severe among them is the generation of translations that are detached from the source content, typically known as *hallucinations* (Raunak et al., 2021; Guerreiro et al., 2022b). Although rare, particularly in high-resource settings, these translations can have dramatic impact on user trust (Perez et al., 2022). As such, researchers have worked on (i) methods to reduce hallucinations either during training-time or even inference time (Xiao and Wang, 2021; Guerreiro et al., 2022b; Dale et al., 2022; Sennrich et al., 2024), and alternatively, (ii) the development of highly effective on-the-fly hallucination detectors (Guerreiro et al., 2022b,a; Dale et al., 2022) to flag these translations before they reach end-users. In this paper, we will focus on the latter.

One immediate way to approach the problem of hallucination detection is to explore high-quality *external* models that can serve as proxies to measure detachment from the source content, e.g., quality estimation (QE) models such as CometKiwi (Rei et al., 2022), or cross-lingual sentence similarity models like LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2022). Intuitively, extremely low-quality translations or translations that are very dissimilar from the source are more likely to be hallucinations. And, indeed, these detectors can perform very effectively as hallucination detectors (Guerreiro et al., 2022b; Dale et al., 2022). Alternatively, another effective approach is to leverage *internal* model features such as attention maps and sequence log-probability (Guerreiro et al., 2022b,a; Dale et al., 2022). The assumption here is that when translation models generate hallucinations, they may reveal anomalous internal patterns that can be highly predictive and useful for detection, e.g., lack of contribution from the source sentence tokens to the generation of the translation (Ferrando et al., 2022). Most importantly, different detectors exhibit complementary properties. For instance, oscillatory hallucinations — translations with anomalous repetitions of phrases or n-grams (Raunak et al., 2021) — are readily identified by CometKiwi, while detectors based on low source contribution or sentence dissimilarity struggle in this regard. Therefore, there is an inherent trade-off stemming from the diverse anomalies different detectors excel at.

In this paper, we address this trade-off by proposing a simple yet highly effective method to aggregate different detectors to leverage their complementary strengths. Through experimentation in the two most widely used hallucination detection benchmarks, we show that our method consistently improves detection performance.

Key contributions are as follows:

- We propose STARE, an unsupervised Simple deTectors AggREgation method that achieves state-of-the-art performance well on two hallucination detection benchmarks.

- We demonstrate that our consolidated detector can outperform single-based detectors with as much as aggregating two complementary detectors. Interestingly, our results suggest that internal detectors, which typically lag behind external detectors, can be combined in such a way that they outperform the latter.

We release our code and scores to support future research and ensure reproducibility.[1]

2 Detectors Aggregation Method

2.1 Problem Statement

Preliminaries. Consider a vocabulary Ω and let (X, Y) be a random variable taking values in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \Omega$ represents translations and $\mathcal{Y} = \{0, 1\}$ denotes labels indicating whether a translation is a hallucination ($Y = 1$) or not ($Y = 0$). The joint probability distribution of (X, Y) is P_{XY}.

Hallucination detection. The goal of hallucination detection is to classify a given translation $x \in X$ as either an expected translation from the distribution $P_{X|Y=0}$ or as a hallucination from $P_{X|Y=1}$. This classification is achieved by a binary decision function $g : X \rightarrow 0, 1$, which applies a threshold $\gamma \in \mathbb{R}$ to a hallucination score function $s : X \rightarrow \mathbb{R}$. The decision function is defined as:

$$g(x) = \begin{cases} 1 & \text{if } s(x) > \gamma, \\ 0 & \text{otherwise.} \end{cases}$$

The objective is to create an hallucination score function s that effectively distinguishes hallucinated translations from other translations.

Aggregation. Assume that we have several hallucination score detectors[2]. When evaluating a specific translation x', our goal is to combine the scores from the single detectors into a single, more reliable score that outperforms any of the individual detectors alone. Formally, this aggregation method, denoted as Agg, is defined as follows:

$$\begin{aligned} \text{Agg} : \quad & \mathbb{R}^K \rightarrow \mathbb{R} \\ & \{s_k(x')\}_{k=1}^K \rightarrow \text{Agg}\left(\{s_k\}_{k=1}^K\right). \end{aligned}$$

[1] Code is available here: https://github.com/AnasHimmi/Hallucination-Detection-Score-Aggregation.

[2] We use the notation $\{s_k\}_{k=1}^K$ to represent a set consisting of K hallucination detectors, where each s_k is a function mapping from $\mathcal{X}$ to $\mathbb{R}$.

2.2 Proposed Aggregation Method

We start with the assumption that we have access to K hallucination scores and aim to construct an improved hallucination detector using these scores. The primary challenge in aggregating these scores arises from the fact that they are generated in an unconstrained setting, meaning that each score may be measured on a different scale. Consequently, the initial step is to devise a method for standardizing these scores to enable their aggregation. The normalization is performed using the min-max normalization based on the entire training dataset $\mathcal{D}_n = \{x_1, \ldots, x_n\}$. Formally, for a given score s_k, the normalized score s'_k is computed as follows:

$$s'_k = \frac{s_k(x') - \min_{z \in \mathcal{D}_n} s_k(z)}{\max_{z \in \mathcal{D}_n} s_k(z) - \min_{z \in \mathcal{D}_n} s_k(z)}.$$

Using these normalized scores, we construct a hallucination detector by summing them.

$$\text{Agg}(x') = \sum_{k=1}^{K} s'_k. \quad (1)$$

We denote this method as STARE.

3 Experimental Setup

3.1 Datasets

In our experiments, we utilize the human-annotated datasets released in Guerreiro et al. (2022b) and Dale et al. (2023). Both datasets include detection scores — both for internal and external detectors — for each individual translation:

LFAN-HALL. A dataset of 3415 translations for WMT18 German→English news translation data (Bojar et al., 2018) with annotations on critical errors and hallucinations (Guerreiro et al., 2022b). This dataset contains a mixture of *oscillatory* hallucinations and *fluent but detached* hallucinations. We provide examples of such translations in Appendix A. For each translation, there are six different detector scores: three are from external models (scores from COMET-QE and CometKiwi, two quality estimation models, and sentence similarity from LaBSE, a cross-lingual embedding model), and three are from internal methods (length-normalized sequence log-probability, Seq-Logprob; contribution of the source sentence for the generated translation according to ALTI+ (Ferrando et al., 2022), and WASS-COMBO, an Optimal

Detector	AUROC ↑	FPR ↓
Individual Detectors		
External		
COMET-QE	70.15	57.24
CometKiwi	86.96	35.15
LaBSE	91.72 🏅	26.86 🏅
Model-based		
Seq-Logprob	83.40	58.99
ALTI+	84.24	66.19
Wass-Combo	87.02	48.38
Aggregated Detectors		
External Only (gap to best single External)		
Isolation Forest	92.61 ↑0.89	19.08 ↓7.78
Max-Norm	92.43 ↑0.71	22.09 ↓4.77
STARE	93.32 ↑1.60	20.67 ↓6.19
Model-based Only (gap to best single Model-based)		
Isolation Forest	88.19 ↑1.17	36.63 ↓11.8
Max-Norm	83.81 ↓3.21	62.94 ↑14.6
STARE	89.07 ↑2.05	42.50 ↓5.88
All (gap to best overall)		
Isolation Forest	92.84 ↑1.12	23.90 ↓2.96
Max-Norm	91.60 ↓0.12	26.38 ↓0.48
STARE	**94.12** ↑2.40	**17.06** ↓9.80

(a) Results on LFAN-HALL.

Detector	AUROC ↑	FPR ↓
Individual Detectors		
External		
COMET-QE	82.22	47.40
LASER	81.11	47.04
XNLI	82.44	33.20
LaBSE	88.77 🏅	34.96 🏅
Model-based		
Seq-Logprob	86.72	28.86
ALTI+	82.26	58.40
Wass-Combo	64.82	84.62
Aggregation Detectors		
External Only (gap to best single External)		
Isolation Forest	71.35 ↓17.4	57.75 ↑22.8
Max-Norm	88.57 ↑0.48	32.59 ↓2.86
STARE	89.76 ↑0.99	32.74 ↓2.22
Model-based Only (gap to best single Model-based)		
Isolation Forest	75.35 ↓11.4	69.71 ↑40.9
Max-Norm	67.70 ↓17.3	83.83 ↑53.1
STARE	89.92 ↑3.20	30.37 ↑1.51
All (gap to best overall)		
Isolation Forest	76.25 ↓12.5	56.28 ↑21.3
Max-Norm	80.67 ↓7.01	41.52 ↑1.91
STARE	**91.18** ↑2.41	**28.85** ↓6.11

(b) Results on HALOMI.

Table 1: Performance, according to AUROC and FPR, of all single detectors available and aggregation methods via combination of external detectors, model-based detectors, or both simultaneously. We represent with 🏅 the best overall single detector and underline the best detectors for each class, according to our primary metric AUROC.

Transport inspired method that relies on the aggregation of attention maps).

HALOMI. A dataset with human-annotated hallucination in various translation directions. We test translations into and out of English, pairing English with five other languages — Arabic, German, Russian, Spanish, and Chinese, consisting of over 3000 sentences across the ten different language pairs. Importantly, this dataset has two important properties that differ from LFAN-HALL: (i) it has a much bigger proportion of fluent but detached hallucinations (oscillatory hallucinations were not considered as a separate category), and (ii) nearly 35% of the translations are deemed hallucinations, as opposed to about 8% for LFAN-HALL.[3] For each translation, there are seven different detection scores: the same internal detection scores as LFAN-HALL, and four different detector scores: COMET-QE, LASER, XNLI and LaBSE.

We provide more details on both datasets in Appendix A.

Aggregation Baselines. The closest related work is Darrin et al. (2023) on out-of-distribution detection methods, using an Isolation Forest (IF; Liu et al., 2008) for per-class anomaly scores. We adapt their method, employing a single Isolation Forest, and designate it as our baseline. Alternatively, we also consider a different way to use the individual scores and normalization weights in Equation 1: instead of performing a sum over the weighted scores, we take the maximum score. We denote this baseline as Max-Norm.

Evaluation method. Following Guerreiro et al. (2022a), we report Area Under the Receiver Operating Characteristic curve (AUROC) as our primary metric, and False Positive Rate at 90% True Positive Rate (FPR@90TPR) as a secondary metric.

Implementation details. For LFAN-HALL, we normalize the metrics by leveraging the held-out set released with the dataset consisting of 100,000 non-annotated in-domain scores. In the case of HALOMI, however, no held-out set was released. As such, we rely on sampling random splits that consist of 10% of the dataset for calibration. We

[3] Given the rarity of hallucinations in practical translation scenarios (Guerreiro et al., 2023), LFAN-HALL offers a more realistic simulation of detection performance.

repeat the process 10 different times. We report average scores over those different runs. We also report the performance variance in the Appendix.

3.2 Performances Analysis

Results on hallucination detection performance on LFAN-HALL and HaloMNI are reported in Table 1.

Global Analysis. STARE aggregation method consistently outperforms (i) single detectors' performance, and (ii) other aggregation baselines. Moreover, we find that the combination of all detectors — both model-based and external-based detectors — yields the best overall results, improving over the STARE method based on either internal or external models only. Importantly, these trends, contrary to other alternative aggregation strategies, hold across both datasets.

Aggregation of External Detectors. STARE demonstrates robust performance when aggregating external detectors on both LFAN-HALL and HALOMI: improvements in AUROC (over a point) and in FPR (between two to six points). Interestingly, we also observe that the best overall performance obtained exclusively with external models lags behind that of the overall aggregation. This suggests that internal models features — directly obtained via the generation process — contribute with complementary information to that captured by external models.

Aggregation of Internal Detectors. Aggregation of internal detectors, can achieve higher AUROC scores than the best single external detector on HALOMI. This results highlights how model-based features — such as attention and sequence log-probability — that are readily and efficiently obtained as a by-product of the generation can, when aggregated effectively, outperform more computationally expensive external solutions.

3.3 Ablation Studies

In this section, our focus is two-fold: (i) exploring optimal selections of detectors, and (ii) understanding the relevance of the reference set's size.

Optimal Choice of detectors. We report the performance of the optimal combination of N-detectors on both datasets in Table 2.[4] We note that including all detectors yields comparable performance to the best mix of detectors. Interestingly, aggregation always brings improvement,

[4]We report the optimal combinations in Appendix C.

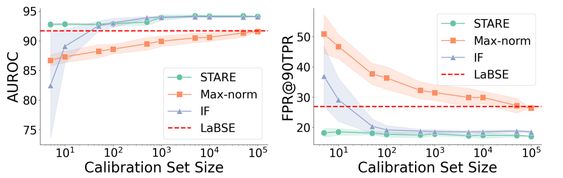

Figure 1: Impact of reference set size on LFAN-HALL.

even when only combining two detectors. As expected, the best mixture of detectors leverages information from different signals: contribution of source contribution, low-quality translations, and dissimilarity between source and translation.

	LFAN-HALL		HALOMI	
N	AUROC	FPR@90	AUROC	FPR@90
LaBSE	91.72	26.86	88.77	34.96
2	93.32	20.67	90.40	27.52
3	94.11	17.27	90.61	27.24
4	94.45	13.69	91.09	26.91
5	94.12	17.06	91.25	28.48
6	—	—	91.40	27.93
STARE	94.12	17.06	91.18	28.85

Table 2: Ablation Study on the Optimal Choice of Detectors when using STARE.

Impact of the size of the references set. The calibration of scores relies on a reference set. Here, we examine the impact of the calibration set size on performance, by ablating on the held-out set LFAN-HALL, which comprises of 100k sentences. Figure 1 shows that the ISOLATION FOREST requires a larger calibration set to achieve similar performance. This phenomenon might explain the drop in performance observed on HALOMI (Table 1). Interestingly, the performance improvement for STARE, particularly in FPR, plateaus when the reference set exceeds 1,000 samples, which suggests that STARE can adapt to different domains with a rather small reference set.

4 Conclusion & Future Perspectives

We propose a simple aggregation method to combine hallucination detectors to exploit complementary benefits from each individual detector. We show that our method can bring consistent improvements over previous detection approaches in two human-annotated datasets across different language pairs. We are also releasing our code and detection scores to support future research on this topic.

5 Limitations

Our methods are evaluated in a limited setup due to the limited availability of translation datasets with annotation of hallucinations. Moreover, in this study, we have not yet studied *compute-optimal* aggregation of detectors — we assume that we already have access to multiple different detection scores.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better. *arXiv preprint arXiv:2212.08597*.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. *arXiv preprint arXiv:2305.11746*.

Maxime Darrin, Guillaume Staerman, Eduardo Dadalto Câmara Gomes, Jackie CK Cheung, Pablo Piantanida, and Pierre Colombo. 2023. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M Guerreiro, Pierre Colombo, Pablo Piantanida, and André FT Martins. 2022a. Optimal transport for unsupervised hallucination detection in neural machine translation. *arXiv preprint arXiv:2212.09631*.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2022b. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *arXiv preprint arXiv:2208.05309*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *Preprint*, arXiv:2309.07098.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

A Model and Data Details

A.1 LFAN-HALL dataset

NMT Model. The model used in Guerreiro et al. (2022b) is a Transformer base model (Vaswani et al., 2017) (hidden size of 512, feedforward size of 2048, 6 encoder and 6 decoder layers, 8 attention heads). The model has approximately 77M parameters. It was trained on WMT18 DE-EN data: the authors randomly choose 2/3 of the dataset for training and use the remaining 1/3 as a held-out set for analysis. We use a section of that same held-out set in this work.

Dataset Stats. The dataset consists of 3415 translations from WMT18 DE-EN data. Overall, there are 218 translations annotated as detached hallucinations (fully and strongly detached — see more details in Guerreiro et al. (2022b)), and 86 as oscillatory hallucinations.[5] The other translations are either incorrect (1073) or correct (2048). We show examples of hallucinations for each category in Table 4.[6]

A.2 HALOMI dataset

NMT model. Translations on this dataset come from 600M distilled NLLB model (NLLB Team et al., 2022).

B Variance of performance on the HALOMI dataset

We report in Table 3 the average performance as well as the standard deviation across the different ten runs on different calibration sets. Despite variance between different runs, the STARE aggregation method consistently outperforms individual detectors and other aggregation techniques.

C Optimal Combination of Detectors via STARE

LFAN-HALL. The optimal set of detectors for various values of N is:

- for $N = 1$: LaBSE
- for $N = 2$: CometKiwi, LaBSE

[5] Some strongly detached hallucinations have also been annotated as oscillatory hallucinations. In these cases, we follow Guerreiro et al. (2022a) and consider them to be oscillatory.

[6] All data used in this paper is licensed under a MIT License.

Detector	AUROC ↑	FPR@90TPR ↓
Individual Detectors		
External		
COMET-QE	82.22 ± 0.28	47.40 ± 0.82
LASER	81.11 ± 0.21	47.04 ± 0.78
XNLI	82.44 ± 0.18	33.20 ± 0.63
LaBSE	88.77 ± 0.21	34.96 ± 0.72
Model-based		
Seq-Logprob	86.72 ± 0.22	28.86 ± 0.64
ALTI+	82.26 ± 0.28	58.40 ± 0.54
Wass-Combo	64.82 ± 0.20	84.62 ± 0.52
Aggregated Detectors		
External Only		
Isolation Forest	71.35 ± 1.62	57.75 ± 4.55
Max-Norm	88.57 ± 0.38	32.59 ± 0.60
STARE	89.76 ± 0.19	32.74 ± 0.50
Model-based Only		
Isolation Forest	75.35 ± 2.32	69.71 ± 5.01
Max-Norm	67.70 ± 1.31	83.83 ± 1.40
STARE	89.92 ± 0.20	30.37 ± 1.84
All		
Isolation Forest	76.25 ± 2.16	56.28 ± 6.29
Max-Norm	80.67 ± 1.37	41.52 ± 5.87
STARE	91.18 ± 0.20	28.85 ± 0.89

Table 3: Performance of individual and aggregated hallucination detectors on the HALOMI dataset, including average performance and standard deviations across ten different calibration sets.

- for $N = 3$: Wass_Combo, CometKiwi, LaBSE
- for $N = 4$: ALTI+, Wass_Combo, CometKiwi, LaBSE
- for $N = 5$: ALTI+, SeqLogprob, Wass_Combo, CometKiwi, LaBSE

HALOMI. The optimal set of detectors for various values of N is:

- for $N = 2$: LaBSE, SeqLogprob
- for $N = 3$: LaBSE, SeqLogprob, Wass-Combo
- for $N = 4$: LaBSE, SeqLogprob, XNLI, COMET-QE
- for $N = 5$: LaBSE, SeqLogprob, XNLI, COMET-QE, ALTI+
- for $N = 6$: LaBSE, Log Loss, XNLI, COMET-QE, ALTI+, Wass-Combo
- for $N = 7$: LaBSE, SeqLogprob, XNLI, COMET-QE, ALTI+, Laser, Wass-Combo

D Quantile transformation instead of min-max normalization

One drawback of min-max scaling is its vulnerability to outliers, as a single outlier can distort the entire distribution. We compare in this section STARE with a quantile transformation which maps all values into the [0, 1] range in a monotonic fashion and also makes the distribution of the resulting values approximately uniform. The results in Tables 5 and 6 show that Quantile-STARE demonstrates competitiveness STARE.

E Comparision with the majority vote

Below (Table 7) are the results (F1 score) for the majority vote baseline as it is not possible to define the AUROC or FPR.

F Contribution of metrics in the decision of STARE

To better understand the strength of STARE, we compare the mean of normalized scores for hallucination and non-hallucination. Tables 8 and 9 show that External detectors are the most discriminative and contribute the most to both benchmarks

G Additional results on other hallucination categories

Category	Source Sentence	Reference Translation	Hallucination
Oscillatory	Als Maß hierfür wird meist der sogenannte Pearl Index benutzt (so benannt nach einem Statistiker, der diese Berechnungsformel einführte).	As a measure of this, the so-called Pearl Index is usually used (so named after a statistician who introduced this calculation formula).	The term "Pearl Index" refers to the term "Pearl Index" (or "Pearl Index") used to refer to the term "Pearl Index" (or "Pearl Index").
Strongly Detached	Fraktion der Grünen / Freie Europäische Allianz	The Group of the Greens/European Free Alliance	Independence and Democracy Group (includes 10 UKIP MEPs and one independent MEP from Ireland)
Fully Detached	Die Zimmer beziehen, die Fenster mit Aussicht öffnen, tief durchatmen, staunen.	Head up to the rooms, open up the windows and savour the view, breathe deeply, marvel.	The staff were very friendly and helpful.

Table 4: Examples of hallucination types. Hallucinated content is shown shaded.

DETECTOR	AUROC ↑	FPR@90TPR ↓
External Only		
STARE	93.32	20.67
Quantile-STARE	93.09	16.03
Model-based Only		
STARE	89.07	42.50
Quantile-STARE	90.30	33.92
All		
STARE	94.12	17.06
Quantile-STARE	94.00	20.46

Table 5: Comparison of STARE with Quantile-STARE on LFAN-Hall

DETECTOR	AUROC ↑	FPR@90TPR ↓
External Only		
STARE	89.76 ± 0.19	32.74 ± 0.50
Quantile-STARE	90.06 ± 0.20	31.73 ± 0.44
Model-based Only		
STARE	89.92 ± 0.28	30.37 ± 1.84
Quantile-STARE	90.15 ± 0.14	28.09 ± 0.60
All		
STARE	91.18 ± 0.20	28.85 ± 0.89
Quantile-STARE	91.79 ± 0.18	29.39 ± 0.43

Table 6: Comparison of STARE with Quantile-STARE on HalOmi

	LFAN-Hall	HalOmi
Majority vote	0.74	0.76 ± 0.01
STARE	0.78	0.78 ± 0.003

Table 7: f1 scores of majority vote and STARE on the two datasets

METRIC	No Hallucinations	With Hallucinations
ALTI+	0.62	0.27
Seq-Logprob	0.57	0.23
Wass-Combo	-0.05	-0.43
CometKiwi	0.75	0.34
LaBSE	0.79	0.36

Table 8: Contribution of metrics in the decision of STARE on LFAN-Hall

METRIC	No Hallucinations	With Hallucinations
Seq-Logprob	0.82 ± 0.03	0.61 ± 0.07
ALTI+	0.69 ± 0.04	0.46 ± 0.03
COMET-QE	0.74 ± 0.03	0.52 ± 0.05
LaBSE	0.83 ± 0.01	0.50 ± 0.01
LASER	0.79 ± 0.01	0.59 ± 0.01
XNLI	0.74 ± 0.00	0.17 ± 0.00
Wass-Combo	0.96 ± 0.01	0.90 ± 0.03

Table 9: Contribution of metrics in the decision of STARE on HalOmi

DETECTOR	AUROC ↑	FPR@90TPR ↑
Individual Detectors		
External		
CometKiwi	91.36	27.17
LaBSE	81.19	53.72
Model-based		
Seq-Logprob	68.26	74.65
ALTI+	71.39	76.63
Wass-Combo	82.07	44.28
Aggregated Detectors		
External Only		
Isolation Forest	88.78	36.53
Max-Norm	88.18	33.16
STARE	89.86	29.02
Model-based Only		
Isolation Forest	68.15	81.14
Max-Norm	70.46	75.51
STARE	78.71	55.84
All		
Isolation Forest	86.60	32.17
Max-Norm	87.16	31.87
STARE	88.02	26.81

Table 10: LFAN-HALL, oscillations

DETECTOR	AUROC ↑	FPR@90TPR ↑
Individual Detectors		
External		
CometKiwi	85.30	37.02
LaBSE	98.05	2.13
Model-based		
Seq-Logprob	94.22	6.84
ALTI+	98.21	2.15
Wass-Combo	95.54	5.52
Aggregated Detectors		
External Only		
Isolation Forest	94.48	13.83
Max-Norm	94.71	16.41
STARE	96.56	7.53
Model-based Only		
Isolation Forest	97.49	2.14
Max-Norm	97.09	1.70
STARE	98.23	1.97
All		
Isolation Forest	97.63	4.99
Max-Norm	95.11	14.53
STARE	98.34	2.21

Table 11: LFAN-HALL, fully detached

DETECTOR	AUROC ↑	FPR@90TPR ↑
Individual Detectors		
External		
CometKiwi	78.90	46.37
LaBSE	85.80	32.53
Model-based		
Seq-Logprob	77.85	66.95
ALTI+	73.76	89.43
Wass-Combo	75.69	68.91
Aggregated Detectors		
External Only		
Isolation Forest	86.82	30.41
Max-Norm	85.81	34.04
STARE	85.01	30.86
Model-based Only		
Isolation Forest	79.96	60.54
Max-Norm	74.45	83.14
STARE	80.70	69.87
All		
Isolation Forest	88.05	29.71
Max-Norm	84.06	43.87
STARE	86.65	35.04

Table 12: LFAN-HALL, strongly detached

DETECTOR	AUROC ↑	FPR@90TPR ↓
Individual Detectors		
External		
score_comet_qe	73.01 $\pm$ 0.27	65.49 $\pm$ 0.59
score_labse	84.67 $\pm$ 0.15	39.40 $\pm$ 0.59
score_laser	75.65 $\pm$ 0.21	52.65 $\pm$ 0.37
score_xnli	83.56 $\pm$ 0.28	55.49 $\pm$ 0.96
Model-based		
score_log_loss	78.11 $\pm$ 0.18	54.99 $\pm$ 0.78
score_alti_mean	68.72 $\pm$ 0.12	79.10 $\pm$ 0.34
score_attn_ot	67.04 $\pm$ 1.31	83.67 $\pm$ 1.53
Aggregated Detectors		
External Only		
Isolation Forest	69.27 $\pm$ 1.80	57.30 $\pm$ 6.29
Max-Norm	84.60 $\pm$ 0.50	49.64 $\pm$ 5.67
Sum-Norm	85.79 $\pm$ 0.26	39.52 $\pm$ 1.33
Model-based Only		
Isolation Forest	65.29 $\pm$ 2.07	83.50 $\pm$ 3.69
Max-Norm	74.39 $\pm$ 1.51	70.16 $\pm$ 2.14
Sum-Norm	78.72 $\pm$ 0.71	62.86 $\pm$ 1.61
All		
Isolation Forest	70.87 $\pm$ 2.63	62.66 $\pm$ 6.34
Max-Norm	85.47 $\pm$ 1.02	49.49 $\pm$ 4.90
Sum-Norm	85.59 $\pm$ 0.25	42.08 $\pm$ 1.36

Table 13: HalOmi, High level language pairs, omissions

DETECTOR	AUROC ↑	FPR@90TPR ↓
Individual Detectors		
External		
score_comet_qe	49.38 $\pm$ 0.21	84.53 $\pm$ 0.41
score_labse	80.19 $\pm$ 0.23	48.89 $\pm$ 0.62
score_laser	70.84 $\pm$ 0.42	69.92 $\pm$ 0.59
score_xnli	59.00 $\pm$ 0.37	76.10 $\pm$ 0.88
Model-based		
score_log_loss	71.47 $\pm$ 0.42	71.01 $\pm$ 1.97
score_alti_mean	65.55 $\pm$ 0.43	77.76 $\pm$ 0.49
score_attn_ot	65.10 $\pm$ 0.44	80.71 $\pm$ 1.06
Aggregated Detectors		
External Only		
Isolation Forest	38.17 $\pm$ 2.27	94.90 $\pm$ 0.70
Max-Norm	75.29 $\pm$ 0.80	65.03 $\pm$ 1.24
Sum-Norm	77.39 $\pm$ 0.65	65.77 $\pm$ 1.69
Model-based Only		
Isolation Forest	60.23 $\pm$ 1.63	84.61 $\pm$ 1.52
Max-Norm	68.67 $\pm$ 1.02	78.98 $\pm$ 1.02
Sum-Norm	73.57 $\pm$ 0.72	70.70 $\pm$ 0.72
All		
Isolation Forest	45.54 $\pm$ 2.11	93.15 $\pm$ 1.04
Max-Norm	70.88 $\pm$ 1.28	75.28 $\pm$ 2.32
Sum-Norm	79.20 $\pm$ 0.58	63.32 $\pm$ 0.78

Table 14: HalOmi, Low level language pairs, hallucinations

Detector	AUROC ↑	FPR@90TPR ↓
Individual Detectors		
External		
score_comet_qe	50.44 ± 0.28	82.16 ± 0.51
score_labse	79.90 ± 0.29	49.44 ± 0.57
score_laser	71.31 ± 0.33	67.88 ± 0.60
score_xnli	61.80 ± 0.33	72.26 ± 0.86
Model-based		
score_log_loss	68.62 ± 0.40	71.91 ± 1.48
score_alti_mean	60.94 ± 0.46	84.44 ± 0.27
score_attn_ot	67.52 ± 0.38	76.24 ± 0.84
Aggregated Detectors		
External Only		
Isolation Forest	35.09 ± 1.67	95.53 ± 0.72
Max-Norm	76.49 ± 0.59	61.00 ± 1.28
Sum-Norm	78.62 ± 0.61	60.61 ± 1.49
Model-based Only		
Isolation Forest	60.55 ± 2.22	83.43 ± 1.90
Max-Norm	70.66 ± 0.82	75.42 ± 0.79
Sum-Norm	69.02 ± 0.81	76.23 ± 0.81
All		
Isolation Forest	42.53 ± 2.26	92.79 ± 1.07
Max-Norm	73.82 ± 1.20	70.49 ± 2.46
Sum-Norm	78.13 ± 0.51	62.33 ± 0.66

Table 15: HalOmi, Low level language pairs, omissions

Detector	AUROC ↑	FPR@90TPR ↓
Individual Detectors		
External		
score_comet_qe	73.41 ± 0.23	50.40 ± 0.48
score_labse	85.91 ± 0.13	40.33 ± 0.32
score_laser	76.22 ± 0.30	57.17 ± 0.50
score_xnli	75.33 ± 0.19	45.47 ± 0.35
Model-based		
score_log_loss	80.64 ± 0.16	49.37 ± 0.45
score_alti_mean	77.45 ± 0.12	60.82 ± 0.52
score_attn_ot	63.93 ± 0.65	84.80 ± 0.67
Aggregated Detectors		
External Only		
Isolation Forest	44.97 ± 1.47	96.32 ± 1.30
Max-Norm	85.02 ± 0.44	40.95 ± 0.76
Sum-Norm	85.41 ± 0.17	40.83 ± 0.32
Model-based Only		
Isolation Forest	65.86 ± 1.68	80.98 ± 2.16
Max-Norm	67.03 ± 0.88	81.95 ± 0.85
Sum-Norm	83.90 ± 0.55	46.37 ± 1.40
All		
Isolation Forest	53.16 ± 2.78	92.92 ± 2.44
Max-Norm	76.17 ± 0.97	51.74 ± 2.84
Sum-Norm	87.06 ± 0.21	38.33 ± 0.36

Table 16: HalOmi, all language pairs, hallucinations

Detector	AUROC ↑	FPR@90TPR ↓
Individual Detectors		
External		
score_comet_qe	64.15 ± 0.23	67.79 ± 0.31
score_labse	80.09 ± 0.13	47.70 ± 0.52
score_laser	74.40 ± 0.24	57.70 ± 0.63
score_xnli	74.09 ± 0.11	49.30 ± 0.26
Model-based		
score_log_loss	75.33 ± 0.16	60.34 ± 0.55
score_alti_mean	66.78 ± 0.14	79.71 ± 0.14
score_attn_ot	65.81 ± 1.28	83.62 ± 2.43
Aggregated Detectors		
External Only		
Isolation Forest	45.86 ± 2.06	95.42 ± 1.76
Max-Norm	81.32 ± 0.23	49.11 ± 0.26
Sum-Norm	78.21 ± 0.14	50.49 ± 0.37
Model-based Only		
Isolation Forest	60.90 ± 0.98	86.63 ± 1.08
Max-Norm	71.32 ± 1.04	73.63 ± 1.32
Sum-Norm	74.82 ± 0.94	66.71 ± 1.52
All		
Isolation Forest	50.65 ± 2.49	93.73 ± 2.29
Max-Norm	78.41 ± 0.65	51.49 ± 3.35
Sum-Norm	78.93 ± 0.25	50.68 ± 0.23

Table 17: HalOmi, all language pairs, omissions