Memorization Inheritance in Sequence-Level Knowledge Distillation for Neural Machine Translation

Verna Dankers[*]
University of Edinburgh
vernadankers@gmail.com

Vikas Raunak[‡]
Microsoft
viraunak@microsoft.com

Abstract

In this work, we explore how instance-level memorization in the *teacher* Neural Machine Translation (NMT) model gets inherited by the *student* model in sequence-level knowledge distillation (SeqKD). We find that despite not directly seeing the original training data, students memorize more than baseline models (models of the same size, trained on the original data)—3.4% for exact matches and 57% for extractive memorization—and show increased hallucination rates. Further, under this SeqKD setting, we also characterize how students behave on specific training data subgroups, such as subgroups with low quality or specific counterfactual memorization (CM) scores, and find that students exhibit greater denoising on low-quality subgroups. Finally, we propose a modification to SeqKD named Adaptive-SeqKD, which intervenes in SeqKD to reduce memorization and hallucinations. Overall, we recommend caution when applying SeqKD: students inherit both their teachers' superior performance *and* their fault modes, thereby requiring active monitoring.

1 Introduction

Memorization of noisy training data creates unexpected failure modes in *neural machine translation* (NMT) models (Raunak and Menezes, 2022), thus presenting a reliability risk when deploying them in the real world. To make NMT models inference-friendly, they are often trained using *sequence-level knowledge distillation* (SeqKD) (e.g., Bapna et al., 2022; Costa-jussà et al., 2022). This is a KD variant in which teachers generate synthetic targets for students (Hinton, 2014; Kim and Rush, 2016). SeqKD yields smaller student models whose performance is competitive with the teacher's performance. Follow-up work has focused primarily on modifying SeqKD objectives to further improve NMT performance (e.g., Wen et al., 2023; Zhang et al., 2023; Wang et al., 2023, 2024). Apart from being used for model compression, SeqKD has proved very beneficial for low-resource and long-tail data (Dabre and Fujita, 2020; Currey et al., 2020; Gumma et al., 2023; Zhou et al., 2024; De Gibert et al., 2024) but its applications extend beyond that as well, e.g., to continual learning (Chuang et al., 2020; Zhao et al., 2022).

☹ Extractive Memorization (ExMem) with respect to the initial parallel corpus increases 57.0%±15.4 in students compared to baselines (i.e. models memorized to emit the target even if we omit the italicized text):

(1) s_C Reprezentacja Trynidadu i Tobago *[w piłce nożnej]*
t_T Trinidad and Tobago national football team
t_S Trinidad and Tobago national football team

☹ Students have 31.0%±25.7 more oscillatory hallucinations (in blue) than baselines:

(2) s_C 1–5 , Stewards are appointed to publish the revelations (. . .)
t_T 1–5, Diener werden ernannt, um die Offenbarungen zu veröffentlichen (. . .)
t_S Die Heiligen sind in der Regel in der Regel in der Regel (. . .)

☹ Students show secondary ExMem (ExMem with respect to the teacher-generated corpus):

(3) s_C Electrical industry in Dominican *[Republic - AmarillasLatinas.net]*
t_T Elektrische Industrie in Dominikanische Republik
t_S Elektrische Industrie in Dominikanische Republik - AmarillasLatinas.net

☺ For low-quality source-target pairs, we observe amplified denoising in students:

(4) s_C La fiche du Pikauba par la Fromagerie Hamel.
t_C Pule » Teuerster Käse der Welt aus Eselsmiclh.
t_T Die Käserei Hamel in Pikauba. (Comet-QE-22=0.47)
t_S Die Geschichte des Pikauba durch die Hamel Käserei. (Comet-QE-22=0.62)
t_B Die Pikauba-Fassung wird von der Käserei Hamel betrieben. (Comet-QE-22=0.47)

Figure 1: An illustration of our findings. Sources (s) are from the corpus (C); translations (t) are from teachers, students and baselines (T, S, B).

[*]Work conducted during an internship at Microsoft.
[‡]Now at Google DeepMind.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 760–774
July 27 - August 1, 2025

Yet, the *understanding* of SeqKD lags behind its *usage*. Prior work in this direction primarily studies why SeqKD is successful, attributing it to mode reduction of the training data (Zhou et al., 2020; Song et al., 2021), or suggesting that SeqKD acts as a regularization technique (Gordon and Duh, 2019). In this work, we better try to understand how model behavior gets transmitted from teacher to student, moving beyond only analyzing average-case performance. In particular, we focus on how the student inherits instance-level memorization. Recent work on image classification suggests that KD inhibits memorization in the student (Lukasik et al., 2024), but also that membership inference attacks on the student are often successful (Jagielski et al., 2024). However, the connection between memorization and SeqKD in NLP is new territory; characterizing this is imperative to mitigate memorization-related failures in students.

Our main contributions are as follows: (1) We provide a quantification of **memorization inheritance** in §2: we identify that even though SeqKD inhibits memorization from teacher to student, the student memorizes more about the initial parallel corpus and hallucinates more than it would have, had it been trained without SeqKD. (2) We perform **subgroup analyses** in §3, for data subsets with specific characteristics, such as examples of different quality levels or with specific counterfactual memorization scores (Feldman, 2020). We identify subgroups for which the student outperforms both the teacher and baseline models. (3) To reduce memorization and mitigate accentuated hallucinations we propose **Adaptive-SeqKD** in §4: a simple intervention in the SeqKD algorithm wherein we adapt the teacher by finetuning it briefly on *intrinsically* obtained high-quality data to reduce memorization and hallucinations in the student.

Figure 1 demonstrates a subset of these findings with examples from our datasets.

2 Memorization inheritance

2.1 Experimental setup

In SeqKD, teacher θ_T is trained on a corpus with source sequences $\mathcal{S}_C$ and targets $\mathcal{T}_C$, and generates translations of $\mathcal{S}_C$ with beam size k. The source sequences and the teacher-generated translations ($\mathcal{S}_C$ and $\mathcal{T}_T$) form the training corpus for student θ_S. We use data from the WMT20 corpus (Barrault et al., 2020), for five language pairs: DE-EN and EN-DE (48M), PL-EN and EN-PL (12M), and FR-DE (14M). Appendix A provides more detail about the WMT corpora, and the validation and test data.

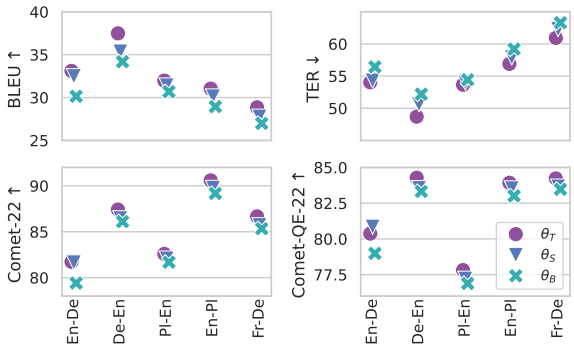


Figure 2: Performance of teacher, student and baseline models for four model quality metrics.

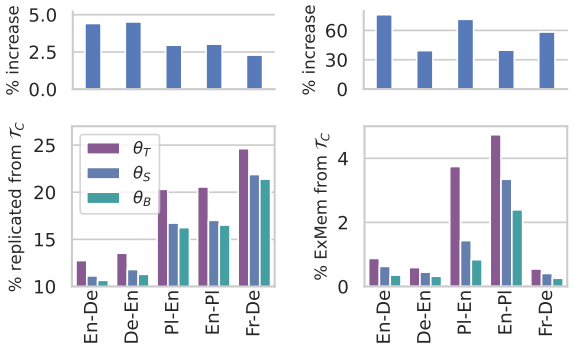
For all languages, a Transformer-large teacher trains for 300k steps on $\mathcal{S}_C$ and $\mathcal{T}_C$, followed by training a Transformer-base student for 100k steps on $\mathcal{S}_C$ and $\mathcal{T}_T$ (k=1).[1] We also train Transformer-base directly on $\mathcal{S}_C$ and $\mathcal{T}_C$, to have a baseline (θ_B) for what the student model would have memorized when exposed directly to WMT20 data. In Appendix C, we furthermore experiment with varying beam size k and the student's model size.

Model quality metrics We first examine the models' quality, to ascertain the SeqKD framework works as intended. We report the following reference-based metrics for translations generated with beam size five: **BLEU** and **chrF**, Translation Error Rate (**TER**), and the **Comet-20** and **Comet-22** metrics that use neural methods for translation quality estimation (Rei et al., 2020, 2022). We supplement this with the reference-free metrics of **Comet-QE-20**, **Comet-QE-22**.

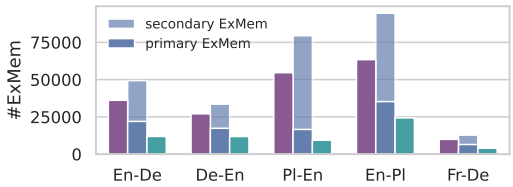
All metrics will be applied to WMT test data, and the reference-free methods are furthermore applied to translations of monolingual data: CommonCrawl data provided by Barrault et al. (2020), and data from the Pulpo poetry corpus (De la Rosa et al., 2023), to examine out-of-domain performance.

Memorization metrics We quantify memorization by comparing greedily translated $\mathcal{S}_C$ to $\mathcal{T}_C$ for all models, and to $\mathcal{T}_T$, for θ_S. We measure the **replication** (exact match) rate, and the **extractive memorization** (ExMem, Raunak and Menezes, 2022) rate. ExMem finds examples for which models memorized to emit the target after seeing at most

[1]Training duration set to equal the setup of Vaswani et al. (2017). We train using `MarianNMT` (Junczys-Dowmunt et al., 2018). For the full training setup see Appendix A and our codebase: `https://github.com/vernadankers/memseqkd`.

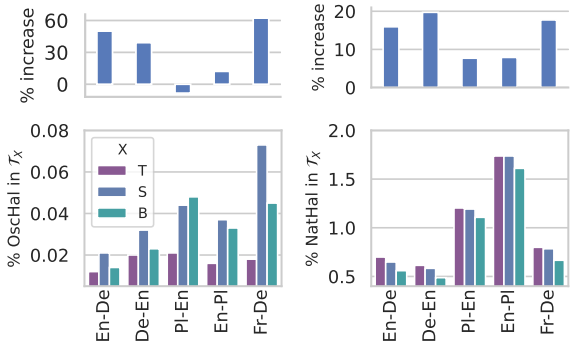


(a) Replication metric (b) ExMem rate



(c) Number of ExMem examples

Figure 3: Memorization metrics for θ_T, θ_S and θ_B and the percentual increase comparing θ_S to θ_B.



(a) OscHal metric (b) NatHal metric

Figure 4: Hallucination metrics for θ_T, θ_S and θ_B and the percentual increase comparing θ_S to θ_B.

75% of the source, e.g., see Example (1). The ExMem rate is the percentage of extractively memorized examples out of the replicated examples.

We also quantify the hallucination rate since hallucinations are often linked to models' memorization capabilities (e.g., Guerreiro et al., 2023; McKenna et al., 2023). We employ binary metrics because they are high-precision and do not require setting a threshold in the absence of ground truth data (Guerreiro et al., 2023; Raunak et al., 2021). We measure the rates of **natural hallucinations** (NatHal) and **oscillatory hallucinations** (OscHal). NatHal is the percentage of source sequences that map to a translation that is repeated in the model's translations at least five times. OscHal is the percentage of translations with bigrams repeated at least 10 times in the target but not the source.[2]

2.2 Results

General model quality Figure 2 and Appendix B provide performance differences for θ_T, θ_S and θ_B. Overall, θ_T outperforms θ_S, and θ_S outperforms θ_B. θ_S and θ_B merely differ in the training targets, which demonstrates that our SeqKD pipeline works as intended. The ordering of models in terms of their quality also holds for CommonCrawl and Pulpo data (see Table 3, Appendix B).

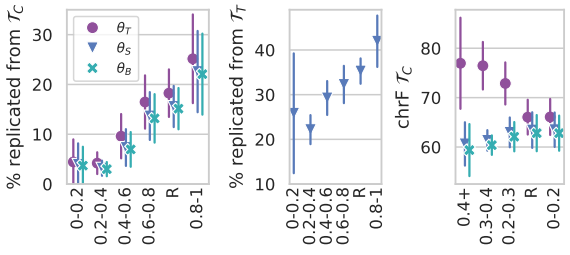
SeqKD facilitates memorization Figure 3 summarizes the memorization results. If we first look at the replication rate, θ_S replicates less from WMT20 than θ_T but **more than** θ_B: students' replication rate with respect to $\mathcal{T}_C$ is 3.4%(±0.9) higher than for θ_B. Students also replicate original material from θ_T: the overall student replication rate for $\mathcal{T}_T$ is 35.3%(±2.7) (see Table 4, Appendix B).

For the ExMem rate with respect to $\mathcal{T}_C$ (Figure 3b), a similar pattern emerges, but with a starker difference between θ_S and θ_B: students extractively memorize less from $\mathcal{T}_C$ compared to θ_T, but **more compared to** θ_B, with a mean increase of 57.0%(±15.4). This is quite surprising; note that by definition, the rate reported here expresses how many of the replicated examples are extractively memorized. Students only observed 18.4% (on average) of $\mathcal{T}_C$ through the SeqKD pipeline (the portion that θ_T replicated) and yet within that smaller pool they still memorized more than θ_B, that was exposed to 100% of the corpus. Not only have students extractively memorized WMT20 examples ('primary ExMem'), they also show ExMem with respect to θ_T ('secondary ExMem', quantified in Table 4, Appendix B). Figure 3c provides the absolute numbers of ExMem examples, distinguishing primary from secondary ExMem that constitute 41% and 59% of all ExMem examples, respectively. Example (3) demonstrates secondary ExMem: the student has memorized to hallucinate "AmarillasLatinas.net" from θ_T's target when merely shown the source's prefix, but θ_T has not.[3]

Why would SeqKD facilitate memorization? We hypothesize that this is due to its denoising function: if noisy data acts as a regularizer during train-

[2]To improve the memorization/hallucination metrics' precision, some training examples are excluded (see Appendix B, along with implementation details and hyperparameters).

[3]ExMem is a more widespread issue affecting commercial translation systems too, see Appendix E.



(a) Quality-based (b) Quality-based (c) CM-based

Figure 5: Illustration of how subgroups (indicated in sub-captions) vary in replication. Error bars indicate standard deviations over language pairs.

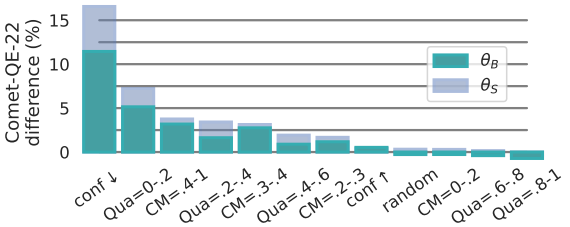


Figure 6: Comet-QE-22 increases compared to the teacher per subgroup, averaged over language pairs.

ing and θ_T partly filters that noise through SeqKD, training θ_S with reduced regularization could lead to increased memorization compared to θ_B. Reduced regularization is traditionally associated with increased memorization in machine learning. In §3, we demonstrate SeqKD's denoising function.

SeqKD amplifies hallucinations It is not just memorization that is amplified through SeqKD; Figure 4 indicates that **hallucinations are also amplified**. For the oscillatory hallucinations, both θ_S and θ_B generate more of these than θ_T, and the student hallucinates more than θ_B, on average (an increase of 31.0%±25.7, with no increase observed for PL-EN). Appendix B presents additional results for monolingual corpora CommonCrawl and Pulpo. Across the board, θ_S hallucinates most there, too.

For the natural hallucinations, θ_S and θ_B hallucinate less than θ_T, but students still hallucinate 13.8%(±5.0) more than θ_B.

3 Subgroup analysis

We now turn our attention to *data subgroups* to describe SeqKD's impact on samples with specific characteristics. We compute replication metrics (**exact match**, **chrF**), neural quality metrics (**Comet(-QE)-22**), and textual diversity (**MSTTR**). Subgroups contain up to 10k examples, with the exception of the random group.

Subgroups We construct 12 subgroups: Firstly, we randomly sample 50k examples from each of the WMT20 corpora. Secondly, we construct five **quality-based subgroups** by bucketing WMT20 examples based on the Comet-QE-22 metric, yielding five subgroups (with ranges <0.2, 0.2-0.4, 0.4-0.6, ≥0.8). Thirdly, we create **counterfactual memorization (CM) subgroups**. Assuming input x and target y, and a model with the parameters θ^{tr} trained on all training data, and θ^{tst} trained on all examples except (x, y), $\texttt{CM}(x, y) = p_{\theta^{\text{tr}}}(y|x) - p_{\theta^{\text{tst}}}(y|x)$ (Feldman, 2020; Feldman and Zhang, 2020). We approximate CM scores for all datapoints for EN-DE, and for 10% of the datapoints for the remaining language pairs, as detailed in Appendix F.1. We create four subgroups of examples (with ranges <0.2, 0.2-0.3, 0.3-0.4, ≥0.4). Finally, we create two **confidence-based subgroups**, by taking θ_T's log-probability averaged over the generated tokens for $\mathcal{T}_T$, selecting the top and bottom 10k examples.

Results We refer the reader to Appendix F for the full results, among which we consider the following patterns to be the most noteworthy:
(1) The random subgroup reflects memorization results: Exact and non-exact match-based metrics for this group (see chrF in Figure 9) follow the overall ranking of θ_T>θ_S>θ_B.
(2) Quality-based subgroups demonstrate (amplified) denoising: The lower the quality of the subgroup, the less models replicate from the corpus (Figure 5a,b). When we look at the Comet-QE-22 quality of the translations the models generate for low-quality subgroups, θ_S and θ_B perform better than θ_T, and θ_S outperforms θ_B (Figure 6). This effect is thus partially attributable to the capacity gap between large and base models, and partially to KD, and confirms prior work (e.g., Zhou et al., 2020) that suggested SeqKD has a denoising function: θ_T denoised WMT20 for θ_S, and now θ_S shows **amplified denoising** on that same data. Example (4) already illustrated this in the introduction.
(3) High counterfactual memorization examples are not replicated: Examples that have high CM for θ_T, do not stand out in terms of replication rates for students or baselines (Figure 5c), and show some amplified denoising (see Figure 6, although substantially less than low-quality subgroups).
(4) Low-confidence examples are not inherited: Finally, for examples for which θ_T generated translations with very low confidence, θ_S has low replication and strong amplified denoising (Figure 6).

Metric	$\theta_{T_{hq}}$	$\theta_{T_{ra}}$	$\theta_{S_{hq}}$	$\theta_{S_{ra}}$
BLEU	$0.0_{\pm 0.5}$	$-1.2_{\pm 0.8}$	$-0.2_{\pm 1.7}$	$-1.2_{\pm 1.6}$
C-QE-22	$+0.2_{\pm 0.3}$	$-0.2_{\pm 0.1}$	$+0.3_{\pm 0.3}$	$-0.1_{\pm 0.2}$
ExMem $\mathcal{T}_C$	$-13.1_{\pm 10.2}$	$-7.5_{\pm 9.3}$	$-23.7_{\pm 24.0}$	$-11.9_{\pm 29.5}$
OscHal	$-55.1_{\pm 11.3}$	$+7.7_{\pm 11.2}$	$-58.9_{\pm 7.5}$	$+9.8_{\pm 15.6}$

Table 1: Percentual performance change following Adaptive-SeqKD, compared to θ_T and θ_S, using high-quality (*hq*) or random (*ra*) data.

4 Adaptive-SeqKD reduces memorization

We previously observed increased memorization and hallucinations in θ_S compared to θ_B, but also amplified denoising on data subsets. We now investigate whether we can reduce students' failures without harming performance otherwise. We apply **Adaptive-SeqKD**, which briefly finetunes the teacher on high-quality data before producing $\mathcal{T}_T$. Instead of costly external metrics like Comet-QE, we use intrinsic metrics for data quality by selecting 500k sequences where θ_T nearly memorized the target (chrF>90), θ_T is confident (normalized translation score > 0.9), and source lengths exceed 5 tokens. We finetuned θ_T per language pair for 200 steps and compared it to finetuning on a random 500k sample meeting the length requirement.

Table 1 shows that teachers finetuned on the high-quality data (and their students) perform similar in terms of BLEU and Comet-QE, but much better in terms of ExMem and OscHal. Finetuning on random data reduced ExMem but not hallucinations. We include additional metrics in Appendix D.

Although we applied Adaptive-SeqKD here by finetuning on a training data subset, the technique could be modified for scenarios in which the training data is unknown. For instance, one could simply take a high-quality subset from a new source and finetune models on it prior to running SeqKD.

5 Related work

Memorization in NLP and NMT Memorization as a general topic has seen increased attention in NLP in recent years. A wide range of work examined how many examples large language models memorize and how to extract those memories (e.g. Carlini et al., 2021, 2023; Nasr et al., 2025), and identified how characteristics of datapoints, models and training techniques relate to memorization (e.g. Mireshghallah et al., 2022; Biderman et al., 2024; Prashanth et al., 2024; Lesci et al., 2024; Li et al., 2024). Alternative lines of related work focused on how memorization affects models internally, for instance, for factual memories (e.g. Geva et al., 2023) or idiomatic expressions (Haviv et al., 2023).

Only a small subset of related work investigated memorization in the NMT context: Raunak et al. (2021) identified that examples with high CM scores tend to elicit hallucinations from models more easily than other examples. In later work, Raunak and Menezes (2022) were the first to discuss the phenomenon of ExMem for NMT systems. Lastly, Dankers et al. (2023) examined the connection between datapoints' features and CM scores in NMT, and emphasized that memorization can sometimes benefit performance.

Knowledge distillation and memorization The connection to memorization and KD had, thus far, only been investigated for the vision domain, where Jagielski et al. (2024) studied membership inference attacks for image classifiers trained through distillation. Although distillation improved the average-case privacy (i.e. reduces memorization), the most vulnerable examples barely benefited from KD. Lukasik et al. (2024) studied CM of image classifiers, concluding that distillation inhibits CM. Some of our findings echo these results, namely that compared to θ_T, memorization is indeed dampened in the student, and that high-CM examples do not stand out in terms of replication.

6 Conclusion

SeqKD is popular for effectively training smaller NMT models, but we show that it also introduces issues like worsened ExMem and hallucinations in θ_S compared to θ_B. At the same time, the subgroup analyses showed that teachers' CM examples are not necessarily replicated by θ_S, and that students exhibit amplified denoising on low-quality examples. This highlights a paradox: through SeqKD, students memorize *more* about the corpus than θ_B, yet also outperform θ_B and θ_T on data where they *did not* memorize. Student improvements thus happen both by mimicking θ_T, and by deviating from θ_T. Future work could suppress memorization during SeqKD, refine Adaptive-SeqKD, and adjust hyperparameters[4] to create more robust SeqKD pipelines. We advise caution with SeqKD: students may inherit not only the teacher's strengths but also its failures, requiring careful monitoring beyond average-case performance.

[4]Increasing k reduces the student's OscHal rate to below that of θ_B, in particular, with Adaptive-SeqKD yielding further improvements (see Appendix C).

7 Limitations

We identify the following three limitations with our work:

- **Limited experimental setup**: inherent to selecting one (although common) experimental setup of distilling a large model into a smaller model, is that the findings need not necessarily transfer to other settings. We experimented with multiple language pairs, and also varied the beam size and student size in Appendix C to comment on this limitation, but recognize that there are other settings that would be interesting to study, such as distilling translation models from LLMs. We opted for the more common SeqKD setup, because of its popularity in the years past; most deployed translation systems are not LLM-based (yet). Besides, SeqKD as a technique is still alive and kicking as demonstrated by, for instance, the recently developed OpusDistillery library[5] (De Gibert et al., 2025).

- **Niche phenomena**: ExMem and hallucinations are phenomena one would only rarely encounter when interacting with NMT systems (Raunak and Menezes, 2022). We consider them worthy of investigation, nonetheless, because they represent extreme system failure that goes far beyond a simple mistranslation. Particularly for high-resource languages, NMT performance is approaching a ceiling, thanks to the sheer volume of translation examples available. Because NMT test performance is so high, it is important that NMT practitioners go beyond standard evaluation and look into domain-specific phenomena, long-tail phenomena (e.g. Raunak et al., 2022), and robustness failures. ExMem and hallucinations represent such failures.

- **Reliability issues not yet solved**: Adaptive-SeqKD substantially reduced the ExMem and hallucination rates, but did not resolve the issue altogether. We conducted limited experimentation in perfecting Adaptive-SeqKD and recognize this could be further expanded upon in the future.

[5] https://github.com/Helsinki-NLP/OpusDistillery

References

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2024. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2020. Lifelong language knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2914–2924.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Anna Currey, Prashant Mathur, and Georgiana Dinu. 2020. Distilling multiple domains for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4500–4511.

Raj Dabre and Atsushi Fujita. 2020. Combining sequence distillation and transfer learning for efficient low-resource neural machine translation models. In *Proceedings of the Fifth Conference on Machine Translation*, pages 492–502.

Verna Dankers, Ivan Titov, and Dieuwke Hupkes. 2023. Memorisation cartography: Mapping out the memorisation-generalisation continuum in neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8323–8343.

Ona De Gibert, Mikko Aulamo, Yves Scherrer, and Jörg Tiedemann. 2024. Hybrid distillation from RBMT and NMT: Helsinki-NLP's submission to the shared task on translation into low-resource languages of spain. In *Proceedings of the Ninth Conference on Machine Translation*, pages 908–917.

Ona De Gibert, Tommi Nieminen, Yves Scherrer, and Jörg Tiedemann. 2025. OpusDistillery: A configurable end-to-end pipeline for systematic multilingual distillation of open NMT models. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 201–208.

Javier De la Rosa, Álvaro Pérez Pozo, Salvador Ros, and Elena González-Blanco. 2023. Alberti, a multilingual domain specific language model for poetry analysis. *Procesamiento del Lenguaje Natural*, 71:215.

Vitaly Feldman. 2020. Does learning require memorization? A short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2881–2891.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Mitchell A Gordon and Kevin Duh. 2019. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *arXiv preprint arXiv:1912.03334*.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075.

Varun Gumma, Raj Dabre, and Pratyush Kumar. 2023. An empirical study of leveraging knowledge distillation for compressing multilingual neural machine translation models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 103–114.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264.

Geoffrey Hinton. 2014. Distilling the knowledge in a neural network. In *Deep Learning and Representation Learning Workshop in Conjunction with NIPS*.

Matthew Jagielski, Milad Nasr, Katherine Lee, Christopher A Choquette-Choo, Nicholas Carlini, and Florian Tramer. 2024. Students parrot their teachers: Membership inference on model distillation. *Advances in Neural Information Processing Systems*, 36.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *Proceedings of ACL 2018, System Demonstrations*, page 116.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2024. Causal estimation of memorisation profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15635.

Bo Li, Qinghua Zhao, and Lijie Wen. 2024. ROME: Memorization insights from text, logits and representation. *arXiv preprint arXiv:2403.00510*.

Michal Lukasik, Vaishnavh Nagarajan, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. What do larger image classifiers memorise? *Transactions on Machine Learning Research*.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In *Conference of the European Chapter of the Association for Computational Linguistics (17: 2023)*, pages 2758–2774.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. 2022. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable extraction of training data from aligned, production language

models. In *The Thirteenth International Conference on Learning Representations*.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir SV, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, et al. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon. *arXiv preprint arXiv:2406.17746*.

Vikas Raunak and Arul Menezes. 2022. Finding memo: Extractive memorization in constrained sequence generation tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5153–5162.

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.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Yuheng Song, Tianyi Liu, and Weijia Jia. 2021. Data diversification revisited: Why does it work? In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks*, pages 521–533. Springer.

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

Jun Wang, Eleftheria Briakou, Hamid Dadkhahi, Rishabh Agarwal, Colin Cherry, and Trevor Cohn. 2024. Don't throw away data: Better sequence knowledge distillation. *arXiv preprint arXiv:2407.10456*.

Shushu Wang, Jing Wu, Kai Fan, Wei Luo, Jun Xiao, and Zhongqiang Huang. 2023. Better simultaneous translation with monotonic knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2334–2349.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10817–10834.

Songming Zhang, Yunlong Liang, Shuaibo Wang, Yufeng Chen, Wenjuan Han, Jian Liu, and Jinan Xu. 2023. Towards understanding and improving knowledge distillation for neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8062–8079.

Yang Zhao, Junnan Zhu, Lu Xiang, Jiajun Zhang, Yu Zhou, Feifei Zhai, and Chengqing Zong. 2022. Life-long learning for multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:2212.02800*.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

Yuhang Zhou, Jing Zhu, Paiheng Xu, Xiaoyu Liu, Xiyao Wang, Danai Koutra, Wei Ai, and Furong Huang. 2024. Multi-stage balanced distillation: Addressing long-tail challenges in sequence-level knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3315–3333.

A Data and experimental setup

WMT data We download the parallel corpora from the WMT20 website, using sources listed in Table 2. For EN-DE we use the validation/test data from Raunak and Menezes (2022), which is a combination of WMT test data from recent years. For the other language pairs, we use the WMT20 test data, along with the WMT19 test data for validation during training. We honor the licensing terms by using the WMT data for research purposes only and citing the shared task article.

Source	EN-DE	FR-DE	PL-EN
Europarl	1.8M	1.8M	632k
ParaCrawl	34.4M	7.2M	6.6M
Common Crawl	2.4M	622k	-
News Commentary	362k	284k	-
Wiki Titles	1.4M	942k	1.0M
Tilde Rapid corpus	1.6M	-	278k
WikiMatrix	6.2M	3.4M	3.1M

Table 2: Composition of the 3 parallel corpora.

Additional data We run monolingual model evaluation using data from additional sources:

1. 1M **CommonCrawl** examples, sampled from the first 100M monolingual CommonCrawl datapoints provided by WMT20. To the best of our knowledge, our use is in line with CC's terms of use.
2. Up to 1M **Pulpo** examples, from De la Rosa et al. (2023)'s multilingual Prolific Unannotated Literary Poetry Corpus containing verses and stanzas. Pulpo contains monolingual sequences for all language pairs, apart from monolingual Polish data. We selected the data because it is expected to be out-of-distribution compared to the WMT20 training corpora. Pulpo contains poems that are copyright-free, or distributed under permissive licenses (De la Rosa et al., 2023, p.3).

Training and evaluation Models are trained using the `marian` toolkit (v1.12.16), using the setup of Raunak and Menezes (2022), that mimics the hyperpameters of Vaswani et al. (2017). θ_T trains for 300k steps; θ_S and θ_B train for 100k steps. We use 8 `Tesla V100-SXM2-32GB` GPUs for model training. When evaluating translations using Comet, we use models `wmt20-comet-da`, `wmt22-comet-da`, `wmt20-comet-qe-da` and `wmt22-cometkiwi-da`, using Comet v1.2.0. Visit our git repository for the training, evaluation and visualization code.

B Model quality and memorization

Model quality metrics Table 3 provides model quality metrics for all language pairs. Generally, θ_T>θ_S>θ_B (except for TER, that should be minimized instead of maximized), and where the results differ we italicized the numbers; this mostly happens for the monolingual CommonCrawl data. Yet, across the board, the pattern is clear for both WMT20 and monolingual data. The Pulpo scores are noticeably lower for both Comet-QE metrics. This could indicate that the poetry data is harder to translate, but we cannot reliably make a direct comparison across domains. Most importantly, the system ranking holds for this OOD data, too.

Lang.	θ	**WMT20**						**CommonCrawl**		**Pulpo**	
		BLEU	chrF	TER	Com-20	Com-22	Com-QE-20	Com-QE-22	Com-QE-20	Com-QE-22	Com-QE-20	Com-QE-22
EN-DE	θ_T	33.11	61.46	54.04	*41.77*	81.73	31.52	*80.37*	35.39	*73.11*	11.56	65.83
	θ_S	32.49	61.01	54.31	*42.00*	81.66	31.11	*80.85*	34.93	*74.00*	10.86	65.37
	θ_B	30.15	59.41	56.44	*34.35*	79.42	28.08	*78.99*	34.77	*72.97*	10.58	63.82
DE-EN	θ_T	37.49	64.95	48.70	68.97	87.42	49.40	84.28	38.19	77.01	9.46	64.72
	θ_S	35.36	63.65	50.65	65.87	86.46	46.27	83.55	37.02	76.77	8.37	63.29
	θ_B	34.18	63.11	52.20	64.68	86.11	45.39	83.32	37.02	76.55	8.36	62.54
PL-EN	θ_T	31.98	59.14	53.66	54.20	82.59	34.70	77.82	*34.35*	*74.58*	n/a	n/a
	θ_S	31.41	58.95	53.99	51.83	82.08	32.54	77.21	*34.08*	*74.72*	n/a	n/a
	θ_B	30.69	58.58	54.44	50.55	81.69	32.23	76.88	*34.33*	*74.44*	n/a	n/a
EN-PL	θ_T	31.04	59.86	56.91	92.07	90.59	74.56	83.93	*33.71*	*70.20*	12.24	61.58
	θ_S	30.13	58.98	57.91	89.09	89.77	71.99	83.53	*34.18*	*71.08*	11.46	61.16
	θ_B	28.94	58.27	59.21	86.52	89.17	70.41	83.01	*34.18*	*70.09*	10.94	59.55
FR-DE	θ_T	28.86	60.70	60.94	60.16	86.67	48.24	84.23	33.86	72.24	9.89	58.44
	θ_S	27.77	60.22	62.31	57.08	85.71	46.60	83.66	33.21	72.04	9.42	58.38
	θ_B	26.99	59.69	63.30	55.95	85.34	46.06	83.47	32.90	71.58	9.03	57.30

Table 3: Model performance for the five language pairs, using the experimental setup described in §2.1.

Memorization-related metrics Table 4 provides memorization metrics for all language pairs. For replication, ExMem and NatHal rates, it holds that θ_T>θ_S>θ_B, whereas for oscillatory hallucinations, θ_S typically hallucinates most, although there are exceptions (highlighted in italics).

Lang.	θ	WMT20						CommonCrawl	Pulpo
		Replic. $\mathcal{T}_C$	Replic. $\mathcal{T}_T$	ExMem $\mathcal{T}_C$ (#)	ExMem $\mathcal{T}_T$ (#)	NatHal	OscHal	OscHal	OscHal
EN-DE	θ_T	12.75	n/a	0.875 (36k)	n/a	0.699	0.012	0.020	*0.008*
	θ_S	11.12	32.75	0.627 (22k)	0.397 (49k)	0.648	0.021	0.029	*0.005*
	θ_B	10.65	n/a	0.356 (12k)	n/a	0.559	0.014	0.021	*0.009*
DE-EN	θ_T	13.53	n/a	0.590 (27k)	n/a	0.614	0.020	0.024	*0.062*
	θ_S	11.80	35.17	0.446 (17k)	0.246 (34k)	0.583	0.032	0.037	*0.048*
	θ_B	11.29	n/a	0.320 (12k)	n/a	0.487	0.023	0.026	*0.024*
PL-EN	θ_T	20.31	n/a	3.744 (55k)	n/a	1.203	*0.021*	*0.106*	n/a
	θ_S	16.73	34.65	1.433 (17k)	3.143 (79k)	1.192	*0.044*	*0.135*	n/a
	θ_B	16.25	n/a	0.836 (9k)	n/a	1.107	*0.048*	*0.136*	n/a
EN-PL	θ_T	20.56	n/a	4.736 (63k)	n/a	1.737	0.016	*0.078*	0.017
	θ_S	17.02	33.64	3.347 (35k)	4.126 (94k)	1.737	0.037	*0.100*	0.029
	θ_B	16.52	n/a	2.394 (24k)	n/a	1.610	0.033	*0.126*	0.027
FR-DE	θ_T	24.62	n/a	0.546 (10k)	n/a	0.799	0.018	0.025	0.190
	θ_S	21.88	40.39	0.407 (6k)	0.330 (13k)	0.784	0.073	0.082	0.223
	θ_B	21.39	n/a	0.257 (4k)	n/a	0.666	0.045	0.046	0.081

Table 4: Memorization metrics for the five language pairs, using the experimental setup described in §2.1.

Additional information on ExMem and hallucination metrics To improve the precision of the ExMem and hallucination metrics, some groups of examples are excluded from the computation:

- **ExMem**: in some cases, it is justified that the model emits the target after having processed only 75% of the source, e.g., if the target paraphrases the source. To improve the precision, we thus exclude the following examples when computing ExMem: a) examples with a source shorter than 4 words, b) examples with incorrect source or target languages, c) examples for which the length ratios between source and target are over 1.3, d) examples for which the source equals the target.
- **OscHal**: examples with source sequences of at least 50 white-space tokenized tokens are excluded, because it becomes more likely that the repeated bigrams might be accurate for longer sequences. This only concerns a small portion of the training data, e.g., only 3.4% for EN-DE. We count a sequence as a hallucination if the most frequent bigram appears more than 10 times in the translation, and at least 4 times more often than in the source. We experimentally verified that when reducing that maximum count of 10, the OscHal rate increases, but the model ranking remains the same.
- **NatHal**: for the natural hallucinations, we exclude examples for which the Comet-QE-22 score for the source and the generated translation is above 0.85. This is to exclude cases where natural hallucinations are detected simply because there are source sequences that are each other's paraphrase, for instance if both "Thank you for your visit at our website." and "Thanks for visiting our website." map to "Vielen Dank für Ihren Besuch auf unserer Website.".

C Varying SeqKD hyperparameters

Figure 7 demonstrates how model quality and memorization metrics change when we increase the beam size used to decode $\mathcal{T}_T$ (for both FR-DE and EN-DE), or change the student's model size (for EN-DE). θ_{S_L} and θ_{S_S} have hidden dimensionalities of 1024 and 256, respectively (with corresponding feedforward sizes of 4096 and 1024). When increasing the beam size: i) the quality of the students' translations decrease (for EN-DE) or remain mostly stable (for FR-DE), ii) the replication rates slightly increase, and iii) ExMem decreases, although not consistently: for FR-DE, the students' ExMem rates all exceed the baseline, whereas for DE-EN and $k \in \{2, 5\}$ the students are slightly below θ_B. This is still concerning, considering that students were exposed to 18.4% (on average) of the original corpus, whereas the baseline has seen 100%, so even similar ExMem rates between θ_S and θ_B suggest KD facilitates memorization. The NatHal rate slightly reduces but still far exceeds the baseline. The only real improvement larger beam sizes appear to make are reducing OscHal, both on WMT20 data (OscHal), and on external data

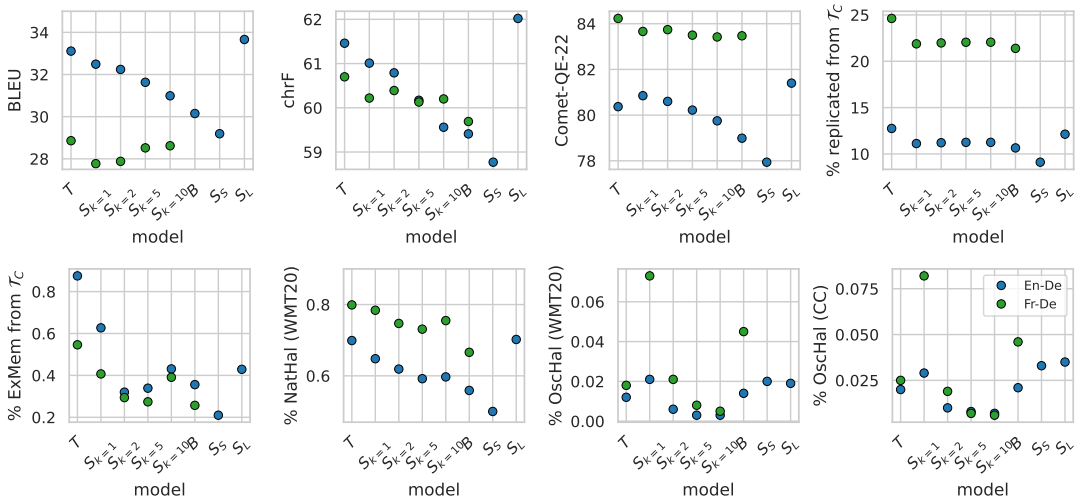


Figure 7: Illustration of how model quality and memorization metrics change as a result of changing the SeqKD beam size for EN-DE.

(OscHal CommonCrawl). With slight performance reduction in EN-DE and no performance reduction in FR-DE this yields a simple KD lesson: even if beam search is more computationally expensive (particularly when applied to the *entire* training corpus): do not use greedy search in KD.

Reducing the student's model size reduces the model's translation quality, and reduces replication, ExMem and NatHal, but still yields similar OscHal rates. The larger model produces better translations than its teacher and has lower ExMem, but also increased hallucinations.

D Adaptive-SeqKD

Figure 8 displays how performance changes per language pair due to Adaptive-SeqKD. Changes to quality metrics like BLEU and Comet(-QE)-22 are relatively minor, but the ExMem and hallucination rates are strongly affected. Finetuning with the high-quality data (*hq* in the graphs) decreases ExMem (for four of five language pairs) and hallucinations (all language pairs); finetuning with random data (*ra* in the graphs) is somewhat effective in reducing ExMem, but mostly fails to effectively reduce the hallucination rates.

In the previous section, we observed that increasing the SeqKD beam size is beneficial for the reduction of the hallucination rate. We thus reran finetuning with high-quality data for $S_{k=5}$ for EN-DE to examine whether finetuning also helps for an increased beam size. Upon doing so, the OscHal rate reduced with 28% for the WMT20 data and 33% for the CommonCrawl data, and the NatHat rate reduced with 10% for the WMT20 data, suggesting the wider applicability of Adaptive-SeqKD.

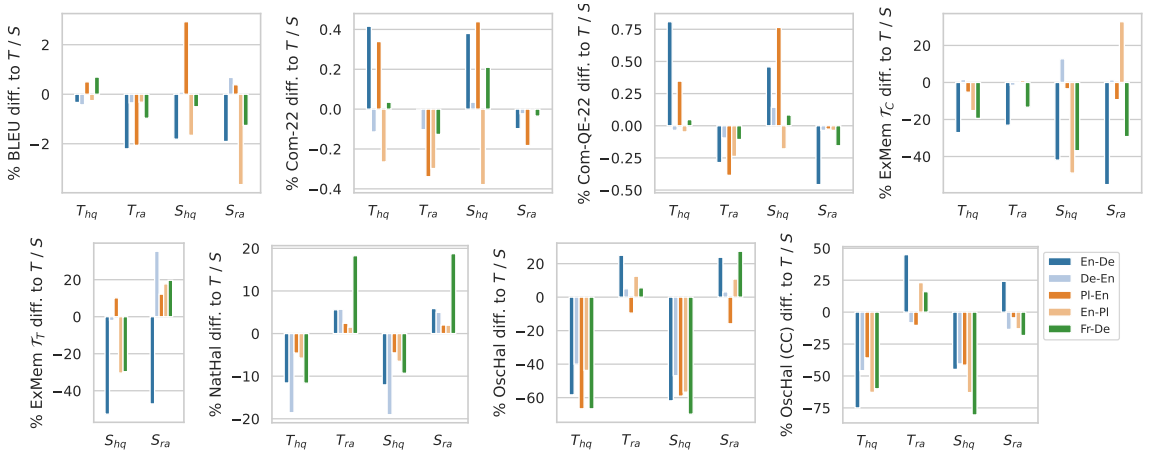


Figure 8: Performance changes observed for the different language pairs when applying Adaptive-SeqKD. Changes are computed as percentages with respect to the original teacher θ_T and θ_S.

E Extractive memorization in commercial systems

In the main paper, we discussed ExMem for the translation systems we trained, but it should be noted that ExMem does not only exist for sytems trained for academic purposes. To assert this, we assessed closed-source translation systems (`Google Translate`, `DeepL` and `Microsoft Bing Translator`),[6] using 70 Europarl examples that led to ExMem for our WMT20 English-German models. ExMem is defined with respect to the training targets, and we do not know what closed-source translation systems are trained on. Still, because of how widespread Europarl is in translation corpora, it may have been a part of the training material for closed-source systems, too. For `DeepL`, the system outputs the target translation without having processed the full source sequence for 25 of those 70 examples. For `Google Translate` and `Microsoft Bing Translator`, that applies to 16 and 20 examples, respectively.

Four examples are provided below. In each source text below, if we omit the italicised part in square brackets, the translations remain the same. Consider the following two examples from `DeepL`:

(5) s I have received seven motions for resolution tabled in accordance with Rule 103(2) *[of the Rules of Procedure]*
t Ich habe sieben Entschließungsanträge erhalten, die gemäß Artikel 103 Absatz 2 der Geschäftsordnung eingereicht wurden

(6) s I have therefore abstained *[from the vote]*
t Ich habe mich daher der Stimme enthalten

Consider the following example from `Microsoft Bing Translator`:

(7) s Mr President, the opposite is *[the case]*
t Herr Präsident, das Gegenteil ist der Fall

Consider the following example from `Google Translate`:

(8) s Mr President, Madam Commissioner, ladies *[and gentlemen]*
t Herr Präsident, Frau Kommissarin, meine Damen und Herren

These are examples copied directly from Europarl, but the underlying robustness issue also pops up when using non-exact variants – e.g. consider the following variant of Example (8): "A cleaner, a teacher and a commissioner said: ladies and". `Google Translate` still translates this including "Meine Damen und Herren" in its translation. It is hard to determine in which other domains ExMem occurs for these models, because we do not have a lot of insight into their training data, but that it is not something that only applies to our models, is for certain.

It is vital that we learn to understand what causes this type of memorization and how to avoid it. We identified knowledge distillation as contributing to exacerbating these types of memorization. We hope our insights can help both scientists and practitioners tackle such problems across different production systems as well, especially since deployed systems tend to be distilled (student) models.

F Subgroup analyses

Before elaborating on the results for the subgroup analysis, we more elaborately explain how we approximated the CM scores.

F.1 Composing the counterfactual memorization subgroups

Assuming input x and target y, and a model with the parameters θ^{tr} trained on all training data, and θ^{tst} trained on all examples except (x, y), CM can be computed as follows (Feldman, 2020; Feldman and Zhang, 2020):

$$\texttt{CM}(x,y) = \underbrace{p_{\theta^{\text{tr}}}(y|x)}_{\texttt{IN}} - \underbrace{p_{\theta^{\text{tst}}}(y|x)}_{\texttt{OUT}}$$

[6] As accessed on the 30th of March, 2025. For `Microsoft Bing Translator`, the 'formal tone' was used for translation.

Leaving out individual datapoints, as this equation suggests, is computationally too expensive given the vast sizes of our WMT20 datasets. We, therefore, approximate CM scores for all datapoints for EN-DE, and for 10% of the datapoints for the remaining language pairs. Following Dankers et al. (2023, p.3), who previously computed CM scores for NMT examples, we replace the probability of the full target with the geometric mean of the target token probabilities. This is more robust to length differences than the full target probability.

For EN-DE, we approximate CM by training 10 teacher models on a randomly sampled 50% of the training corpus, while evaluating it on the remaining 50%, such that for each datapoint, the 'IN' and 'OUT' quantities in this equation are both estimated using five models.

For the remaining language pairs, we use the original θ_T model to estimate the 'IN' quantity, and train another model according to the same training procedure on 90% of the training data, to estimate the 'OUT' quantity for 10% of the remaining training data. This is a rather coarse estimation, but should suffice to determine generic relations between CM and our evaluation metrics of interest.

Using the CM scores, we create six subgroups of interest. The last four were also contained in the main paper, and here we add the first two, that separate examples with a low CM score into two groups:

- IN and OUT performances ≤ 0.2, marked 'L(ow)-L(ow)' in the figures;
- IN and OUT performances ≥ 0.8, marked 'H(igh)-H(igh)' in the figures;
- $\{[0, 0.2), [0.2, 0.3), [0.3, 0.4), [0.4, 1.0]\}$.

F.2 Results

Random subgroup Our baseline subgroup consists of 50k examples randomly sampled from each of the WMT20 corpora. Replication metrics (chrF in Figure 9) follow the overall patterns of θ_T>θ_S>θ_B, consolidating that SeqKD dampens memorization in θ_S compared to θ_T but amplifies it compared to θ_B. The quality and diversity metrics (Comet-QE-22 and MSTTR in Figure 9) emphasize that, compared to the corpus (column $\mathcal{T}_C$), the models generate higher-quality translations, but with lower textual diversity.

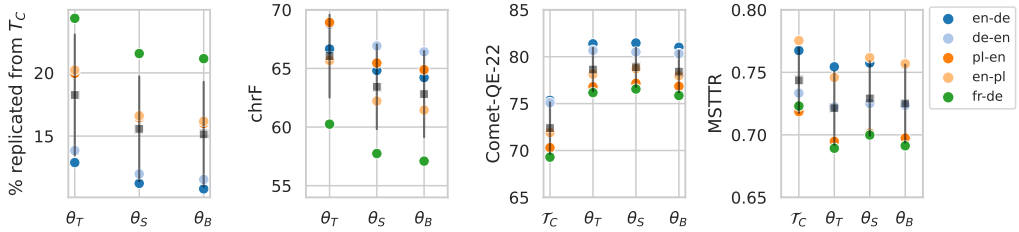


Figure 9: Evaluation metrics applied to the random subgroup, for all five language pairs. The square marker indicates the mean and standard deviation.

Quality-based subgroups How does KD affect examples from the WMT20 corpus with a certain quality? To categorize WMT20 examples based on quality, we applied the Comet-QE-22 metric using the corpus's targets as translations (because the reference is now the translation, we apply Comet's 'reference-free' method). We examine five subgroups: $\{[0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1.0]\}$.

In Figure 10, we first notice that for the low-quality subgroups, fewer translations are replicated from $\mathcal{T}_C$ (all models) and from $\mathcal{T}_T$ (for θ_S, see Figure 5b). Secondly, for groups with a quality below 0.6, both θ_S and θ_B generated better translations than θ_T, as per Comet-QE-22 applied to the model-generated translations. Textual diversity metrics here mostly follow the same trend as Comet-QE-22, with θ_S showing the highest diversity.

Figure 13 visualizes the relative differences of students/baselines to the teacher more explicitly, as a percentual increase. For Comet-QE-22, if the student's bar exceeds the baseline's bar, this signifies that the improvement over the teacher is partially attributable to the capacity gap between large and base models, and partially to the SeqKD process. This improvement over the teacher holds for the lower quality groups (0-0.2, 0.2-0.4, 0.4-0.6), but not for the higher quality groups. Example (4) from the introduction already illustrated a sample from the lowest quality subgroup: in the corpus, the target was likely misaligned. θ_T's

translation is slightly better but still wrong, and θ_S's translation is slightly better than θ_B's. Compared to θ_B, θ_S benefits from being presented with the teacher's corpus, which is a **denoised** version of WMT20, since θ_T replicates the least for the lowest-quality (noisiest) examples. θ_S, in turn, replicates less from $\mathcal{T}_T$ for these lowest-quality subgroups, too, which allows for **amplified denoising** compared to θ_T.

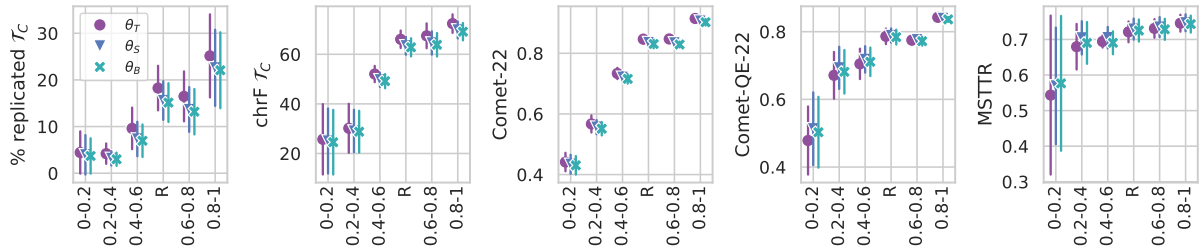


Figure 10: Evaluation metrics applied to the quality subgroups, aggregated over language pairs with error bars indicating standard deviation over language pairs.

Counterfactual memorization subgroups If we now focus on the high CM subgroups (introduced in the previous subsection) in Figure 11, we firstly observe that in terms of memorization metrics, the teacher shows increased replication and chrF. This follows from the definition of CM as the `IN` metric is expected to correlate highly with replication: the higher the target probability, the more likely it is that the teacher can replicate the target when generating translations. We note, however, that the student and baseline replicate examples with high CM less than examples with lower CM. This is likely due to their lower capacity: by definition, CM highlights examples for which a model assigns a low probability to the target when that example is not in the training set, thus requiring more capacity/parameters from the original θ_T to learn that target. The reduced replication also leads to lower Comet-22 scores for these examples, since that metric partially relies on the corpus's target.

When looking at the Comet-QE-22 scores, we observe that the higher CM groups typically also have lower Comet-QE-22 scores, although the quality scores are still above the lowest quality subgroup discussed in the previous paragraph. The baseline and student models outperform θ_T here, likely because they struggle to replicate the somewhat noisy targets as well as the teacher did. The student shows some amplified denoising compared to θ_B, but that does apply to CM groups across the board, and does not seem specific to individual CM subgroups.

All three models show more textual diversity for subgroups with higher CM scores.

Finally, if we inspect the two groups with very low and very high `IN` and `OUT` scores, the models—as expected based on the CM definition—do not replicate the 'low' group, and have very high replication rates for the 'high' group. For the remaining metrics, the 'low' group underscores the findings we previously reported for the low-quality subgroups: the student model shows amplified denoising, as reflected by the Comet-QE-22 metric, and also shows more textual diversity than the other two models. Examples that score badly in terms of both `IN` and `OUT` performance are typically low-quality, misaligned examples (Dankers et al., 2023); if the teacher does not replicate those low-quality targets, but partially denoises by translating them more accurately than the corpus did, the student can further improve upon that.

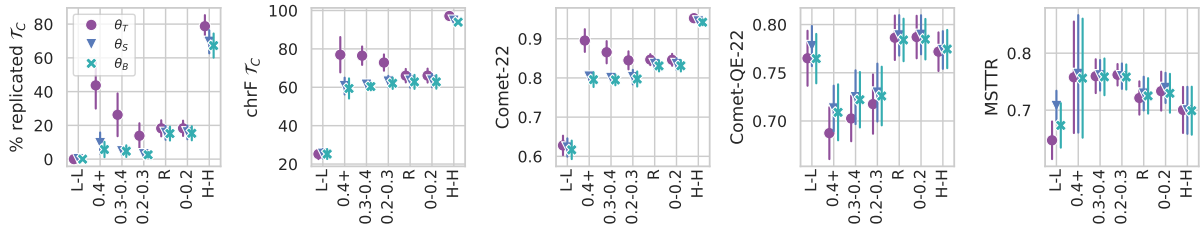


Figure 11: Evaluation metrics applied to the CM subgroups, aggregated over language pairs with error bars indicating standard deviation over language pairs.

Confidence subgroups Finally, we turn to the confidence-based groups, created using the lowest 10k and highest 10k examples based on the mean token-level log-probability of the translations the teacher generated. Figure 12 demonstrates that the low confidence examples stand out for all metrics: none of the models replicate their targets, all three models generate translations with a relatively low Comet-QE-22 score (akin to the lowest-quality subgroups discussed above), and all three models show an exceptionally low MSTTR. If we inspect the examples, they contain many misaligned examples from the corpus for which the teacher also does not generate an adequate translation (e.g., Example (9)), and a lot of cases where the teacher hallucinates in its translation (e.g., Example (10)). The student's improvement here thus does not stem from learning from the superior targets θ_T generated but more likely from deviating from θ_T. Yet, note that even if θ_S does not necessarily copy these hallucinations from its teacher, we did find that the students inspected in the main paper showed amplified hallucination capabilities, so low-quality teacher targets like this could have other unanticipated effects downstream.

(9) s_C Evergreen Terrace - Almost Home 22.
t_C Zum Ernst-Gettke-Haus, Hausnummer 68, siehe unten.
t_T Die alte Burganlage – heute noch 22 Hektar.
t_S Die Evergreen Terrace ist fast die Heimat 22.
t_B In: Evergreen Terrace – Fast Home 22.

(10) s_C Because the race is restricted to Canadian-bred horses, it is not eligible for grading, despite being one of Canada's most prestigious races Northern Dancer Turf Stakes, (. . .)
t_C Kammerherr, Land- und Obergerichtsrat Magnus Graf von Moltke, Ständedeputierter der Stadt Schleswig, Präsident der konstituierenden Ständeversammlung des Jahres 1836, (. . .)
t_T Stakes, (. . .)
t_S Depots of Canada'hintergrund für kanadische Pferde, ist es nicht berechtigt für das Einstufung, obwohl es eines der renommiertesten Rennen Northern Dancer Turf Stakes, (. . .)
t_B Stakes, (. . .)

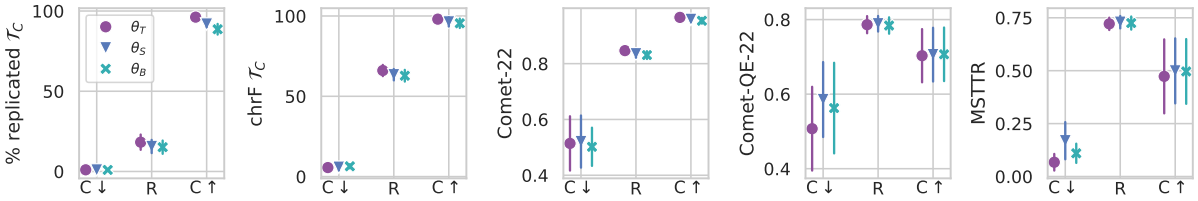


Figure 12: Evaluation metrics applied to the confidence subgroups, aggregated over language pairs with error bars indicating standard deviation over language pairs.

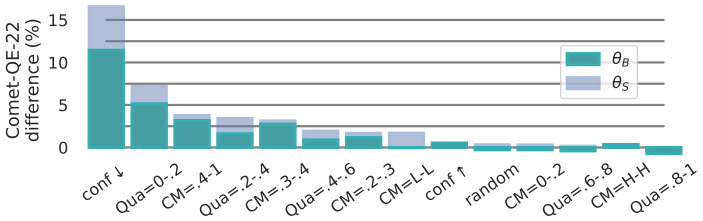


Figure 13: Relative increases comparing students and baselines to the teacher models, for the Comet-22-QE metric.