Some Assembly Required 🧩 Pointer-Decoded Seq2Seq Models for German Without a Native Seq2Seq Backbone

Anonymous ACL submission

Abstract

Pointer-decoded Seq2Seq transformers remain a strong recipe for aspect-based sentiment analysis (ABSA), but no native German Seq2Seq backbone exists, and current pretraining trends favor decoder-only architectures. Prior German work in this direction has therefore relied on multilingual mBART, leaving assembly from independently pretrained encoder and decoder LMs as the practical route to a native model. Reports on such assemblies disagree on the right recipe, with no agreed mechanism for why some pairings work. We study assembly quality on German ABSA, where structured pointer outputs make failures easier to measure through both tuple F1 and grammar validity. Pairing three native German encoder and decoder backbones in all 3×3 combinations shows that under standard cross-attention encoder capability dominates and direction-mismatched cells underperform; T5Gemma2-style merged attention repairs those cells and the dominance flips to the decoder. We isolate in a reductive ablation what carries this performance in our case: not the weight reuse credited by T5Gemma2, but the fused residual update. A further sweep over the parameter budget distribution places it on the encoder side: a grid-best 1B encoder, 120M decoder assembly reaches near parity with mBART-large-50 on German without pretraining a native German Seq2Seq. The same pattern generalizes to a homogeneous English LLaMa assembly.

1 Introduction

Encoder-decoder Seq2Seq transformers remain the standard backbone for tasks that consume a full input before producing an output, like machine translation (Tiedemann et al., 2024) or pointer-network structured prediction, where the decoder selects *positions* in the encoded input to emit spans and labels (Yan et al., 2021). Pointer decoding follows a fixed output grammar instead of generating a distribution over a subword vocabulary, so these comparably small models stay competitive with much larger decoder-only LLMs: e.g., BARTABSA++ (Pfister et al., 2025a), a pointer-decoded Seq2Seq for aspect-based sentiment analysis (ABSA), outperforms a GPT-4o finetune on the same task. Yet modern pretraining rarely produces encoder-decoder backbones from scratch (Zhang et al., 2025a), and many languages including German lack native Seq2Seq models. One alternative to pretraining from scratch is to assemble such a model: Rothe et al. (2020) build Vaswani et al.'s (2017) Seq2Seq from independently pretrained encoder/decoder LMs by inserting a cross-attention and fine-tuning.

But we find that these assemblies do not always train: some pairings collapse to near-zero F1 with structurally invalid output that further training does not recover, while others remain valid but extract poorly. Reports in the literature disagree on when Seq2Seq assembly works how well. Rothe et al. (2020) achieve their best results when both sides initialize from an encoder, leaving assemblies that use decoders throughout or decoders as encoders largely unprobed, which were later found to hit a performance ceiling that scaling does not fix (Pfister et al., 2025a). T5Gemma (Zhang et al., 2025a) trains its two-decoder homogeneous assembly through extensive further-pretraining, while T5Gemma2 (Zhang et al., 2025b) fuses the two attention blocks of the same configuration into a single "*merged attention*" that reuses the decoder's pretrained projections, motivated by the functional similarity of self- and cross-attention.

We study what governs assembly success on German aspect-based sentiment analysis (ABSA), where the structured pointer output makes failures easier to measure: every prediction has both a tuple F1 score and a grammar-validity status. Pointer-decoded Seq2Seq has been (close to) the SotA recipe for ABSA (Yan et al., 2021; Wunderle et al., 2023; Pfister et al., 2025a), yet no na-

tive German pretrained Seq2Seq exists (Pfister and Hotho, 2024), so Wunderle et al.'s (2023) prior German work relied on a multilingual BART as backbone. However, strong native German encoder LMs (ModernGBERT) and decoder LMs (LLäMmlein) are available in multiple sizes (Wunderle et al., 2026; Pfister et al., 2025b), opening assembly as a viable route to a native German Seq2Seq. A successful assembly is a usable artifact by itself: to our knowledge, the first pointer-decoded Seq2Seq for German ABSA built entirely from native components.

We organize this paper around four questions:

RQ1: Which encoder/decoder combinations train into a working pointer-decoded Seq2Seq for German ABSA, and which work best?

RQ2: What makes an assembled Seq2Seq train, and which feature of the merged attention mechanism is load-bearing?

RQ3: Where does the parameter budget belong, in the encoder or the decoder?

RQ4: How do assemblies compare against pretrained Seq2Seq models like mBART, and does the recipe transfer to English?

Our contributions are:

(a) We identify the fused residual update as the load-bearing feature of merged attention, refining T5Gemma2's weight-reuse account.

(b) We build the first working assembled pointer-decoded Seq2Seq for German ABSA from native German components.

(c) We establish a scaling rule for pointer-decoded structured prediction: scale the encoder, shrink the decoder.

(d) We replicate the same pattern on a homogeneous English LLaMa+LLaMa assembly, and release a reimplementation of T5Gemma2 merged attention that extends to heterogeneous configurations across model families.

Note: Throughout the paper, we highlight interesting findings and insights we gained during the process in little boxes like this one.

2 Related Work

Pointer networks (Vinyals et al., 2015) emit pointers to indices from the input rather than tokens from a vocabulary, a natural fit for span-based tasks. BARTABSA (Yan et al., 2021) therefore applies this to aspect-based sentiment analysis (ABSA, example in section 3.1), on top of which BARTABSA++ (Pfister et al., 2025a) introduces three architectural enhancements. BARTABSA++ also report initial experiments combining this recipe with assembled Seq2Seq models (Rothe et al., 2020), but only with older English RoBERTa, BERT, and GPT2 backbones. Further, Wunderle et al. (2023) apply the pointer-network recipe with a multilingual mBART backbone on a German ABSA task (Wojatzki et al., 2017). Some sentiment-tuple datasets generalize the ABSA triplet with holders, yielding (holder, target, expression, polarity) tuples (Barnes et al., 2021); SemEval-2022 Task 10 (Barnes et al., 2022) provides the opener_en dataset we use in section 5.6.

Rothe et al. (2020) warmstart both sides of a Seq2Seq from independently pretrained encoder and decoder LMs. A contemporaneous line of work on neural machine translation leverages a pretrained encoder-only LM (BERT) on the source side, either by warmstarting the encoder (Imamura and Sumita, 2019) or by fusing BERT representations into the NMT model via attention (Zhu et al., 2020). T5Gemma (Zhang et al., 2025a) pursues a related goal from the opposite direction, warmstarting both sides of its Seq2Seq from a single pretrained decoder-only Gemma checkpoint and continuing pretraining at length; T5Gemma2 (Zhang et al., 2025b) introduces *merged attention* on the same configuration, motivated by the functional similarity of self- and cross-attention. Merged attention is in spirit related to multi-query (Shazeer, 2019) and grouped-query attention (Ainslie et al., 2023), although those share keys and values across heads of one attention block rather than between self- and cross-attention. The pretrained multilingual Seq2Seq backbones available are mBART (Liu et al., 2020; Tang et al., 2020) and mT5 (Xue et al., 2021), and serve as our baselines (section 5.5).

Assembled Seq2Seq for German is tractable because both sides have strong, natively pretrained options in multiple sizes: LLäMmlein (Pfister et al., 2025b), ModernGBERT, and LLäMmlein2Vec (Wunderle et al., 2026).

3 Methodology

3.1 Pointer-Decoded Seq2Seq for Aspect-Based Sentiment Analysis

We build on BARTABSA++ (Pfister et al., 2025a), a pointer network on top of a Seq2Seq backbone that emits sentiment triplets under a *fixed output grammar*. The encoder reads the input once, the

Figure 1: An example sentence for ABSA with aspect, opinion, sentiment, and word indices annotated. Reproduced from Pfister et al. (2025a), adapted from Yan et al. (2021).

decoder then autoregressively emits *pointers*: at decoding step t its hidden state h_t^d is scored against the n encoder token representations H^e and a set of special-token embeddings S (EOS plus task labels such as grammar delimiters and sentiment classes),

$$P_t = \mathrm{softmax}\big([H^e; S]\, h_t^d\big). \tag{1}$$

P_t is the model's entire output space: the decoder never generates over its subword vocabulary. BARTABSA++ extends the original BARTABSA (Yan et al., 2021) pointer head with three modifications to the inputs of eq. (1): feature normalization, parametrized gating, and an additional cross-attention block before the pointer head. We adopt all three throughout but do not analyze them further in this work, and defer to Appendix C.

Output grammar. Each sentiment triplet (like "Drivers" / "okay" / Positive in fig. 1) is serialized as a tuple of the form

<ASP_BEGIN> a_s a_e <OPN_BEGIN> o_s o_e <SENT>

where each span pointer is an index into the encoder positions, the holder block is optional, and <SENT> stands for one of <POS>, <NEU>, <NEG>. The full target sequence is a run of such tuples closed by EOS. For the SCARE sentence *"Ohne dieses neue update kann ich gar nichts mehr machen !"* (word indices 0–10) with gold aspect "update" (word 3), opinion "gar nichts mehr machen" (words 6–9), and negative polarity, the word-level rendering of the decoded pointer sequence is

<ASP_BEGIN> 3 3 <OPN_BEGIN> 6 9 <NEG> EOS.

One visible failure mode is *grammar collapse*, where the model breaks this grammar and emits structurally invalid output (catalogued in Appendix J). Appendix B details how an emitted pointer is turned back into a decoder input embedding for the next decoding step.

3.2 Assembling a Seq2Seq from Pretrained Encoder and Decoder LMs

In this work, we assemble Seq2Seq models from three native German backbones: ModernGBERT (an encoder), LLäMmlein (a decoder), and LLäMmlein2Vec (a decoder converted into an encoder using LLM2Vec (BehnamGhader et al., 2024)). LLäMmlein2Vec shares its backbone weights with LLäMmlein but is adapted toward the encoder role, giving us a controlled third option that sits between the dedicated encoder and the dedicated decoder used as encoder. This requires extending Rothe et al.'s (2020) recipe, originally implemented only for (Ro)BERT(a) and GPT2, to more modern architectures such as LLaMa and ModernBERT.

To assemble a Seq2Seq from existing encoder and decoder models, we insert a cross-attention block into the decoder stack, and initialize the encoder and decoder weights from their respective checkpoints. To support arbitrary pairings, each checkpoint is adapted to its assigned side regardless of pretraining direction: encoder-side models run bidirectionally, and decoder-side models run causally with added cross-attention to the encoder states. This yields both *homogeneous* configurations (initializing from the same backbone on both sides) and *heterogeneous* ones (using a dedicated encoder and a dedicated decoder). Each has natural appeal: homogeneous for symmetry and (intuitively) "compatible" encoder/decoder representations from a single backbone, heterogeneous for letting each side use a backbone pretrained for its assigned role. When the two hidden sizes differ, pointer scoring uses a learned linear projection from the encoder representation space (H^e and S) into the decoder hidden size. These projection parameters are trained jointly with the rest of the model.

A further degree of freedom assembly opens up is the per-side layer count: independently of backbone choice and scale, we are free to truncate either side and do not need to use its full pretrained depth. Since the pointer head selects over encoder positions rather than decoding over a subword vocabulary, we hypothesize that intermediate encoder representations may already suffice for the decoder to point against. We therefore treat the per-side layer count as a design axis alongside backbone choice and cross-attention type, sweeping both as truncations to the first N pretrained layers (the *layer-cutoff sweep*), and write $\mathrm{e}N_e\mathrm{d}N_d$ for the cell with an encoder truncated to N_e layers and a decoder to N_d (e.g. e11d11).

3.3 Cross-Attention: Separate vs. Merged

The final design axis assembly exposes is the cross-attention block itself: the only freshly inserted component, and the one that bridges the encoder and decoder states inside each decoder layer. We compare two variants of how to insert it. A standard Seq2Seq decoder layer has two attention blocks that write into the residual stream in sequence: self-attention over the decoder states, then cross-attention to the encoder states, each with its own input norm and residual addition, following Vaswani et al. (2017). We call this the *separate* variant, the default in Rothe et al.'s (2020) assembly recipe, where the encoder-decoder attention is added as a second attention block to each pretrained decoder layer with freshly initialized projections.

T5Gemma2 (Zhang et al., 2025b) contracts these two blocks into one *merged attention* with a single fused residual write (fig. 4). Only the decoder states H^d supply the query vectors, while the concatenation $X = [H^d; H^e]$ of decoder and encoder states along the sequence axis supplies both, key and value vectors. The output of a single softmax over the joint logits and a single output projection is then written back to the residual stream:

$$\mathrm{softmax}\left(\frac{H^d W_Q (X W_K)^\top}{\sqrt{d_h}} + M\right) X W_V W_O, \quad (2)$$

where the mask M is causal within the self block and padding-only over the cross block. All four projections W_Q, W_K, W_V, W_O are reused unchanged from the decoder's pretrained self-attention, so the merged block introduces no freshly initialized attention projection. T5Gemma2 attributes the merged advantage to this weight sharing: for homogeneous backbones, the decoder's pretrained self-attention projections are claimed to transfer directly to the cross part. The decoder's pretrained pre-attention pathway also carries over, with the decoder's rotary positional encoding (RoPE) on the shared query and self-attention keys, which we revisit in the ablation of section 5.3.

We treat *separate* and *merged* as two design options for both homogeneous and heterogeneous configurations, and compare them empirically in section 5.

4 Experimental Setup

4.1 Used Datasets

All our datasets annotate sentiment as triplet structures with a target or aspect, an opinion, and a polarity class (fig. 1). For German we use SCARE (Sänger et al., 2016) (app-store reviews) and USAGE (Klinger and Cimiano, 2014) (product reviews), both with three-way polarity. The English replication (section 5.6) adds opener_en (Barnes et al., 2022), an English structured sentiment corpus with two-way polarity, and the two refined SemEval ABSA triplet sets (Xu et al., 2020) 14lap and 14res, the same English ABSA benchmark used by BARTABSA(++) (Yan et al., 2021; Pfister et al., 2025a). All five share the aspect/opinion/polarity triplet structure, so the pointer grammar and the tuple-F1 metric carry over unchanged across languages; opener_en additionally annotates opinion holders, which we emit as the optional third span block of the grammar (section 3.1) and which the tuple metric scores alongside aspect and opinion. Dataset details are listed in Appendix E.

4.2 Backbone Models

We draw encoders and decoders from three native German backbones, each available in two sizes: LLäMmlein (120M, 1B), a native decoder (Pfister et al., 2025b); ModernGBERT (134M, 1B), a native encoder; and LLäMmlein2Vec (120M, 1B), the LLäMmlein decoder repurposed as an encoder (the latter two from Wunderle et al. (2026)). Any of the three can sit on either side of the assembly.[1]

As pretrained Seq2Seq baselines we compare against the multilingual mBART-large-50 (610M) (Liu et al., 2020; Tang et al., 2020) and mT5-small (300M) (Xue et al., 2021).

4.3 Training and Evaluation

Since our aim is to conceptually establish whether/when these assemblies train and where the parameter budget belongs, not to extract maximal benchmark performance, we do not further tune hyperparameters per configuration. We inherit the BARTABSA++ working hyperparameters throughout and build on its implementation (section 3.1), which exposes the three enhancements as toggleable components. The full hyperparameter list is in Appendix D.

The metric is a weighted-overlap tuple F1 (Yan et al., 2021; Barnes et al., 2021): a predicted tuple

[1] We name a configuration encoder2decoder, abbreviating ModernGBERT, LLäMmlein, and LLäMmlein2Vec as mg, ll, and lv: e.g. mg2lv is a ModernGBERT encoder with a LLäMmlein2Vec decoder, and ll2ll is LLäMmlein on both sides. The small backbone is the default; a 1B side is marked inline, as in mg_{1B}2ll (a 1B ModernGBERT encoder with a small LLäMmlein decoder).

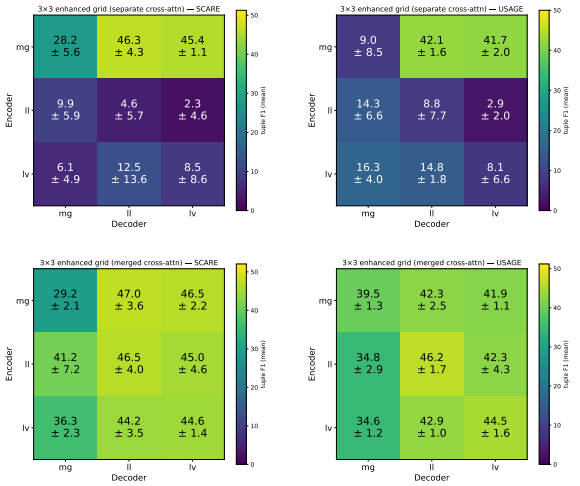

Figure 2: 3×3 encoder (rows) × decoder (columns) tuple F1 grid under separate (top) and merged (bottom) cross-attention, on SCARE (left) and USAGE (right). 5-seed mean ± std per cell.

counts as a true positive when its polarity matches the gold and each annotated span shares at least one token with the corresponding gold span, with the true-positive then weighted by the average per-span token-overlap fraction. We call a run *collapsed* when its test tuple F1 falls below 5. In this pointer-decoded setting, collapse can additionally be audited against grammar validity (section 3.1), so we report the invalid-output rate alongside tuple F1 where it is informative.

5 Results and Insights

5.1 Which Combinations Can Be Assembled?

Three native German backbones can occupy either side of the assembly, giving nine candidate pairings: which of them actually train into a working pointer-decoded Seq2Seq, and what predicts the failures? We screen all of them in a 3×3 grid: each of the three German backbones (ModernGBERT, LLäMmlein, LLäMmlein2Vec) as the encoder, crossed with each as the decoder, at the small parameter size and with *separate* cross-attention (RQ1), 5 seeds per cell.

We find the grid resolves into a clear pattern that runs along the *encoder*, not the decoder (fig. 2, top two plots). The three backbones order by the degree to which each was pretrained for the encoder role. ModernGBERT, the only native encoder, produces the strongest cells with either LLäMmlein-family decoder (mg2ll SCARE 46.26 ± 4.31, USAGE 42.05 ± 1.62; mg2lv SCARE 45.38 ± 1.07, USAGE 41.69 ± 1.96). LLäMmlein forced to encode collapses every cell below 15 F1, while LLäMmlein2Vec, the LLäMmlein decoder adapted into an encoder, interpolates but still trails the ModernGBERT row. The encoder-row ordering thus reads "encoder-pretrained > adapted > decoder-as-encoder" rather than tracking backbone identity alone. The same direction mismatch appears on the decoder side: with a working encoder, LLäMmlein and LLäMmlein2Vec decoders are interchangeable, but ModernGBERT forced to decode drops its own row sharply, with mg2mg trailing mg2ll and mg2lv by ∼18 F1 on SCARE (28.2 vs 46.3, 45.4) and ∼33 F1 on USAGE (9.0 vs 42.1, 41.7). This mirrors Weller et al. (2026), who report that decoder-from-encoder conversions underperform native decoders on generative tasks: repurposing a model away from the task its pretraining was shaped for is the harder direction.

Across the weak cells, we examine the three same-backbone cells in detail as an illustrative slice (Appendix J) and find two distinct failure modes. ll2ll and lv2lv break their own output grammar, emitting structurally invalid sequences at high rates (invalid rate 0.31–0.65 for lv2lv). mg2mg instead stays mostly valid but extracts poorly and unstably. These are the failure modes section 5.2 sets out to repair.

> **Pretraining direction dominates**: A model is worst at the direction opposite its pretraining, on whichever side it sits. This refines Rothe et al.'s (2020) best configuration of two encoders (BERT+BERT): our analogous mg2mg cell instead underperforms, since the ModernGBERT-encoder is forced to decode.

5.2 Rescuing Direction-Mismatched Failures with Merged Attention (RQ2)

T5Gemma2 (Zhang et al., 2025b) motivates merged attention on a homogeneous two-decoder assembly, but the cells that fail in our grid (section 5.1) are a superset: every LLäMmlein- and LLäMmlein2Vec-encoder cell plus mg2mg on the ModernGBERT row, spanning both homogeneous (ll2ll, lv2lv, mg2mg) and heterogeneous off-diagonal cells. Does the rescue extend to all of them? We swap separate for merged attention (section 3.3) and rerun the full 3×3 grid (fig. 2, bottom row).

Merged attention **repairs every direction-mismatched** cell on a single hyperparameter pass. The LLäMmlein- and LLäMmlein2Vec-encoder rows that bottomed out at 2–17 F1 under sepa-

rate jump into the 34–46 F1 band; the largest single-cell gap is ll2ll (4.57 → 46.45 on SCARE, +41.88; 8.81 → 46.19 on USAGE, +37.38), and no previously collapsed run stays below the 5 F1 threshold under merged. ModernGBERT-encoder cells, which already trained under separate, hold steady, except for mg2mg USAGE which itself jumps from 8.97 to 39.49. Seed variance compresses across the grid: median std drops from > 5 F1 under separate to ≈ 2.5 F1 under merged, and most cells now have std ≤ 5. On the rescued LLäMmlein- and LLäMmlein2Vec-encoder rows, the same-family cells now narrowly top their row on both datasets (ll2ll 46.45/46.19, lv2lv 44.65/44.45 on SCARE/USAGE), though many within-row gaps here sit inside seed std.

Layer-cutoff sweeps on lv2lv and ll2ll confirm the rescue is grid-wide rather than a single-cell artefact: every depth configuration trains under merged, while the corresponding pre-rescue grid remains broadly underperforming and unstable under separate (fig. 5 in Appendix G).

Reading the top two panels against the bottom two panels of fig. 2 surfaces a symmetric importance story. Under separate cross-attention (top), encoder identity dominates: swapping the encoder shifts the row mean by ~ 30 F1, swapping the decoder barely moves it. Under merged (bottom), the dominance flips to the decoder columns: encoder row means contract into a tighter 37–46 F1 band, while ModernGBERT-as-decoder becomes the bottleneck (column means on USAGE: ≈ 36 F1 for ModernGBERT decoders, ≈ 44 for LLäMmlein, ≈ 43 for LLäMmlein2Vec). Separate cross-attention *requires* a strong encoder; merged cross-attention *requires* a strong decoder.

🧩 **Merged repairs direction-mismatched failures**: T5Gemma2-style merged attention rescues every underperforming cell.

On heterogeneous configs that already train under separate, merged is mostly a no-op within seed noise (controlled $mg_{1B}2ll$ 5-seed comparison: SCARE $\Delta = +1.32$, USAGE $\Delta = -1.02$, both inside seed std ≈ 7), so merged is a free upgrade in our assemblies: it rescues the direction-mismatched cells and matches separate on the rest.

5.3 What Is Load-Bearing in Merged Attention? (RQ2)

Merged attention rescues every direction-mismatched cell, but which of its structural differences from separate is actually carrying the rescue? The mechanism question is clearest where separate and merged differ most, so we run the ablation on ll2ll, a clean collapse case (separate near 5 F1, merged near 45).

T5Gemma2 (Zhang et al., 2025b) attributes the merged advantage to weight reuse: the decoder's pretrained Q, K, V, and O projections transfer cleanly to the cross part when both sides share a backbone. We test this by starting from the working merged baseline R0 (the five features of section 3.3: single attention block, single softmax over concatenated keys, single residual write, all four projections shared with the decoder's self-attention, RoPE on the shared query; 44.90 ± 3.91 F1, 0/10 models collapse) and stripping one feature at a time toward Rref, the standard separate decoder (two softmaxes, two residual writes, reinitialized cross projections, no cross-Q RoPE, dedicated cross-attention norm; 8.17 ± 6.65 F1, 4/10 models collapse). R0 and Rref differ in six structural features. We evaluate each single-feature reduction on ll2ll with 5 seeds on both German datasets (10 runs per variant); a *collapse* is a run below 5 F1.

R1, split softmax (two independent softmaxes over self- and cross-keys): 44.76 ± 2.98 F1, 0 collapses, indistinguishable from R0; the shared softmax is not the mechanism. **R2, fresh cross-Q** (reinitialized cross-attention query): 44.26 ± 5.38, 0 collapses, within seed noise. **R4, fresh cross-O** (reinitialized cross-attention output): 45.76 ± 3.32, 0 collapses, if anything slightly above R0. **R5, extra cross-attn LN** (dedicated layer norm before the cross-attention query, alongside the shared input norm): 42.68 ± 3.02, 0 collapses. Together, R2, R4, and R5 indicate that weight reuse is not what carries the rescue: any cross-attention projection can be reinitialized and a separate cross-norm added, and merged still trains with no noticeable drop in performance. **R3, no cross-Q RoPE** (rotary embeddings removed from the cross-attention query): 31.87 ± 14.57, 1/10 collapses, a high-variance partial contributor that still trains on the remaining seeds. **R6, split residual update** (the single attention block split into self-attention followed by cross-attention, each with its own residual write): 7.59 ± 8.20, 4/10 below threshold, matching Rref's collapse rate within seed variance while retaining merged's shared $Q/K/V/O$ projections and shared input norm. This isolates the split residual write as the load-bearing change.

R6 is not strictly atomic: running the two blocks

separately entails both a split softmax (R1) and no RoPE on the cross-Q (R3), since the cross block reuses the shared q_{proj} without RoPE. Neither alone collapses the model reliably, so the split residual write, possibly compounded by the entailed RoPE removal, carries the effect.

The R0 vs R6 contrast transports to mg2mg. On mg2mg R6 drops from 35.81 to 19.50 ($\Delta -16.3$, no seeds collapse but variance widens substantially). The fused residual update is therefore load-bearing across different backbones, not just on ll2ll. The R6 floor itself depends on the backbone, though, because each cell falls back to its own separate-cross-attention floor: R6 lands at $\approx$ 5 F1 on ll2ll and $\approx$ 18 F1 on mg2mg, mirroring their respective separate-baseline performances (ModernGBERT-encoder cells already partially train under separate, fig. 2).

> **The load-bearing feature**: Merged's advantage comes from its fused residual update and residual topology, not weight reuse.

5.4 Scaling: Where Should Parameters Go?

Two priors point to the encoder. Prior work reports that autoregressive NMT benefits from deep encoders and shallow decoders (Kasai et al., 2021), and that the same encoder-heavy preference holds for pointer-decoded ABSA (Pfister et al., 2025a); our pointer head produces a softmax over encoder positions rather than tokens (eq. (1)), so the decoder effectively "only" need to deliver one query vector against H^e. We test this on two axes. **Depth** truncates each side to a prefix of its pretrained layers at fixed scale. **Scale** swaps each side between its small and 1B checkpoint, with the layer-cutoff sweep re-run for each scale, so each scale comparison reports the grid-best cell rather than a fixed layer count. The following scaling sweeps all use merged attention, so all configurations are in its post-rescue version from section 5.2.

The **depth** grid is sharply asymmetric (fig. 3, ll2ll, SCARE). F1 drops sharply once the encoder is truncated below $\sim$10 layers, but is insensitive to decoder depth: at full encoder, the top five SCARE cells span decoder cutoffs 1 to 10 within a 3 F1 band (e12d6=49.4, e12d4=49.2, e11d10=48.2, e12d3=47.7, e11d1=46.5). The same shape holds on lv2lv (best cells e11d11 / e12d7) and on heterogeneous mg2ll (fig. 6 in Appendix H), so the asymmetry is not specific to one backbone pairing. Decoder depth is largely free to cut, encoder depth is not.

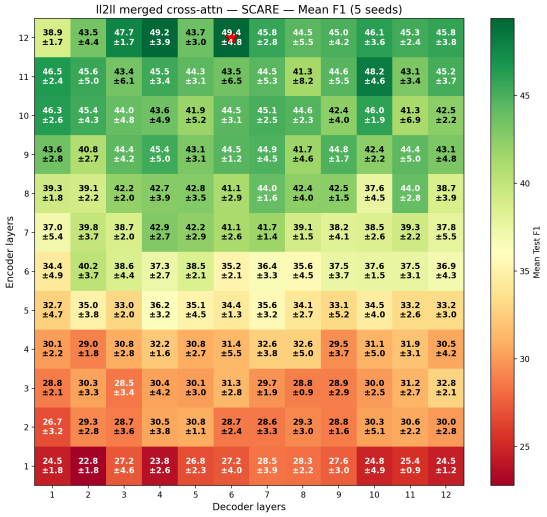

Figure 3: Encoder $\times$ decoder depth grid on ll2ll (post-rescue merged, SCARE, 5-seed mean tuple F1). The same asymmetry holds on lv2lv and on USAGE.

For the **scale** axis we report each cell as the grid-best over its own encoder/decoder layer-cutoff sweep (full per-combo numbers in table 4). We pair each 1B encoder with the small decoder that topped its row in the rescued 3$\times$3 of section 5.2: the LLäMmlein-120M decoder for the LLäMmlein and ModernGBERT encoder rows, and the LLäMmlein2Vec-120M decoder for the LLäMmlein2Vec row.

The three principled encoder-swap cells (1B encoder, row-best small decoder) read ll_{1B}2ll 54.29 / 52.47, lv_{1B}2lv 55.07 / 53.24, and mg_{1B}2ll 62.27 / 53.91 (SCARE / USAGE; table 4). Holding the encoder-side backbone fixed (LLäMmlein and its derivatives), retraining for the encoder role (LLäMmlein2Vec) buys $\sim$1 F1, and swapping to the native encoder (ModernGBERT) buys $\sim$7 F1 SCARE, reproducing the "trained for the role" ordering of section 5.1 at 1B encoder scale. Scaling the decoder instead is the wrong direction: ll2ll_{1B} 46.13 and mg2ll_{1B} 49.74 SCARE both lose against their small-decoder counterparts, and scaling both sides to 1B adds nothing on top of scaling only the encoder ($\text{ll}_{1B}\text{2ll}_{1B}$ 53.16 vs ll_{1B}2ll 54.29; $\text{mg}_{1B}\text{2ll}_{1B}$ 54.44 vs mg_{1B}2ll 62.27 on SCARE; all from table 4). The strongest cell in the whole grid is therefore mg_{1B}2ll at 62.3 SCARE (section 5.5), pairing the strongest 1B encoder with its row-best small decoder.

model	params	SCARE F1	USAGE F1
baselines			
mT5-small (enhanced)	300M	21.64 ± 2.96	18.42 ± 3.86
mBART-large-50 (enhanced)	610M	60.79 ± 1.76	**56.12** ± **2.19**
ours			
mg2ll (grid-best)	∼250M	50.39	45.77
mg_{1B}2ll controlled	∼1.1B	53.97 ± 6.45	45.51 ± 6.66
mg_{1B}2ll grid-best	∼1.1B	**62.27** ± **1.00**	53.91 ± 2.08

Table 1: Assembled mg_{1B}2ll reaches parity with mBART-large-50 at grid-best, and small mg2ll clears mT5-small at matched parameter count, both without a pretrained German Seq2Seq. *Controlled*: fixed full-layer merged config (5 seeds), single-config comparison. *Grid-best*: best per-dataset layer-cutoff cell (3 seeds), achievable upper bound.

Shrink the decoder, scale the encoder: Tuple F1 tracks encoder size and is largely insensitive to decoder size. Our strongest German assembly is mg_{1B}2ll, a 1B encoder with a 120M decoder.

The pointer wiring (eq. (1)) makes this concrete: the decoder produces one query vector per step against H^e, so its capacity saturates early, while the encoder representations those scores select over keep paying off with scale.

5.5 Comparison to Pretrained Seq2Seq Baselines (RQ4)

With a recipe in hand (merged attention, big encoder, small decoder), how close does the strongest assembled German Seq2Seq get to a multilingual one that was pretrained for the language? We compare mg_{1B}2ll against mBART-large-50 (610M) and mT5-small (300M) under the same training protocol (table 1), reading mBART two ways: a grid-best cell (achievable upper bound) and a fixed full-layer single configuration (since mBART itself is not tuned over layer cutoffs). Grid-best mg_{1B}2ll reaches mBART parity within seed noise (+1.48 SCARE, −2.21 USAGE), without pretraining a native German Seq2Seq. The controlled reading is more conservative: full-layer mg_{1B}2ll trails by ∼7 (SCARE) and ∼11 (USAGE) F1, with seed std several times mBART's. At matched parameter count, the small mg2ll (∼250M) clears mT5-small with ease (∼27–29 F1) on both German datasets.

5.6 Replication on English

Is the collapse-and-rescue pattern of section 5.2 a German backbones artefact, or a property of homogeneous direction-mismatched assemblies in general (RQ2, RQ4)? We replicate two LLaMa-family configurations on English (14lap, 14res, opener_en; section 4.1), 5 seeds per cell (full numbers in Appendix F). A heterogeneous ModernBERT-large (400M) + Llama-3.2-1B assembly (the closest English analogue to mg2ll, although decoder-heavy against the section 5.4 preference for parameters on the encoder side) trains under both cross-attention variants with $|\Delta| \leq 0.71$ F1 between merged and separate, matching the no-op behaviour found on the German mg2ll in section 5.2. A homogeneous Llama-3.2-1B on both sides (parallel to ll2ll) collapses under separate on all 15 seeds (5 × 3 datasets, mean F1 ≤ 3.5), and merged brings 14 of those 15 above the 5 F1 collapse threshold (per-dataset numbers in table 3). The collapse-and-rescue pattern of section 5.2 is therefore not specific to the German backbones, but a property of homogeneous direction-mismatched assemblies that merged attention repairs.

6 Conclusion

Competitive pointer-decoded Seq2Seq models for languages without a native Seq2Seq backbone can be assembled from independently pretrained encoder and decoder LMs. On a 3×3 grid of German backbones (RQ1), each side prefers a backbone pretrained for its role: the LLäMmlein decoder fails when forced to encode, ModernGBERT (a native encoder) is the weakest decoder, and LLäMmlein2Vec, the LLaMa adapted into an encoder, interpolates on either side. The direction-mismatched cells fail under standard cross-attention but are repaired by T5Gemma2-style merged attention, and the encoder-vs-decoder dominance flips between the two paths: separate needs a strong encoder, merged needs a LLaMa-family decoder. A reductive ablation on same-backbone stress tests (RQ2) localizes the rescue to the fused residual update, not the weight reuse credited by T5Gemma2.

Because the decoder emits pointers rather than generating over the subword vocabulary, its depth and width saturate early while encoder size keeps paying off, so the parameter budget belongs in the encoder (RQ3). The strongest grid best German assembly, mg_{1B}2ll, reaches rough parity with mBART-large-50 and the small mg2ll clears mT5-small at matched parameter count (RQ4), both without pretraining a German Seq2Seq. A homogeneous LLaMa+LLaMa English transfer reproduces the same collapse-and-rescue pattern, so our observations are not specific to the German backbones.

Limitations

Scope of the empirical claim Our findings are anchored to pointer-decoded structured sentiment analysis on five datasets (SCARE, USAGE in German; opener_en, 14lap, 14res in English), three German backbone families (ModernGBERT, LLäMmlein, LLäMmlein2Vec), and two English ones (ModernBERT-large, Llama-3.2-1B). The encoder-heavy scaling rule (section 5.4) and the parity claim against mBART (section 5.5) rest on the pointer-decoding wiring of eq. (1), where the decoder emits a single query per step against the encoder positions the encoder positions rather than generating over a subword vocabulary. We do not test whether the same scaling rule transports to standard subword-vocabulary generation tasks like machine translation, summarization, or open-ended generation, where the decoder side carries additional load. The collapse-and-rescue pattern (section 5.2) is verified on a homogeneous English LLaMa assembly in addition to the German configurations.

No further pretraining of the assembled Seq2Seq. Unlike T5Gemma and T5Gemma2 (Zhang et al., 2025a,b), which continue pretraining the assembled Seq2Seq on a denoising objective before downstream finetuning, our assemblies go from initialization straight to the ABSA pointer task. The mBART comparison in section 5.5 is therefore not strictly like-for-like: mBART is a pretrained Seq2Seq with denoising-trained in-layer cross-attention, while our assemblies insert fresh cross-attention and train it only through the downstream signal. The reading we draw, that a fairly cheap finetune of an assembly reaches rough parity with a pretrained multilingual Seq2Seq, is conditional on this setup, and further Seq2Seq pretraining might either widen or close the gap. A single attempt to warm the cross-attention through a generative pretraining stage (summarization on MLSUM, Appendix L) hurt downstream F1 across all six conditions tested, which suggests the choice of further-pretraining objective is itself non-trivial rather than a free improvement.

Grid-based readings. Several headline cells are read off a grid rather than a single training run: the rescued 3×3 (section 5.2) reports 5-seed means per cell, and the 1B-scale sweeps (table 4) use 3 seeds per cell. Reproducing the grid-best parity result in section 5.5 requires sweeping the encoder/decoder depth grid at 1B encoder scale; the controlled reading in table 1 is the corresponding single-configuration anchor.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. 2023. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4895–4901, Singapore. Association for Computational Linguistics.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402, Online. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong. Association for Computational Linguistics.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Roman Klinger and Philipp Cimiano. 2014. The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218, Reykjavik, Iceland. European Language Resources Association (ELRA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training

for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Jan Pfister and Andreas Hotho. 2024. SuperGLEBer: German language understanding evaluation benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico. Association for Computational Linguistics.

Jan Pfister, Tom Völker, Anton Vlasjuk, and Andreas Hotho. 2025a. BARTABSA++: Revisiting BARTABSA with decoder LLMs. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 115–128, Vienna, Austria. Association for Computational Linguistics.

Jan Pfister, Sebastian Wankerl, and Andreas Hotho. 2022. SenPoi at SemEval-2022 task 10: Point me to your opinion, SenPoi. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1313–1323, Seattle, United States. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025b. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Mario Sänger, Ulf Leser, Steffen Kemmerer, Peter Adolphs, and Roman Klinger. 2016. SCARE — the sentiment corpus of app reviews with fine-grained annotations in German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1114–1121, Portorož, Slovenia. European Language Resources Association (ELRA).

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *Preprint*, arXiv:1911.02150.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Preprint*, arXiv:2008.00401.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with opus-mt. *Language Resources and Evaluation*, 58(2):713–755.

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.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2026. Seq vs seq: An open suite of paired encoders and decoders. In *The Fourteenth International Conference on Learning Representations*.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. GermEval 2017: Shared Task on Aspect-based Sentiment in Social Media Customer Feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12, Berlin, Germany.

Julia Wunderle, Anton Ehrmanntraut, Jan Pfister, Fotis Jannidis, and Andreas Hotho. 2026. New encoders for german trained from scratch: Comparing moderngbert with converted llm2vec models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10424–10446, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Julia Wunderle, Jan Pfister, and Andreas Hotho. 2023. Pointer networks: A unified approach to extracting German opinions. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 127–138, Ingolstadt, Germany. Association for Computational Lingustics.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025a. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *Preprint*, arXiv:2504.06225.

Biao Zhang, Paul Suganthan, Gaël Liu, Ilya Philippov, Sahil Dua, Ben Hora, Kat Black, Gus Martins, Omar Sanseviero, Shreya Pathak, Cassidy Hardin, Francesco Visin, Jiageng Zhang, Kathleen Kenealy, Qin Yin, Xiaodan Song, Olivier Lacombe, Armand Joulin, Tris Warkentin, and Adam Roberts. 2025b. T5gemma 2: Seeing, reading, and understanding longer. *Preprint*, arXiv:2512.14856.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

A T5Gemma2 Architecture Overview

Referenced from sections 3.3 and 5.3.

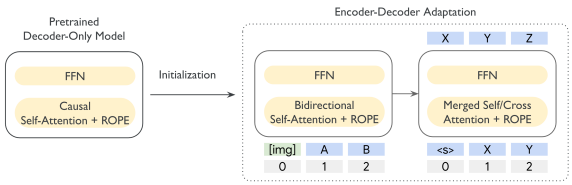

Figure 4: Overview of T5Gemma2. Reproduced from Zhang et al. (2025b): encoder and decoder parameters are initialized from a pretrained decoder-only model and then adapted with UL2. All word embeddings (in blue) are tied across encoder and decoder, and the decoder's self- and cross-attention sub-layers are merged into a single block.

B Pointer Decoder Inputs

At each decoding step the pointer head emits a single index y_t into the joint set $[H^e; S]$ of eq. (1): either a pointer to an encoder position (a span boundary) or to a special-token slot (a sentiment label, a grammar delimiter, or EOS). The fixed sequence in which these slots are expected to appear (a span tuple closed by a sentiment, the whole sequence closed by EOS) is the output grammar; a wrong slot at the wrong position is what Appendix J catalogues as a structurally invalid sequence. The decoder itself, however, consumes *token embeddings* rather than pointer indices, so each emitted y_t has to be turned back into an input-token id $\hat{y}_t$ before being fed back as the next decoder input. A more complete treatment of the pointer-decoded Seq2Seq is given by Pfister et al. (2025a); we summarize only what a reader needs to interpret section 3.1 and the invalid shapes in Appendix J.

Concretely, if y_t indexes the special-token block, $\hat{y}_t$ is that sentiment or boundary token; if y_t indexes an encoder position, $\hat{y}_t$ is the input subword token sitting at that position. From the decoder's perspective each step therefore looks like ingesting a regular subword sequence, and the pointer mechanism stays confined to the output side. Special tokens (`<ASP_BEGIN>`, etc.) are added to the tokenizer and their embeddings initialized as the mean of their constituent sub-tokens.

Training uses teacher forcing with cross-entropy on the pointer logits P_t.

C BARTABSA++ Enhancements We Inherit

The three modifications BARTABSA++ (Pfister et al., 2025a) makes to the inputs of eq. (1), referenced from section 3.1, are:

(i) **feature normalization**, L2-normalizing H^e and S before the pointer score (turning it into a cosine similarity) and applying an RMSNorm to the decoder output;

(ii) **parametrized gating** of the encoder representations: the original BARTABSA blends the MLP-processed encoder output $\hat{H}^e$ with the raw input embeddings E^e through a fixed scalar α, $\overline{H}^e = \alpha\hat{H}^e + (1-\alpha)E^e$; BARTABSA++ replaces α with a learned, input-dependent sigmoid gate (and adds an analogous decoder-side gate over H^d); and

(iii) an **additional cross-attention block** placed between the decoder stack and the pointer head, querying the decoder hidden states H^d against the encoder representations $[H^e; S]$, whose projection weights are initialized from the pretrained final-layer cross-attention when one exists (e.g. in BART, mBART, mT5).

We adopt all three throughout and do not analyze them further in this work.

D Hyperparameters

All runs use the BARTABSA++ (Pfister et al., 2025a) defaults unchanged: AdamW at a learning rate of 5×10^{-5} with weight decay 0.01, batch size 16, gradient clipping to norm 5.0, a linear warmup-only learning-rate schedule, dropout 0.1, and bf16 mixed precision. We train for up to 50 epochs and keep the checkpoint with the best dev tuple F1. Nothing is tuned per configuration. We run our experiment on L40 and L40s NVIDIA GPUs, each experiment takes between 30 minutes and 3 hours depending on the backbone size.

E Dataset Statistics

Table 2 lists sentence and triplet counts per split for the five datasets used in section 4.1. SCARE and USAGE (German) use three-way polarity; opener_en uses two-way polarity; 14lap and 14res use the three-way ABSA polarity inventory. opener_en additionally annotates opinion holders, which we emit as the optional third span block of the grammar (section 3.1).

	sentences			triplets		
dataset	train	dev	test	train	dev	test
SCARE	1100	140	133	1415	170	166
USAGE	1827	259	238	2970	405	395
opener_en	1744	249	499	2884	400	865
14lap	906	219	328	1460	346	543
14res	1266	310	492	2338	577	994

Table 2: Sentence and triplet counts per split for the five datasets.

One example per dataset. To illustrate the shared triplet structure, one annotated example per dataset (word indices of respective span in parentheses, polarity in braces):

- **SCARE**, *"Ohne dieses neue update kann ich gar nichts mehr machen!"*: aspect *update* (3), opinion *gar nichts mehr machen* (6–9), {NEG}.
- **USAGE**, *"Super Waschmaschine"*: aspect *Waschmaschine* (1), opinion *Super* (0), {POS}.
- **opener_en**, *"They were there just to keep us happy."*: aspect *They* (0), opinion *to keep … happy* (4–5, 7), holder *us* (6), {POS}.
- **14lap**, *"Boot time is super fast, … "*: aspect *Boot time* (0–1), opinion *fast* (4), {POS}.
- **14res**, *"The bread is top notch as well."*: aspect *bread* (1), opinion *top notch* (3–4), {POS}.

opener_en carries the optional holder block of the grammar in section 3.1; the four other datasets do not annotate holders.

F English Transfer: Per-Dataset Numbers

Table 3 gives the per-dataset tuple F1 underlying the summary in section 5.6, with monolingual BART-base (Lewis et al., 2020) and multilingual mBART-large-50 (Tang et al., 2020; Pfister et al., 2022) as references. mBART matches the native BART-base on both SemEval splits (14res 77.52 vs 76.29; 14lap 65.99 vs 66.56) and is slightly above on opener_en (52.47 vs 50.49), so the multilingual reference is the appropriate baseline to compare the assemblies against on English, as it is on German. Heterogeneous ModernBERT+LLaMa trails the multilingual reference by roughly 13 to 15 F1, and the merged homogeneous LLaMa+LLaMa transfer trails by roughly 7 to 19 F1.

config (cross-attn)	14lap	14res	opener_en
native BARTABSA ++ references			
BART-base	66.56 ± 1.36	76.29 ± 1.97	50.49 ± 1.94
mBART-large-50	65.99 ± 2.03	77.52 ± 1.01	52.47 ± 2.19
heterogeneous init			
ModernBERT+LLaMa (separate)	51.86 ± 2.12	64.71 ± 1.52	38.75 ± 1.99
ModernBERT+LLaMa (merged)	51.24 ± 3.28	64.66 ± 2.72	39.46 ± 1.28
homogeneous within-family init			
Llama+Llama (separate)	2.45 ± 2.44	3.34 ± 3.15	1.05 ± 1.02
Llama+Llama (merged)	$\mathbf{47.46 \pm 25.50}$	$\mathbf{70.06 \pm 5.54}$	$\mathbf{35.65 \pm 16.14}$

Table 3: English transfer (section 5.6), 5 seeds per cell. Heterogeneous ModernBERT+LLaMa (parallel to mg2ll): merged ≈ separate, no failure mode to repair. Homogeneous LLaMa+LLaMa (parallel to ll2ll): separate collapses on 15/15 seeds (5 per dataset × 3 datasets), merged brings 14/15 above the collapse threshold. BART-base (Lewis et al., 2020) is the monolingual English reference; mBART-large-50 is the multilingual reference (parallel to table 1 on German) and matches BART-base on all three datasets.

G ll2ll Depth Grid Before and After the Merged-Attention Rescue

Figure 5 compares the ll2ll encoder/decoder depth grid before (top row) and after (bottom row) the merged-attention rescue of section 5.2, on SCARE (left column) and USAGE (right column). Pre-rescue cells show 5-seed mean tuple F1, std, and the fraction of seeds above the 5 F1 collapse threshold (*%ok*); post-rescue cells show mean and std, since %ok is uniformly 100 across the grid. Under separate cross-attention, mean F1 stays low across the grid with %ok rarely above 80, often 20–40: the underperformance is grid-wide, not a single-cell artefact, so no depth configuration would have recovered the assembly to the merged-attention band. Once merged attention is enabled, every cell trains,

the encoder-heavy asymmetry surfaces, and the best cells cluster at deep encoders with moderate decoder cutoffs (the SCARE post-rescue panel here is the same as fig. 3).

H Layer-Cutoff Sweep on a Heterogeneous Config

Figure 6 shows the merged depth grid for the heterogeneous mg2ll on both German datasets, complementing the homogeneous ll2ll grid in fig. 3. The same encoder-heavy asymmetry obtains: F1 drops sharply once the encoder is truncated below $\sim$10 layers, and is largely flat across decoder depth, so the pattern is not specific to the homogeneous configuration shown in the main paper.

I Per-Combo Grid-Best at 1B Scale

Table 4 lists the grid-best cell of each 1B-scale layer-cutoff sweep used in section 5.4, on SCARE and USAGE. The two blocks are the LLäMmlein-family grid (six combos crossing {120M, 1B} encoders and decoders, plus the cross-family ModernGBERT encoder rows) and the LLäMmlein2Vec-family grid (six combos with at least one LLäMmlein2Vec side at 1B). All cells use merged cross-attention with hidden-size projection where the encoder and decoder hidden dimensions differ (section 3.3).

J Structurally Invalid Outputs Under Separate Cross-Attention

A working ll2ll or lv2lv run under merged attention reproduces the grammar of section 3.1, while the collapsed separate-cross-attention runs do not. Inspecting ll2ll checkpoints from collapsed separate-cross-attention runs and from a structural-loss upweighting follow-up experiment (the cross-entropy term on the structural pointer targets, i.e. sentiment and grammar-delimiter slots, is multiplied by a weight w to push the model into emitting the scaffold even when the content pointers collapse) shows three recurring failure shapes, against the grammar of section 3.1:

- **Truncated tuple.** The decoder emits one opening class token and one span pointer, then jumps straight to a sentiment token and EOS, dropping the rest of the tuple: e.g. <OPN_BEGIN> i <NEG>. This is the dominant failure shape on ll2ll once the run has collapsed: every test prediction settles into the same three-token pattern by epoch ten and persists through training. Validation fails because the span body has length one.
- **Class-token-only collapse.** The decoder emits a single class token (typically <OPN_BEGIN>) and immediately EOSs, producing a length-one output (under structural-loss weighting factor at w=10). Roughly half of all test predictions reduce to the literal sequence [<OPN_BEGIN>] in this regime.
- **Structural skeleton, empty content.** The decoder produces class delimiters but no usable content pointers in between, e.g. <OPN_BEGIN> i j <HOL_BEGIN>, then stops before producing a usable tuple (structural-loss weighting factor at w=100). Upweighting the structural targets in the loss can force the delimiters back in, but the content pointers do not recover.

All three shapes share the same end state: the output does not decode into a usable sentiment tuple. The validator rejects malformed tuples, for example when they lack <OPN_BEGIN> or contain an empty or odd-length span body, and sequences that stop before a tuple closes simply yield no triplet. In either case, the run gets no true positive credit on that example. On lv2lv the per-seed invalid rate ranges from 0.31 to 0.65, but we recorded only aggregate invalid rates for that cell. mg2mg is the qualitative exception: its outputs typically pass the validator and decode as well-formed tuples, but the content pointers extract poorly and the cell is unstable across seeds (fig. 2, same-backbone ModernGBERT cell: SCARE 28.17 ± 5.57, USAGE 8.97 ± 8.50). Once merged attention is enabled the recorded invalid rates fall to roughly 0 to 0.10 on the homogeneous cells we audited (section 5.2) and these shapes effectively disappear from the test predictions.

K Remaining Error Type Analysis

Tuple F1 (section 4.3) credits each prediction by its per-span token-overlap with the gold conditional on polarity match, but it is still a single number that conflates several qualitatively different failure modes. To see which of these failure modes the assembled models actually drive their remaining errors with, we train five representative configurations to convergence at a single fixed seed, generate

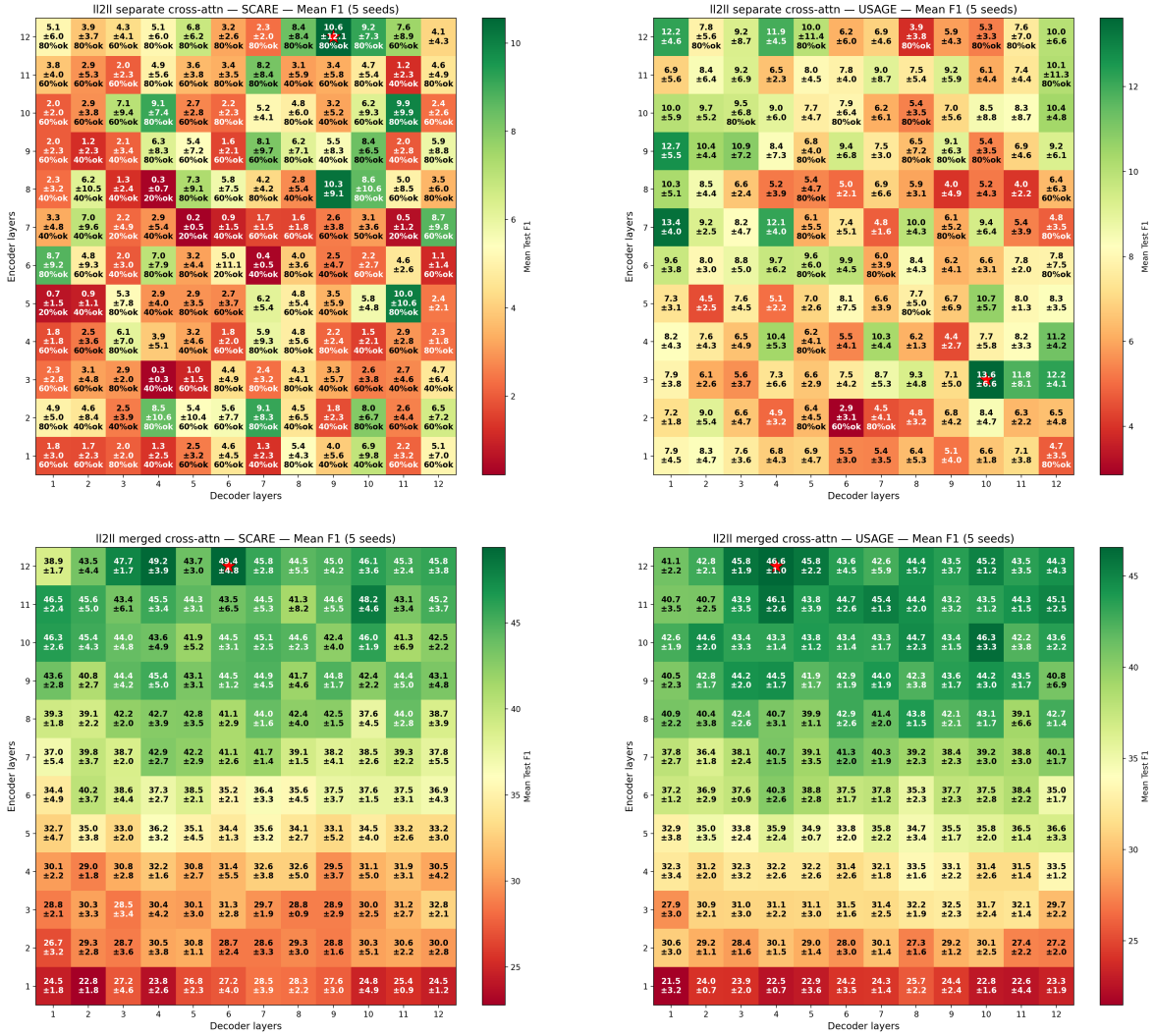

Figure 5: ll2ll encoder × decoder depth grid before (top, separate) and after (bottom, merged) the rescue, on SCARE (left) and USAGE (right), 5 seeds. Pre-rescue cells report mean F1 / std / %ok (fraction of seeds above the 5 F1 collapse threshold). Post-rescue cells report mean F1 / std (%ok is uniformly 100). Pre-rescue grids are broadly underperforming and unstable. Post-rescue grids show the encoder-heavy asymmetry of fig. 3.

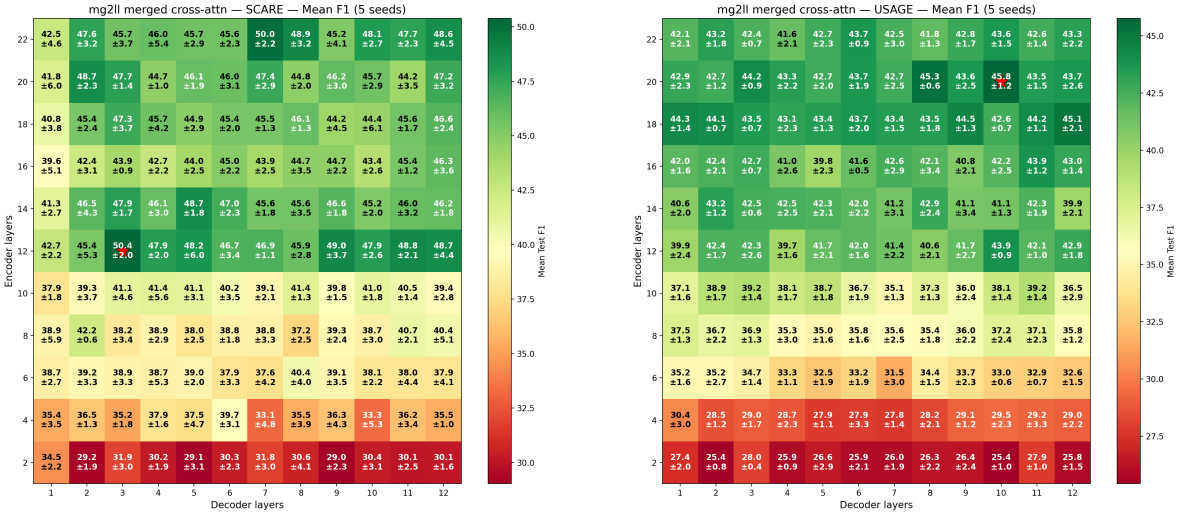

Figure 6: Encoder × decoder depth grid on mg2ll (post-rescue merged, 5-seed mean), SCARE (left) and USAGE (right). Encoder rows extend to 22 because ModernGBERT-134M is deeper than LLäMmlein-120M. The encoder-heavy asymmetry of fig. 3 holds on this heterogeneous config.

full test-set predictions, and categorize each non-matching output as a wrong sentiment class on overlapping spans, a span boundary error with partial overlap, a pure false positive, or a structurally invalid sequence. The five configurations span the strongest baseline and the strongest and weakest assemblies in our grid: mBART-large-50 (enhanced), $mg_{1B}2ll$, mg2ll, ll2ll, and lv2lv. Table 5 reports the resulting error counts on the SCARE and USAGE test sets.

Two patterns emerge. First, span boundary errors are the dominant single error category for every model on both datasets, often accounting for roughly 40–50% of the FN plus FP volume. A boundary error is one where the model overlaps a gold span but trims, extends, or otherwise shifts it: “Spiel” instead of “Super Spiel”, or “Bedienung” instead of “unlogische und unübersichtliche Bedienung”. The pattern holds across the entire parameter range we cover, from the 240M ll2ll up to the 1.1B $mg_{1B}2ll$, and is just as prominent for the mBART baseline as for the assemblies: it is a property of the ABSA task rather than a quirk of the assembled architecture.

Second, finding overlapping spans but predicting the wrong polarity is rare: between two and nineteen cases in any model/dataset cell, and always a small share of the FN plus FP volume. Combined with the structurally invalid rate of zero to four sequences per dataset (driven close to zero by merged attention, section 5.2), this leaves span boundary detection as the high-leverage direction for further improvement of pointer-decoded ABSA; gains in sentiment classification or in grammar adherence alone cannot move the metric by more than a few F1 points.

A cross-model view sharpens this further. Even in the five-model comparison, no system, mBART included, recovers any gold triplet on roughly 35% of SCARE and 44% of USAGE test sentences. Multi-triplet sentences are the clearest high-difficulty slice: on SCARE no model solves any sentence with three or more gold triplets, and on USAGE even the strongest model in that slice reaches only 19% accuracy. Stabilizing these longer roll-outs is the natural complement to better boundary detection; together they bound the headroom available to the pointer-decoded recipe on the current German ABSA benchmarks.

Two ceilings, both task-intrinsic: Span boundary errors and multi-triplet roll-outs account for most remaining headroom. Wrong sentiment and invalid outputs are small terms.

L Warm Start from Summarization

A natural question is whether the assembled models trail mBART because the in-layer cross-attention starts from scratch (section 3.1), while mBART’s was pretrained end-to-end as part of its denoising objective. We tested this cross-attention warm-up hypothesis directly: warm-start the assembled Seq2Seq, finetune it on German summarization (MLSUM, 3 epochs) so the freshly inserted cross-attention is exercised on a real generation task, then

Combo	SCARE best (e/d, F1±std)	USAGE best (e/d, F1±std)
LLäMmlein-family grid (3 seeds per cell)		
$mg_{1B}2ll$	e24/d10, 62.27 ± 1.00	e28/d8, 53.91 ± 2.08
$ll_{1B}2ll$	e22/d2, 54.29 ± 2.81	e22/d10, 52.47 ± 1.88
$mg_{1B}2ll_{1B}$	e28/d22, 54.44 ± 4.20	e20/d2, 54.07 ± 2.66
$ll_{1B}2ll_{1B}$	e22/d10, 53.16 ± 2.19	e18/d10, 50.29 ± 2.11
$mg2ll_{1B}$	e14/d22, 49.74 ± 0.86	e14/d6, 43.80 ± 2.32
$ll2ll_{1B}$	e10/d18, 46.13 ± 1.56	e12/d2, 46.65 ± 2.75
LLäMmlein2Vec-family grid (3 seeds per cell)		
$mg_{1B}2lv$	e16/d8, 55.26 ± 2.26	e28/d4, 53.94 ± 0.39
$lv_{1B}2lv$	e22/d8, 55.07 ± 1.48	e22/d8, 53.24 ± 0.90
$lv_{1B}2lv_{1B}$	e18/d14, 54.47 ± 0.59	e18/d22, 51.55 ± 0.57
$mg_{1B}2lv_{1B}$	e12/d2, 51.76 ± 2.07	e16/d18, 51.27 ± 0.49
$mg2lv_{1B}$	e22/d10, 49.33 ± 1.14	e18/d2, 44.16 ± 1.69
$lv2lv_{1B}$	e12/d18, 46.20 ± 4.39	e12/d22, 45.91 ± 4.19

Table 4: Per-combo grid-best cells for the two 1B-scale layer-cutoff sweeps (3 seeds per cell, post-rescue merged). The three combos cited in section 5.4 as the principled encoder-swap comparison are $mg_{1B}2ll$, $ll_{1B}2ll$, and $lv_{1B}2lv$; each pairs its 1B encoder with the small decoder that topped the corresponding row of the rescued 3×3 (section 5.2). Cells `e/d` are the encoder and decoder layer cutoffs of the grid-best position; encoder layers naturally extend further for ModernGBERT-1B (28 layers) than for LLäMmlein-1B (22 layers).

model	FN	FP	wrong sent.	boundary	invalid
SCARE (133 sentences, 166 gold triplets)					
mBART (enhanced)	77	29	6	45	2
$mg_{1B}2ll$	72	24	3	48	1
mg2ll	90	26	2	66	2
ll2ll	89	38	2	47	0
lv2lv	101	42	6	52	0
USAGE (238 sentences, 395 gold triplets)					
mBART (enhanced)	172	56	13	94	0
$mg_{1B}2ll$	198	82	17	124	1
mg2ll	236	70	19	136	2
ll2ll	212	83	6	146	4
lv2lv	230	70	12	143	1

Table 5: Span boundary errors dominate for every model on both datasets. Wrong sentiment on overlapping spans and structurally invalid outputs together account for under 15% of all errors. Single-seed (42), full-test predictions. The FP column counts unmatched predictions with no span overlap, and the boundary column counts unmatched predictions with partial span overlap.

proceed to the standard ABSA pointer pipeline. Across three configurations (mg2ll, ll2ll, $mg_{1B}2ll$) and both SCARE and USAGE, the five-seed mean F1 drops in all six conditions, by 3.8 to 13.6 points, and seed-to-seed variance roughly doubles for the small assemblies. The damage is largest for the smaller cells (mg2ll and ll2ll lose about 13 F1 on SCARE) and smallest for the 1B-encoder cell ($mg_{1B}2ll$, about 6 F1 on both datasets). A plausible reading is that summarization biases the encoder toward document-level semantics and trains cross-attention to attend broadly, both of which conflict with the per-token pointing the head needs, but ruling out scale, corpus, or objective effects would be a separate study. We did not pursue it further, as our takeaway for this work is that warming the cross-attention up on a summary generation task is not necessarily beneficial.

Summarization pretraining hurts: Pretraining on summarization to "align" the cross-attention before ABSA significantly *hurts* downstream F1.