Token Distribution versus Data Volume: Domain Balancing in Multi-Domain Meeting Summarisation

Ashima Sood **Bryan Gardiner** **Joan Condell**
School of Computing, Engineering and Intelligent Systems, Ulster University
Londonderry, Northern Ireland, United Kingdom
{sood-a1, b.gardiner, j.condell}@ulster.ac.uk

Abstract

Jointly fine-tuning an LLM on meeting-summarisation corpora of widely varying size raises a question that prior work leaves confounded: when a domain-balanced training mixture helps, is the gain due to the distribution of tokens across domains, or merely to the volume of data seen? We disentangle these factors by constructing balanced and natural (native-proportional) token mixtures at matched token budgets (2-32M) over five English meeting corpora, fine-tuning Mistral-7B with QLoRA, and evaluating per domain. Balancing redistributes quality, improving the data-scarce minority domains at a low cost to the data-rich ones. The trade favours balancing whenever the minority domains matter: their share under proportional allocation is fixed at 1-2% regardless of budget, so matching balanced quality on those domains requires far more total data. We further find that pruning low-value transcript lines removes $\sim$15% of tokens from the conversational corpora at no measurable cost, and that balancing by tokens is not the same as balancing by examples. Fine-tuning one model per domain is competitive only on the data-rich domains and falls below the zero-shot model on the data-scarce ones. A two-annotator study of 741 judge-labelled facts validates our fact-level evaluation. Together these results give practitioners a basis for deciding when to balance an imbalanced multi-domain mixture, and on what unit.

1 Introduction

Meeting summarisation is increasingly handled by large language models (Fu et al., 2024; Laskar et al., 2024; Kirstein et al., 2025), but real deployments must support multiple meeting domains such as project meetings, academic discussions, parliamentary sessions, and others that differ by orders of magnitude in the amount of available training data. When such corpora are pooled to jointly fine-tune a single model, the larger domains dominate the training signal, and the smaller domains are drowned out. A natural instinct is to balance the training mixture so that each domain contributes comparably (Li et al., 2025).

However, whether balancing actually helps is difficult to establish, as it entangles two effects. To give data-scarce domains a larger share within a fixed training budget, a balanced mixture reallocates tokens away from data-rich domains; granting scarce domains more data outright instead changes the total volume. Any change in quality could therefore stem from the distribution of tokens across domains or from the volume of data, and the two typically move together. Until the two are separated, the central question cannot be answered: is it the token distribution that helps, or the amount of data? The distinction matters wherever models are built from heterogeneous datasets, where imbalance is the norm rather than the exception.

This paper separates them through a controlled analytical study. Holding data volume fixed and varying only the token distribution, we examine how domain balancing affects per-domain summarisation quality: when it helps, for which domains, and at what budget. We study five English meeting corpora spanning project, academic meetings, parliamentary, and municipal proceedings, whose training sizes are severely imbalanced. We keep volume constant by constructing the natural (native-proportional) and balanced distributions at the same token budgets, across a ladder of 2, 4, 8, 16, and 32M tokens, so that a balanced-versus-natural comparison at a matched budget isolates the distribution effect. We fine-tune Mistral-7B-Instruct-v0.3 (Mistral-7B) and evaluate per domain. We release all fine-tuned adapters, including the per-domain baselines.[1]

[1] https://huggingface.co/soodashima91/meeting-summarization-domain-balancing

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 1348–1382
October 17-21, 2026 ©2026 Association for Computational Linguistics

Research Questions. We organise the study around six research questions.

- **RQ1 (Distribution).** Independent of data volume, to what extent does domain-balanced token allocation affect summarisation quality relative to the natural distribution?
- **RQ2 (Token efficiency).** How does quality scale with the token budget (2→32M), and is balancing more valuable at smaller budgets?
- **RQ3 (Pruning).** Does removing low-value transcript lines from the training input preserve quality at a reduced token count?
- **RQ4 (Unit of balancing).** Does balancing by token count differ from balancing by example count?
- **RQ5 (Model scale).** Do the findings hold on a smaller model (Llama-3.2-3B-Instruct) from a different family?
- **RQ6 (Joint versus per-domain).** Does jointly fine-tuning one model on a domain mixture outperform fine-tuning a separate model for each domain?

Contributions. **(i)** By comparing balanced and natural mixtures at matched token budgets, we separate token distribution from volume and show that balancing redistributes quality, raising minority domains while lowering majority domains, rather than adding it uniformly (RQ1). **(ii)** We characterise quality across the 2→32M budget ladder and find that the distribution gap persists at nearly every budget, closing only when natural allocation spends more budget (RQ2). **(iii)** We show that pruning low-value transcript lines preserves quality while cutting ∼15% of tokens from the conversational corpora (RQ3). **(iv)** We find that balancing by token rather than examples weights domains differently as equal meeting counts produce very unequal token amounts (RQ4). **(v)** We show these trends hold for a smaller model from a different family (Llama-3.2-3B), indicating they are not artifacts of a single backbone (RQ5). (vi) Under an identical protocol, per-domain fine-tuning matches joint training only on the data-rich domains; the data-scarce domains are learned only through joint training (RQ6).

2 Related Work

Meeting summarisation. Abstractive meeting summarisation was first benchmarked on AMI (Kraaij et al., 2005) and ICSI (Janin et al., 2003), with early encoder-decoder models such as HMNet (Zhu et al., 2020) and DialogLM (Zhong et al., 2022) achieving strong single-domain results. Later resources broadened the setting to longer, more heterogeneous transcripts, including MeetingBank (Hu et al., 2023) and the ELITR Minuting Corpus (Nedoluzhko et al., 2022), and QMSum (Zhong et al., 2021) which spans product, academic and committee meetings and reports both joint and single-domain training, and the AutoMin shared tasks (Ghosal et al., 2021, 2023; Shinde et al., 2025) established automatic minuting as a community benchmark. Its most recent edition reports an unspecified GPT-4 baseline leading the minuting task, though the organisers attribute this to self-preference in their GPT-based judge, and the lead does not hold under the reference-based metrics they also report. Recent LLM-based work targets summarisation quality (Fu et al., 2024) rather than data allocation across domains; such systems typically train a separate model per dataset, and none addresses the cross-domain imbalance that arises when one model is jointly fine-tuned on corpora of widely varying size.

Data mixtures for fine-tuning. How to weight sources of differing size is a standard question in mixture construction: proportional allocation samples each source by its native token mass, uniform allocation gives each an equal share. The trade-off is well studied at pretraining scale where domain weights are learned (Xie et al., 2023) or predicted from mixing laws (Ye et al., 2025), but data mixing for supervised fine-tuning remains comparatively underexplored (Li et al., 2025; Renduchintala et al., 2024), and pretraining methods do not transfer directly. The closest work to ours is budget-constrained instruction tuning: ADAPT (Kadasi et al., 2025) compares uniform (equal-token) and size-proportional sampling under a token budget, the two schemes we call balanced and natural, and reports diminishing returns beyond a small fraction of the budget, as we do. But that line of work learns the mixture, whereas we hold it fixed and use it as an analytical probe: we instantiate both schemes at matched token budgets, so allocation is the only variable and its effect is not confounded with volume (Section 3.2), and we mix domains rather than task types, on real meeting corpora ranging from tens of meetings to several thousand.

Positioning and PEFT. We fine-tune with QLoRA (Dettmers et al., 2023) under a fixed compute budget. The closest meeting-specific work, MeetMulti-X (Sood et al., 2025), benchmarks LLMs across meeting corpora zero-shot, without fine-tuning or a mixture strategy. Outside meetings, Wang et al. (2026) fine-tune a 7B model with LoRA on five similarly imbalanced scientific disciplines and find that the joint model trails one adapter per discipline on average, with instability concentrated in the data-scarce disciplines. Their smallest discipline has tens of thousands of examples; ours have tens of meetings (RQ6). To our knowledge no prior work fine-tunes LLMs on token-level, domain-balanced mixtures of meeting corpora, or isolates token distribution from volume under matched budgets. No external system is jointly fine-tuned across these corpora, so our comparisons are internal: the difference between two allocation schemes at a fixed budget, which no external number can inform.

3 Methodology

We study how the distribution of training data across domains affects multi-domain meeting summarisation, keeping the total amount of data fixed. Because our five corpora vary widely in size, it is hard to tell whether domain balancing itself helps or whether the gains come from other factors it usually gets mixed up with (the amount of data, the input length, and the unit of balancing). To separate these factors, we build controlled training mixtures at a matched token budget and fine-tune a model on each.

Figure 1 provides an overview of our pipeline, summarising the complete experimental workflow. This section describes each part in turn: transcript pruning (§3.1), the three budget-allocation schemes (§3.2), baselines and model selection (§3.3), fine-tuning (§3.4), inference (§3.5) and evaluation (§3.6).

3.1 Pruning

Meeting transcripts contain a large amount of conversational filler, such as greetings, backchannels, and false starts, that adds length without adding content. To reduce input length before training, we apply a pruning that decides, for each transcript line, whether to keep or delete it. This is a line-retention task, not summarisation: we remove noise but keep every line that carries any substantive content.

We prune with gemma-3-27B-it[2] under greedy decoding, using the same model and prompt for all five corpora (prompt in Appendix O.1; detailed statistics in Appendix A). This model was selected due to its strong instruction-following capability while remaining computationally practical. The prompt sets KEEP as the default and deletes only lines that are pure filler, keeping anything with a decision, action, number, date, name, question, answer, or opinion. To guard against over-deletion, a deterministic rule force-keeps any line containing a digit, a decision or action keyword, or at least twelve content words. Pruning is applied only to the training and development data; test transcripts are always used in full, so every model is evaluated on the same inputs. Pruning removes 6.8% of training tokens overall, see Appendix Table 7 for details. A two-annotator study of 500 lines (100 per corpus, 50 Keep and 50 Delete) validates these decisions against human judgement; inter-annotator agreement is substantial and the pruner tracks annotators on the conversational corpora, over-deleting only on EuroParlMin (Appendix B).

3.2 Budget Allocation

We construct training mixtures by allocating a fixed token budget $B \in \{2, 4, 8, 16, 32\}$M tokens across the five corpora and sampling within each corpus to fill its share. The budget fixes the size of the mixture, i.e. the tokens seen per epoch; the number of tokens processed in training additionally depends on the epoch at which early stopping selects the checkpoint (Table 19). The three schemes differ only in how the budget is split across corpora; they share the same budgets, the same pruned and unpruned variants, and the same prompt, tokeniser, and random seed. This shared construction is what lets a comparison at a matched budget isolate one factor at a time. Table 1 summarises how the schemes weight the smallest corpus (ICSI) across budgets; full per-corpus, per-budget, and per-seed statistics are in Appendix C.

Balanced (equal-token) allocation. Each corpus receives an equal share of the budget (B/5 or 20% each). Meeting this quota means subsampling the large corpora, which always have more than enough data. For the small corpora the behaviour depends on the budget: at the smallest budget every meeting is used at most once, but as the budget

[2] https://huggingface.co/google/gemma-3-27b-it

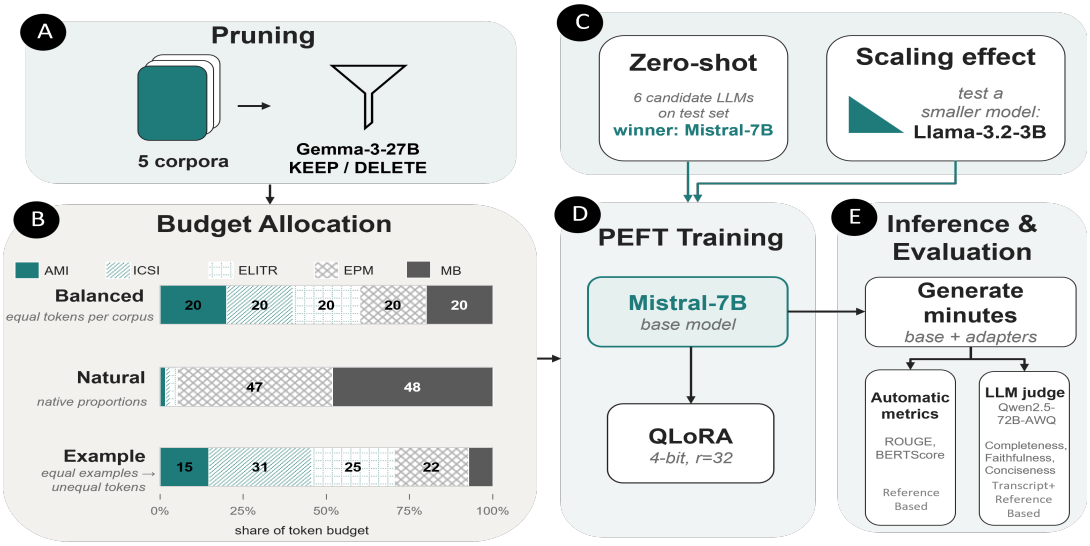


Figure 1: Pipeline. **(A)** Transcripts are pruned of conversational filler with Gemma-3-27B-it. **(B)** Training mixtures are built under three allocation schemes (balanced, natural, example) at a fixed token budget. **(C)** Zero-shot screening over six candidate LLMs selects Mistral-7B-Instruct-v0.3; Llama-3.2-3B is retained as a scaling control. **(D)** Mistral-7B is fine-tuned with QLoRA. **(E)** Generated minutes are scored with automatic metrics and an LLM judge.

grows they must be oversampled (their meetings repeated) to fill the equal share. At our largest budget (32M) this reaches roughly 11× for ICSI.

Natural (proportional) allocation. Each corpus receives a share proportional to its own token mass, so the two large corpora (EPM, MB) together take about 95% of every budget and the three small corpora only 1-2% each. These shares fit within each corpus, so natural allocation almost never oversamples. At a matched budget, natural is the volume-controlled counterpart to balanced: comparing the two measures the effect of distribution alone.

Example-level (equal-count) allocation. Instead of balancing tokens, this scheme balances example count: each corpus contributes an equal number of meetings (43, capped by the smallest corpus), giving about 1.9M tokens in total. Because transcripts differ in length, equal example counts produce very unequal token counts (ICSI contributes 578K tokens, MB only 137K (seed 1)). Comparing this scheme with the balanced one at a similar budget therefore isolates the unit of balancing (tokens versus examples). We use it only for unit comparison (RQ4).

3.3 Baselines and Model Selection

Before fine-tuning, we screen candidate base models zero-shot to choose a primary model. We prompt five instruction-tuned open models and one closed model to summarise the raw test transcripts, using the same prompt and decoding settings for all, and score their outputs with our automatic metrics. The six candidates span a range of sizes and families: Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Qwen3-8B[3], gemma-4-12B-it from HuggingFace[4] and one closed model, GPT-4o[5]. GPT-4o ranks last of the six on macro BERTScore-F1, its minutes padded with markdown structure our references do not contain (15-26% of lines), which on MeetingBank yields the highest recall of any candidate alongside the lowest precision (Appendix F). Mistral-7B leads on both micro and macro BERTScore-F1, so we select it as our primary base model for all subsequent experiments. We additionally keep the

Scheme		2M	4M	8M	16M	32M
Balanced	tokens	397K	794K	1.6M	3.2M	6.4M
	os	1.0×	1.4×	2.8×	5.6×	11.1×
Natural	tokens	26K	51K	105K	215K	436K
	os	1.0×	1.0×	1.0×	1.0×	1.0×

Table 1: Token allocation and oversampling (os) for the smallest corpus (ICSI). Balanced holds ICSI at a ∼20% budget share throughout; natural gives it ∼1.4%. Per-corpus breakdowns in Appendix C.

[3]Default non-thinking mode suggested settings.
[4]https://huggingface.co/models
[5]We used the gpt-4o-2024-11-20 model.

smaller Llama-3.2-3B as a scaling control, to test whether our findings hold at a lower model capacity (RQ5). The full zero-shot comparison across all six candidates is in Appendix F.

Per-domain fine-tuned baselines. Joint fine-tuning assumes that a single model must serve all domains. The alternative is to fine-tune one adapter per corpus. We therefore train five adapters, each on the full training set of one corpus, and evaluate each on the test meetings of its own domain. All five runs follow the protocol of §3.4 unchanged, with early stopping on the development loss of the corresponding domain, so the training data is the only variable relative to the mixture runs. Each adapter processes its corpus once per epoch, i.e. 0.66M (AMI), 0.58M (ICSI), 0.94M (ELITR), 19.7M (EPM) and 20.4M (MB) tokens, compared with 6.4M tokens per domain under balanced allocation at 32M. This baseline establishes whether joint fine-tuning, under either allocation scheme, improves on fine-tuning each domain in isolation (RQ6; Appendix G).

3.4 Fine-tuning

We fine-tune Mistral-7B-Instruct-v0.3 with QLoRA, which back-propagates through a frozen 4-bit base into low-rank adapters. Beyond cost, this suits our design: with the base model fixed and every run trained under identical hyperparameters and different training mixtures. Loss is cross-entropy over summary tokens, with prompt tokens masked so that optimisation targets generation rather than prompt memorisation. A single fixed prompt, instructing the model to ground minutes only in the transcript, is applied identically across all domains, budgets, and conditions (Appendix O.2).

3.5 Inference

At test time we generate with greedy decoding and a fixed generation budget of `max_new_tokens` $= 2048$, applied identically across all models and conditions. This budget covers every test reference in AMI, ICSI, ELITR, and MeetingBank in full, and 98.8% of test references overall; the few longer references are undivided full-chapter documents in EuroParlMin rather than typical per-session minutes (Appendix H). Test transcripts are never pruned, so zero-shot and fine-tuned conditions see identical inputs. For ELITR, which supplies multiple reference minutes per meeting, we score each generated minute as the maximum over its references, so ELITR contributes 38 meetings to evaluation (Appendix I).

3.6 Evaluation

We evaluate generated minutes with automatic metrics and a fact-level LLM judge.

Automatic metrics. We report ROUGE-1/2/L/Lsum (Lin, 2004) and BERTScore-F1 (Zhang et al., 2019), computing each per domain and then aggregating two ways. The macro average weights the five domains equally (the mean of per-domain scores); the micro average weights every meeting equally (the mean over all test meetings). The distinction is central to our analysis because the test sets are as imbalanced as the training sets: MeetingBank (862 meetings) and EuroParlMin (242) together account for 94% of test meetings, against 38 for ELITR, 20 for AMI and 6 for ICSI. The micro average is therefore dominated by the majority domains, while the macro average gives equal weight to the data-scarce domains where domain balancing has its largest effect, at the cost of inheriting the wide confidence intervals of their small test sets. We report both averages throughout, with 95% bootstrap confidence intervals where relevant, and read per-domain results before either aggregate.

Fact-level LLM judge. Automatic overlap metrics can miss factual correctness, so we additionally evaluate with an LLM-as-judge protocol adapted from (Zhou et al., 2026) using an open judge model (Qwen2.5-72B-Instruct-AWQ[6]), chosen from a different model family than our summarisers to avoid self-preference bias. The judge decomposes each summary into atomic facts and scores three dimensions: completeness, faithfulness, and conciseness. Full prompts and protocol are in Appendix O.3 and Appendix L respectively.

Human validation of the judge. To check that the judge's per-fact labels track human judgement, two annotators independently re-labelled 741 of its decisions on the balanced-32M system (30 test meetings, six per domain; facts stratified by judge label). Inter-annotator agreement is substantial on all three dimensions (Cohen's $\kappa = 0.79$-0.84), and the judge agrees with the annotators at $\kappa = 0.71$–0.76 on completeness and $\kappa = 0.57$-0.63 on faith-

[6] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct-AWQ

fulness and conciseness. The errors are one-sided where both annotators agree, the judge over-accepts facts in 87-94% of disagreements, so the absolute scores it reports are best read as upper bounds. Protocol, annotation guidelines, and full agreement results are in Appendix M.

4 Experiments

4.1 Corpora

The work is done on five English meeting-summarisation corpora that vary in genre, scale, transcription provenance, and reference convention. AMI (Kraaij et al., 2005) (scenario-driven design meetings), ICSI (Janin et al., 2003) (academic research meetings), MeetingBank (MB) (Hu et al., 2023) (ASR-transcribed municipal proceedings), the ELITR Minuting Corpus (Nedoluzhko et al., 2022) (technical project meetings), and EuroParlMin (EPM) (Ghosal et al., 2023) (parliamentary debate). They span more than two orders of magnitude in scale (Table 17), reproducing the domain imbalance that motivates our study. ELITR supplies multiple references per meeting; we consolidate each to a single reference for training and take the maximum over all references at test time. Per-corpus construction, splits, units, and the consolidation procedure are given in Appendix I.

4.2 Experimental settings

The main experimental grid is $2 \times 2 \times 5$ (allocation scheme $\times$ pruning condition $\times$ token budget), giving 20 runs (§3.2). All grid runs use seed 42. We repeat both schemes under the pruned condition at 2M, 8M, and 32M with two further seeds (2,15), confirming the per-domain pattern is not seed-specific (12 runs; Appendix E.1). To this we add the example-level baseline (§3.2), run under both pruning conditions across three seeds ($\{1, 2, 15\}$; Appendix E.2) and reported as mean $\pm$ std, for 6 runs. Unlike the token-budget schemes, it draws only 43 meetings per corpus from corpora that supply far more, so the particular draw affects both which meetings enter the mixture and its resulting token mass (Appendix Table 10); the seeds quantify that sensitivity rather than leaving it unmeasured. Finally, the Llama-3.2-3B scaling control (§3.3) repeats the pruned condition of both schemes at the 2M, 8M, and 32M budgets, for 6 runs.

4.3 Implementation Details

All runs use QLoRA with 4-bit NF4 quantisation and low-rank adapters of rank r=32 and scaling α=16, applied to all seven attention and MLP projections (q, k, v, o, gate, up, down). This trains $\approx$83.9M parameters, 1.14% of the 7B base model. Hyperparameters (Table 18) are held fixed across every scheme, budget, and condition, so that differences across runs reflect the training data rather than the training setup. We train for up to 5 epochs with early stopping on the development loss (patience 2) and retain the lowest-development-loss checkpoint; per-run loss curves and the selected epoch for each run are in Appendix K. Each run uses a single NVIDIA A100-SXM4-80GB. The QLoRA formulation is given in Appendix J.

5 Results and Discussion

We report the findings in RQ order. Our corpora are highly imbalanced, so aggregate averages are majority-weighted and can hide the effect of allocation on the smaller domains. We therefore read every result per domain and treat macro and micro averages as summaries rather than as primary evidence.

5.1 RQ1: Domain Balancing Redistributes Quality Across Domains

Finding. At matched token budgets, balancing does not lift every domain; it trades between them. It improves all three minority domains (AMI, ICSI, ELITR) on both metrics. The fact-level judge shows the same pattern.

Table 2 contrasts the two schemes per domain at 32M. Balanced leads every minority domain on both metrics, and the gains are substantial: in ROUGE-Lsum, AMI $+0.102$, ICSI $+0.164$, ELITR $+0.163$. On the majority side the movement is smaller and one-sided: MB shifts to natural by 0.046, while EPM is close. The same per-domain split holds across the full budget ladder (Table 22; Figure 2). The two aggregates therefore disagree by construction: the macro average, weighting domains equally, favours balanced (ROUGE-Lsum 0.500 vs. 0.421; BERTScore-F1 0.877 vs. 0.872), while the micro average, dominated by MB and EPM, favours natural (0.637 vs. 0.615). The effect is visible only once the domains are read separately.

The fact-level judge (Table 3) reproduces this at both budgets. On completeness, the recall of reference key facts, balanced leads on macro ($+0.017$ at 2M, $+0.041$ at 32M) and trails on micro (-0.016,

Scheme	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
ROUGE-Lsum							
Balanced	**0.495**	**0.445**	**0.367**	**0.547**	0.648	**0.500**	0.615
Natural	0.393	0.281	0.204	0.534	**0.694**	0.421	**0.637**
BERTScore-F1							
Balanced	**0.875**	**0.852**	**0.846**	0.887	0.928	**0.877**	0.915
Natural	0.865	0.833	0.837	**0.889**	**0.938**	0.872	**0.923**

Table 2: RQ1: balanced vs. natural allocation per domain at the matched 32M budget (pruned, seed 42). Bold marks the higher-scoring scheme, italics the test-set meeting counts. Macro weights domains equally, micro weights meetings. Full grid in Table 22; three-seed means in Table 27.

System	Completeness	Faithfulness	Conciseness
Macro			
balanced 2M	$\mathbf{0.276}_{[.25,.30]}$	$\mathbf{0.765}_{[.74,.79]}$	$\mathbf{0.515}_{[.48,.54]}$
natural 2M	$0.259_{[.24,.28]}$	$0.739_{[.70,.78]}$	$0.467_{[.43,.51]}$
balanced 32M	$\mathbf{0.373}_{[.35,.40]}$	$\mathbf{0.742}_{[.73,.76]}$	$0.472_{[.45,.49]}$
natural 32M	$0.332_{[.31,.36]}$	$0.728_{[.70,.75]}$	$\mathbf{0.506}_{[.46,.56]}$
Micro			
balanced 2M	$0.427_{[.41,.45]}$	$\mathbf{0.707}_{[.69,.73]}$	$0.573_{[.55,.59]}$
natural 2M	$\mathbf{0.443}_{[.42,.46]}$	$0.685_{[.67,.70]}$	$\mathbf{0.575}_{[.55,.60]}$
balanced 32M	$0.507_{[.49,.53]}$	$\mathbf{0.642}_{[.62,.66]}$	$0.589_{[.57,.61]}$
natural 32M	$\mathbf{0.522}_{[.50,.54]}$	$0.636_{[.62,.66]}$	$\mathbf{0.625}_{[.61,.64]}$

Table 3: Fact-level LLM-judge scores at the smallest and largest budgets, with bootstrap 95% CIs (10,000 resamples). Balanced leads on macro completeness and trails on micro, the same sign flip the automatic metrics show. Per-domain breakdown in Table 25.

-0.015), the same sign flip as in the automatic metrics, now measured over facts rather than surface overlap. Faithfulness is comparable between the schemes but falls as the budget grows (balanced $0.765 \rightarrow 0.742$, natural $0.739 \rightarrow 0.728$), most sharply on EPM, where longer generations draw on transcript beyond our 16K truncation. Natural is more concise at 32M (0.506 vs. 0.472 macro): balanced recovers more reference content at a cost in per-fact salience.

5.2 RQ2: Token efficiency and the budget dependence of balancing

Finding. Balancing's minority-domain advantage does not close with budget. Proportional allocation fixes their share at 1-2%, so at 32M it gives ICSI barely more tokens than balancing gives it at 2M.

Figure 2 plots per-dataset quality across the budget ladder. Both metrics show the diminishing-returns shape of a token-efficiency curve, rising with budget and flattening at the top, most clearly in BERTScore-F1. In the minority domains balanced climbs with the budget and leads natural at

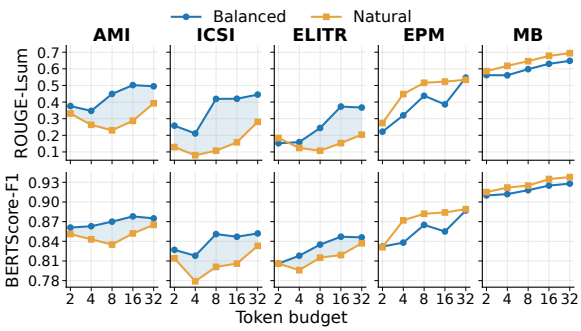


Figure 2: RQ2: Per-dataset quality across the token budget ladder (2→32M) for balanced (blue) and natural (orange) allocation, on ROUGE-Lsum (top) and BERTScore-F1 (bottom). Shading marks balanced's advantage on the minority panels. Quality rises with budget and flattens; the minority gap is open at nearly every budget and does not close by 32M.

nearly every budget, while natural stays low and flat, held down by the ∼1-2% share proportional allocation assigns those domains at every budget.

As the budget grows, natural routes more absolute tokens to the scarce domains and closes part of the gap, but does not catch balanced: at 32M it still trails on every minority domain. The asymmetry is stark in absolute terms. Proportional allocation gives ICSI 436K tokens at 32M, barely more than the 397K balanced already assigns it at 2M (Table 1), so natural must spend sixteen times the budget to buy the same exposure. On the majority side the schemes differ: on MB natural stays ahead throughout, while on EPM natural's early lead narrows with budget and at 32M the two are indistinguishable, the seed-42 difference (0.547 vs. 0.534 on ROUGE-Lsum) lying within seed variation (Table 27). Balanced allocation is therefore attractive whenever the minority domains matter, since natural can close the gap only by spending budget it may not have.

5.3 RQ3: Pruning preserves quality at fewer tokens

Finding. Pruning low-value transcript lines removes 6.8% of training tokens, most of it from the conversational corpora, while preserving quality: at matched budgets, pruned and unpruned runs are near-identical on BERTScore-F1 and converge on ROUGE-Lsum as the budget grows.

We compare pruned and unpruned training at matched budgets. Equal token counts give equal compute (FLOPs differ by 0.04% on average), so any quality difference reflects what the tokens contain rather than how many there are. Per-corpus reductions are in Appendix A.

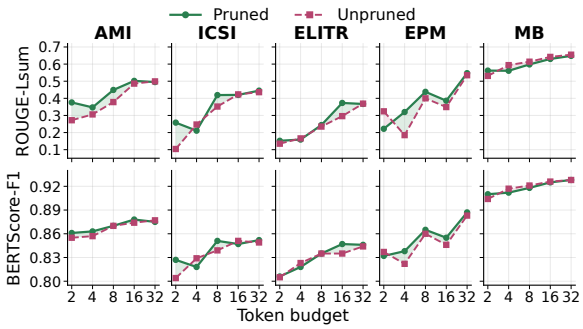


Figure 3: RQ3: Pruned vs. unpruned training quality per dataset across the budget ladder (balanced allocation), on ROUGE-Lsum (top) and BERTScore-F1 (bottom). The two conditions are near-identical on BERTScore-F1 and converge on ROUGE-Lsum as the budget grows; the visible gaps are confined to the smallest budgets.

Figure 3 plots the two conditions across the budget ladder. On BERTScore-F1 the curves are near-identical throughout, so removing the conversational fillers leaves semantic adequacy intact. In ROUGE-Lsum they track closely at the mid and large budgets and meet by 32M (mean absolute difference under 0.01); the gaps at 2M and 4M, running in both directions across ICSI, AMI, and EPM, are variance rather than a cost of pruning.

Pruning also interacts with balancing. Pruned transcripts are 7-16% shorter, so a fixed budget spans 8-19% more meetings. Where the balanced mixture oversamples the minority corpora heavily (up to 11× for ICSI; Table 8), the shorter transcripts fill the same quota with less repetition and a higher share of distinct content, so pruning is most useful exactly where balancing is most aggressive.

Our matched-budget design cannot show this saving, since equalising tokens equalises compute by construction. The saving appears when the corpus rather than the budget is fixed: pruning then trains on 6.8% fewer tokens for proportionally less compute, at no measurable cost.

5.4 RQ4: The unit of balancing matters

Finding. Tokens, not examples, are the unit that matters. Equal meeting counts hand ICSI four times MB's token mass, and per-domain quality follows the tokens, moving the two domains in opposite directions. Balancing on examples therefore allocates by accident; balancing on tokens allocates by design.

Example-level allocation gives every corpus the same number of meetings (43, capped by ICSI). But transcripts vary in length, so equal meeting counts turn into very unequal token mass: ICSI brings 578K tokens and MB only 137K (seed 1),

Scheme	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro
ROUGE-Lsum						
Balanced	.376	.258	.152	.222	**.562**	.314
Natural	.331	.129	.184	.274	.586	.301
Example	$.328_{\pm.044}$	$\mathbf{.292}_{\pm.036}$	$.229_{\pm.014}$	$.271_{\pm.044}$	$.505_{\pm.008}$	$.325_{\pm.017}$
BERTScore-F1						
Balanced	.861	.827	.806	.832	**.916**	.847
Natural	.851	.814	.806	.831	.915	.843
Example	$.860_{\pm.004}$	$\mathbf{.834}_{\pm.003}$	$.826_{\pm.006}$	$.833_{\pm.013}$	$.898_{\pm.000}$	$.850_{\pm.004}$

Table 4: RQ4: Example-level balancing at ∼2M. Example-level (43 meetings per corpus) is mean±std over seeds {1,2,15}; Balanced and Natural are seed 42. Equal meeting counts give ICSI ∼578K tokens but MB only ∼137K. Full results in Appendix Table 21.

a ∼4× gap (Table 10). Token-balancing does the reverse, holding every corpus to the same token share.

The difference shows up exactly where transcript lengths are most extreme (Table 4). MB has short transcripts, is starved of tokens under equal example counts, and drops accordingly (0.505 vs. 0.562), a change well outside its seed variance (±0.008). ICSI has long transcripts, so counting by meetings hands it far more tokens than token-balancing would, and it moves in the opposite direction (0.292 vs. 0.258); with only six test meetings its seed variance is large (±0.036), so we read the direction rather than the magnitude. The two domains move opposite ways on the same data. Both moves track token mass, not meeting counts. On BERTScore-F1 the gap is small, so the unit mainly moves surface overlap rather than semantic adequacy.

ICSI's higher score under example-level balancing is therefore not a point in that scheme's favour, but rather transcript length leaking into the allocation. Token-balancing instead assigns each domain the same budget by construction, which is the control our matched-budget study relies on: because example counts do not correspond to a fixed token budget, the RQ1-RQ3 comparisons require balancing on tokens.

5.5 RQ5: Generalisation to a smaller model

Finding. The redistribution replicates on Llama-3.2-3B. Balancing helps the same three minority domains and costs the same two majority domains on both metrics. Absolute quality is lower, so what transfers is the direction of the effect, not its magnitude.

An allocation effect is only useful if it survives a change of model and capacity, so we repeat the comparison on Llama-3.2-3B, a different family at half the scale. Across the three shared budgets (2M/8M/32M), Table 5 reports the mean

	ROUGE-Lsum		BERTScore-F1	
Dataset	M-7B	L-3B	M-7B	L-3B
AMI	+0.122	+0.155	+0.018	+0.034
ICSI	+0.202	+0.141	+0.027	+0.040
ELITR	+0.089	+0.072	+0.010	+0.032
EPM	−0.039	−0.178	−0.006	−0.031
MB	−0.039	−0.061	−0.007	−0.014

Table 5: RQ5: per-dataset mean balanced−natural gap for M-7B (Mistral-7B) and L-3B (Llama-3.2-3B), averaged over the three shared budgets (2M/8M/32M; pruned, seed 42). Positive favours balancing. Per-budget numbers in Table 26.

balanced−natural gap per dataset. The sign agrees on every dataset and both metrics: positive on the minority domains and negative on the majority domains. That the same pattern holds under a different, smaller model indicates the redistribution comes from the data allocation, not from Mistral-7B.

5.6 RQ6: Joint versus per-domain fine-tuning

Finding. Per-domain fine-tuning exceeds balanced allocation on the two data-rich domains and falls below it on the three data-scarce domains. Joint training is what makes the minority domains learnable.

Table 6 compares the five per-domain adapters with balanced allocation at 32M. On EPM and MB the per-domain adapters lead on both metrics, and the per-domain system attains the highest micro average in this study (0.647 vs. 0.615). On AMI, ICSI and ELITR the ordering reverses: the per-domain adapters trail by 0.11–0.24 ROUGE-Lsum, with non-overlapping confidence intervals on ICSI and ELITR, and the macro average falls from 0.500 to 0.397. Trained on 43–98 meetings under the fixed protocol, these adapters acquire each corpus's reference format but not its content (Appendix G.1 for details). Per-domain fine-tuning therefore serves only majority-weighted deployments; balanced joint training remains the single configuration that serves all five domains.

6 Conclusion

EuroParlMin and MeetingBank account for roughly 95% of the token mass across our five domains, so under proportional allocation a jointly fine-tuned model is trained almost exclusively on them, and its per-domain quality follows each domain's share of the training signal. Domain balancing removes this dependence: at a matched 32M budget it leads on all three data-scarce domains on both metrics. Balancing thus does not raise quality on every domain but reallocates it across them, and the reallocation is asymmetric. Proportional allocation holds the minority domains at a 1-2% token share at every budget, and narrows the gap only by increasing their absolute volume; at 32M it has still not matched balancing on any of the three. Two further results bear on how the budget is spent: pruning conversational filler preserves quality at no measurable cost, and the unit of balancing is consequential, as equal example counts assign ICSI four times the token mass of MeetingBank. The effect replicates on a smaller model from a different family. Fine-tuning each domain in isolation matches joint training only on the data-rich domains; on the data-scarce domains it falls below the untuned model, so joint training is what makes those domains learnable at all.

Imbalance of this severity is characteristic of real multi-domain deployments rather than an artefact of our design. Our findings offer practitioners a basis for allocation decisions: retain the native distribution when the deployment is majority-weighted, balance when all domains must be served, balance by tokens rather than examples, and prune prior to training in either case.

System	**AMI** *20*	**ICSI** *6*	**ELITR** *38*	**EPM** *242*	**MB** *862*	**Macro**	**Micro**
ROUGE-Lsum							
Per-domain	0.382	0.210	0.126	**0.564**	**0.703**	0.397	**0.647**
Balanced 32M	**0.495**	**0.445**	**0.367**	0.547	0.648	**0.500**	0.615
BERTScore-F1							
Per-domain	0.868	0.817	0.772	**0.897**	**0.939**	0.858	**0.923**
Balanced 32M	**0.875**	**0.852**	**0.846**	0.887	0.928	**0.877**	0.915

Table 6: RQ6: per-domain fine-tuned baselines vs. balanced allocation at 32M (pruned, seed 42). Each per-domain adapter is trained on the full training set of one corpus under the protocol of Table 18 and evaluated on its own domain. Bold marks the higher system per column; italics are test-set meeting counts. Macro weights domains equally, micro weights meetings. Bootstrap CIs in Table 14.

Limitations

Meeting transcripts are long, and a full transcript with its minute exceeds the 16,384-token window we can fit on a single device. Truncation falls unevenly across the corpora: 54% of ICSI and 33% of ELITR training transcripts are cut, against 24% of EuroParlMin's, but severity runs the other way. ICSI loses 18% of a median 15.8K-token transcript, whereas EuroParlMin's truncated transcripts retain

only 65% of content against a maximum length twelve times the window. The consequence is a training target not fully supported by its input: the reference minute covers the whole session while the input is cut, so on EuroParlMin the model is trained to assert outcomes it cannot observe. Its low faithfulness on that corpus is consistent with this, though we do not isolate it as the cause; ICSI, truncated more often but less severely, records the highest faithfulness of any corpus. At test time the constraint is far milder, affecting 3.3% of transcripts, so evaluation inputs are near-complete.

Our fact-level judge is validated against two annotators on one system, and its errors are one-sided: it accepts facts the annotators reject in 87–94% of disagreements, so its absolute scores are too high. But every system is scored by the same judge under the same rubric, so the differences between systems still hold.

The findings are established up to 7B parameters under low-rank adaptation. Whether the redistribution effect holds at larger scale or under full fine-tuning is untested; our cross-family control, Llama-3.2-3B, reproduces the direction of the effect on every domain.

Our automatic metrics reward agreement with each corpus's reference convention as well as content coverage, as the zero-shot screening makes visible. This bounds what the absolute scores mean, though not the balanced-versus-natural comparisons, which hold the reference form fixed.

All experiments use English meeting transcripts. Whether domain balancing interacts with language balancing, or transfers to other multi-domain tasks, is left open.

Ethics

All corpora are publicly available and used under their released licences. We perform no re-identification and add no annotation beyond schema unification and the reference-selection and pruning steps described. The two annotators for the pruning and judge validation studies are colleagues of the authors with backgrounds in NLP, not involved in developing the model or producing the results. They were compensated at a locally appropriate rate and participated voluntarily. Our pruning and mixture procedures are designed to preserve substantive content, but automatic minuting can omit or distort information and should not be relied on as an authoritative record without human review. We report compute and seed budgets transparently and concentrate repeated runs on cheaper configurations to limit energy use. All fine-tuning and open-model inference run on a single A100, and the fact-level judge on a single H100; the only external cost is the GPT-4o zero-shot baseline, a single closed-model comparison run that incurred approximately USD $\approx$ 15.5 in API charges.

We used LLM-based tools for editorial assistance, such as rewriting for clarity, shortening text, improving grammar, and ensuring stylistic consistency. All experimental design, analyses, results, and claims are the authors' own, and the authors verified all AI-assisted text and take full responsibility for the content.

Acknowledgments

The authors would like to thank the anonymous reviewers for the their valuable feedback which helped improve the quality of this paper. In addition, the authors are grateful for access to the Tier 2 High-Performance Computing resources provided by the Northern Ireland High Performance Computing (NIHPC) facility funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant No. EP/T022175/1.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394.

Tirthankar Ghosal, Ondřej Bojar, Marie Hledíková, Tom Kocmi, and Anna Nedoluzhko. 2023. Overview of the second shared task on automatic minuting (AutoMin) at INLG 2023. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 138–167, Prague, Czechia. Association for Computational Linguistics.

Tirthankar Ghosal, Ondřej Bojar, Muskaan Singh, and Anja Nedoluzhko. 2021. Overview of the first shared task on automatic minuting (automin) at interspeech

2021. *Proceedings of the First Shared Task on Automatic Minuting at Interspeech*, pages 1–25.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Pritam Kadasi, Abhishek Upperwal, and Mayank SIngh. 2025. Adapt: Learning task mixtures for budget-constrained instruction tuning. *arXiv preprint arXiv:2512.04555*.

Frederic Kirstein, Terry Ruas, and Bela Gipp. 2025. What's wrong? refining meeting summaries with LLM feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2100–2120, Abu Dhabi, UAE. Association for Computational Linguistics.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2024. Query-OPT: Optimizing inference of large language models via multi-query instructions in meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1140–1151, Miami, Florida, US. Association for Computational Linguistics.

Yuan Li, Zhengzhong Liu, and Eric Xing. 2025. Data mixing optimization for supervised fine-tuning of large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 35419–35437. PMLR.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. Elitr minuting corpus: A novel dataset for automatic minuting from multi-party meetings in english and czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182.

H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Ganesh Ramakrishnan. 2024. SMART: Submodular data mixture strategy for instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12916–12934, Bangkok, Thailand. Association for Computational Linguistics.

Kartik Shinde, Laurent Besacier, Ondrej Bojar, Thibaut Thonet, and Tirthankar Ghosal. 2025. Findings of the third automatic minuting (automin) challenge. *arXiv preprint arXiv:2509.13814*.

Ashima Sood, Muskaan Singh, Bryan Gardiner, and Joan Condell. 2025. Meetmulti-x: A benchmark analysis of scaling and prompting large language models on automatic minuting. *Expert Systems with Applications*, page 130428.

Lintao Wang, Zhuqiang Lu, Yilin Zhu, Kun Hu, Zhenfei Yin, Shixiang Tang, Zhiyong Wang, Wanli Ouyang, and Xinzhu Ma. 2026. Charting empirical laws for llm fine-tuning in scientific multi-discipline learning. *arXiv preprint arXiv:2602.11215*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. 2025. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and 1 others. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Weixiao Zhou, Gengyao Li, Xianfu Cheng, Junnan Zhu, Feifei Zhai, and Zhoujun Li. 2026. A large-scale multi-dimensional empirical study of llms for conversation summarization. *arXiv preprint arXiv:2606.15974*.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

Appendix

The appendix is organised as follows.

Appendix Contents

A		Pruning Statistics
B		Human Validation of Pruning
C		Data-Mixture Construction Details
D		Transcript Length and Truncation
E		Multi-Seed Robustness
	E.1	Balanced and Natural Schemes
	E.2	Example-Level Baseline
F		Model Selection
G		Per-Domain Fine-tuned Baselines
	G.1	Setup and Training Trajectories
	G.2	Full Results
H		Summary Lengths (Reference and Generated)
I		Corpus Construction
J		Fine-tuning Hyperparameters
K		Training Trajectories and Checkpoint Selection
L		LLM Judge Protocol
M		Human Validation of the LLM Judge
	M.1	Annotation Guidelines
N		Full Per-Domain Results
O		Prompts
	O.1	Pruning Prompt
	O.2	Summarisation Prompt
	O.3	LLM-as-a-Judge Prompts

A Pruning Statistics

Table 7 reports the effect of transcript pruning on the training data, with token counts measured using the Gemma-3-27B-it tokenizer. Token reduction is strongly domain-dependent: pruning removes 14–16% of tokens from the conversational corpora (AMI, ICSI, ELITR), which contain substantial backchannel and filler, but only 2.3% from EuroParlMin, whose formal parliamentary prepared and edited speeches carries little removable noise. The aggregate reduction (6.8%) is dominated by the two largest corpora and understates the savings on the smaller conversational domains. The final column reports the number of lines the deterministic content guard force-kept. The guard overrides the pruning model whenever a line contains a digit, a decision or action keyword, or at least twelve content words, ensuring such lines are never deleted regardless of the model's decision. Across all corpora it force-kept 225,191 lines, bounding the recall of substantive content independently of the pruning model. The full pruning prompt is given in Appendix subsection O.1

Corpus	Transcripts	Tokens (pre)	Reduction	Guard-saved
AMI	118	863,292	15.6%	3,030
ICSI	52	771,303	14.7%	2,586
ELITR	94	1,164,265	14.4%	3,770
MB	6,030	28,616,375	9.8%	69,493
EPM	2,252	23,878,795	2.3%	146,312
Total	**8,546**	**55,294,030**	**6.8%**	**225,191**

Table 7: Transcript pruning on the training data (train+dev; test is unpruned), Gemma-3-27B-it tokenizer. Reduction is the fraction of tokens removed; Guard-saved is the number of lines force-retained by the content guard.

B Human Validation of Pruning

Setup We validate the pruning decisions against two annotators, mirroring the judge-validation design. Pruning is applied to the training split, where KEEP dominates, so a uniform sample would leave few DELETE items on the majority corpora and let a constant-KEEP annotator agree by default. We therefore sampled 50 KEEP and 50 DELETE lines per corpus (500 items total), drawing each stratum across many transcripts so no single meeting dominates, and restricting KEEP items to non-guard lines—those the deterministic content guard did not force-keep—so that agreement reflects the pruning model's judgement rather than the guard. Each line was shown in ±4 lines of context with the model's decision hidden; the annotated line is the one judged, the context is reference only. Both annotators labelled every item independently, and items were shuffled so the KEEP/DELETE composition was not visible. The following annotation guidelines were given verbatim to both annotators.

> **Your task.** Each row of the spreadsheet shows one meeting-transcript line. You decide whether that line should be kept or deleted from the transcript. The goal is to remove pure conversational noise while keeping anything with real content. In the `human_decision` column, write `Keep` or `Delete` for each row.

Which line to judge. Judge only the target line – the one marked with »> in the `context` column (also shown in `target_line`). The other lines in `context` are there only to help you understand the target line. Do not mark them.

How to decide. Delete a line only if it is pure conversational noise with no standalone content:

- bare acknowledgements / fillers: "okay", "right", "yeah", "mm-hmm", "uh", "cool";
- greetings, goodbyes, thanks, social pleasantries;
- connection/logistics chatter: "can you hear me?", "you're muted";
- meaningless fragments or false starts that carry no information.

Keep a line if it contains any real content, even briefly:

- a topic, fact, opinion, reason, problem, suggestion, or description;
- a decision, agreement, action item, deadline, number, price, name, or date;
- a question, or an answer to one — even a short one ("Friday", "Yes", "25 euros").

Rules.

- Keep is the default. Only delete when you are confident the line is pure noise.
- Use the context to judge short lines. "Friday." is Keep if it answers a question, Delete if it is just filler.
- When unsure, Keep.
- Do not confer with the other annotator; make every call yourself.
- Do not leave any row blank.

Results Inter-annotator agreement is substantial (Cohen's κ=0.67 pooled; 0.62 on the four conversational corpora), establishing a reliable ground truth. Against it the pruner agrees with the annotators on KEEP decisions (ratified in 72–89% of cases) and, on the conversational corpora, on DELETE decisions (52–84% ratified; Figure 4). EuroParlMin is the sole failure: annotators agreed almost perfectly there (κ=1.0) yet ratified only 4% of the model's deletions, the pruner removing substantive parliamentary content that both annotators kept. This matches EuroParlMin's formal, low-filler style, a pruner tuned for conversational filler has little to remove and misfires and is consistent with the corpus's lower faithfulness under both allocation schemes. Because EuroParlMin loses only 5.6% of lines (2.3% of tokens) to pruning, this affects a small fraction of the corpus and does not bear on the matched-budget comparisons, which equalise tokens by construction.

C Data-Mixture Construction Details

This appendix section provides the complete per-corpus token and meeting counts for every mixture,

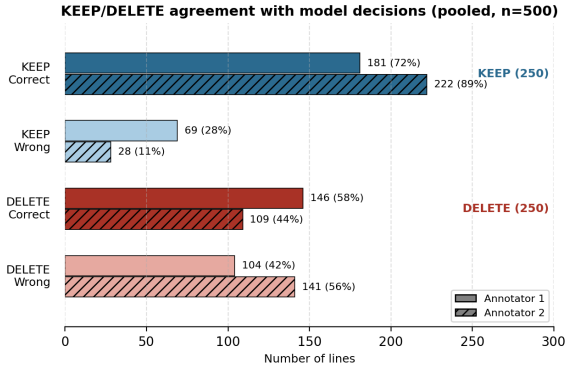


Figure 4: Annotator ratification of the pruner's KEEP-/DELETE decisions, pooled over 500 sampled lines (250 KEEP, 250 DELETE; 100 per corpus). Bars show how many of the model's KEEP and DELETE calls each annotator judged correct vs. wrong.

Corpus		2M	4M	8M	16M	32M
AMI	tokens	399K	800K	1.6M	3.2M	6.4M
	oversample	1.0	1.19	2.45	4.86	9.73
ICSI	tokens	397K	794K	1.6M	3.2M	6.4M
	oversample	1.0	1.37	2.79	5.56	11.09
ELITR	tokens	399K	800K	1.6M	3.2M	6.4M
	oversample	1.0	1.0	1.70	3.38	6.82
EPM	tokens	400K	800K	1.6M	3.2M	6.4M
	oversample	1.0	1.0	1.0	1.0	1.0
MB	tokens	400K	800K	1.6M	3.2M	6.4M
	oversample	1.0	1.0	1.0	1.0	1.0

Table 8: **Balanced scheme**: per-corpus token allocation and oversampling ratio across budgets. Each corpus receives an equal $\sim B/5$ share, so the minority corpora (AMI, ICSI, ELITR) are increasingly oversampled to fill it while the majority corpora (EPM, MB) always draw distinct meetings ($1.0\times$).

across all budgets and (for example-level) all seeds. Table 8 gives the balanced scheme, Table 9 the natural scheme, and Table 10 the example-level seed variation.

D Transcript Length and Truncation

Table 11 reports transcript lengths and truncation counts per corpus and split, measured with the Mistral-7B tokeniser. The budget available to the transcript is not the full 16,384-token window: the chat template consumes 70 tokens, and at training time the reference minute must also fit, so the transcript budget is 16,384 minus template, summary, and end-of-turn token. Test-time inputs are subject only to the template overhead.

Truncation is a training-side constraint. Across train and development, 12.3% of transcripts exceed the window; at test time only 3.3% do, so

Corpus		2M	4M	8M	16M	32M
AMI	tokens	31K	61K	125K	249K	500K
	meetings	5	10	20	37	76
ICSI	tokens	26K	51K	105K	215K	436K
	meetings	3	4	8	15	33
ELITR	tokens	43K	87K	176K	353K	710K
	meetings	5	9	16	34	64
EPM	tokens	932K	1.86M	3.73M	7.46M	14.9M
	meetings	102	195	405	790	1554
MB	tokens	965K	1.93M	3.86M	7.72M	15.4M
	meetings	239	514	997	1968	3903

Table 9: **Natural scheme**: per-corpus token allocation and unique meeting count across budgets. Shares follow native corpus size (AMI 1.6%, ICSI 1.4%, ELITR 2.2%, EPM 46.6%, MB 48.3%), so the majority corpora dominate every budget. The scheme never oversamples: even at 32M, ICSI draws only 33 of its 43 meetings.

Corpus	Meetings	s1	s2	s15
AMI	43	276K	283K	292K
ICSI	43	578K	578K	578K
ELITR	43	464K	473K	457K
EPM	43	417K	402K	417K
MB	43	137K	155K	167K
Total	215	1.87M	1.89M	1.91M

Table 10: **Example-level scheme**: per-corpus token counts across three seeds (pruned). Each corpus contributes 43 meetings, but token counts differ because transcripts vary in length: ICSI ($\sim$4$\times$ MB). ICSI is identical across seeds as it has exactly 43 meetings; other corpora vary with the random draw.

evaluation inputs are near-complete on every corpus. Rate and severity also come apart. ICSI is truncated most often (53.8%) but least severely, retaining 82% of a median 15.8K-token transcript: its meetings sit just above the window, so crossing it costs little. EuroParlMin is truncated less often (23.9%) but far more heavily, retaining 65% on average, with a longest transcript of 209K tokens, twelve times the window. AMI is never truncated in any split.

This bears on the training target rather than on evaluation. For a truncated example the reference minute still covers the whole session while the input does not, so the model is trained to produce content its input does not contain. The effect is concentrated where severity is highest, and EuroParlMin records the lowest faithfulness of any corpus under both allocation schemes (Table 25); ICSI, truncated more often but far less severely, records the highest. We note the association without isolating it as the cause. Because test transcripts

Split	Corpus	n	#Trunc.	%	Median (tok)	Kept (%)
Train	AMI	98	0	0.0	7,558	—
	ICSI	43	22	51.2	15,667	81.9
	ELITR	84	28	33.3	12,105	71.3
	EPM	2065	529	25.6	9,077	64.7
	MB	5169	397	7.7	1,890	62.3
	All	**7459**	**976**	**13.1**	**3,261**	—
Dev	AMI	20	0	0.0	7,212	—
	ICSI	9	6	66.7	16,596	83.4
	ELITR	10	3	30.0	10,746	76.9
	EPM	187	9	4.8	517	60.8
	MB	861	58	6.7	1,589	63.4
	All	**1087**	**76**	**7.0**	**1,402**	—
Test	AMI	20	0	0.0	8,632	—
	ICSI	6	2	33.3	13,352	72.7
	ELITR	38	5	13.2	11,200	91.3
	EPM	242	1	0.4	394	67.5
	MB	862	31	3.6	1,294	62.6
	All	**1168**	**39**	**3.3**	**1,119**	—

Table 11: Transcript lengths and truncation per corpus and split (Mistral-7B tokeniser, unpruned transcripts). *#Trunc.* counts transcripts exceeding the available budget; *Kept* is the mean percentage of a transcript retained, computed over truncated transcripts only. Train and development use the training budget (window minus template, summary, and end-of-turn token); test uses the inference budget (window minus template). Truncation is far heavier in training (12.3% of train+dev) than at test (3.3%), and rate and severity diverge: ICSI is truncated most often but retains the most, EuroParlMin least often of the three affected corpora but retains the least.

are almost never truncated, the annotation guideline for faithfulness that addresses truncated tails (Appendix M.1, Task B) applies to very few items in practice.

E Multi-Seed Robustness

E.1 Balanced and Natural Schemes

The main grid (Section 4.2) runs on seed 42. To check that the per-domain crossover of RQ1 is a property of the allocation rather than of one draw, we repeat both schemes under the pruned condition at the 2M, 8M, and 32M budgets with two further seeds, $\{2, 15\}$, giving 12 additional runs. The seed governs adapter initialisation, dropout, and training order, as well as which meetings are sampled to fill each corpus's share of the budget; every other setting is unchanged (Table 18). Table 27 reports the resulting scores as mean $\pm$ std over the three seeds, across the full metric family.

The crossover reproduces. At 8M and 32M, on all five metrics without exception, balanced leads each of the three data-scarce domains (AMI, ICSI, ELITR) and natural leads the two data-rich ones (EPM, MB), so the macro average favours balanced

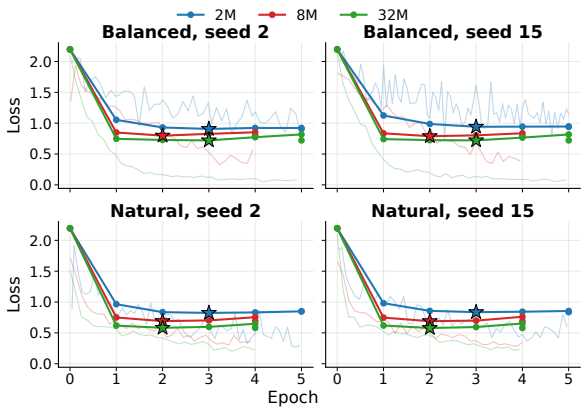


Figure 5: Multi-seed training trajectories (pruned, seeds 2 and 15): per-epoch training loss (faint) and development loss (bold, with markers) for the balanced (top) and natural (bottom) schemes at the 2M, 8M, and 32M budgets. The star marks the selected checkpoint, at the lowest development loss. As in the seed-42 grid (Figure 7), development loss bottoms out and then rises while training loss keeps falling, and each run stops two epochs after its minimum under patience-2 early stopping.

while the micro average favours natural, the same sign flip reported in Section 5.1. The gaps are large relative to seed variation on the minority domains: at 32M on ROUGE-Lsum, balanced leads AMI by 0.168, ICSI by 0.227, and ELITR by 0.189, against seed standard deviations below 0.07. At 2M the picture is weaker, as the earlier budget rows also show: the minority advantage is smaller, ELITR falls within seed variation, and the macro averages of the two schemes are indistinguishable. On the majority side, only MeetingBank is a consistent cost to balancing; the EuroParlMin gap narrows with budget and at 32M is smaller than its own seed spread, matching the erosion of natural's early EPM lead noted in Section 5.2.

E.2 Example-Level Baseline

The three example-level seeds {1,2,15} were selected to maximise the diversity of the sampled data on the corpora where sampling variance is real. For each candidate seed we simulated the builder's per-corpus selection of 43 meetings and measured the mean pairwise overlap on the large corpora (EPM, MB), where the pool greatly exceeds 43; the chosen trio minimises that overlap. Table 12 summarises the outcome: the large corpora are fully distinct across seeds, while the small corpora overlap only because the pool is too small to draw three disjoint 43-subsets.

Corpus	Pairwise overlap	Coverage across seeds
EPM	0% (fully distinct)	—
MB	0% (fully distinct)	—
AMI	(arithmetic)	83/98
ELITR	(arithmetic)	74/84
ICSI	100% (forced)	43/43

Table 12: Example-level seed selection {1,2,15}. Large corpora are fully distinct across seeds (genuine sampling variance); small-corpus overlap is arithmetically unavoidable (pool $\leq 3 \times 43$) but offset by near-complete coverage of the minority data. For ICSI ($|\text{pool}| = 43$) all meetings are used, so only training order varies across seeds.

F Model Selection

Table 13 reports zero-shot screening on ROUGE-Lsum and BERTScore-F1, per corpus, with macro (equal weight per corpus) and micro (weighted by number of meetings) averages; per-corpus and per-average bests are in **bold**. n is the number of test meetings, with ELITR collapsed to one prediction per meeting (scored as the maximum over its references). We screen five open models (Mistral-7B-Instruct-v0.3[7], Llama-3.1-8B-Instruct[8], Llama-3.2-3B-Instruct[9], Qwen3-8B[10], gemma-4-12B-it[11]) and one closed model (GPT-4o), the latter added because the third AutoMin challenge reports closed frontier models leading its shared task; we evaluate it under exactly the same prompt, decoding settings, and metrics as the open candidates.

Choice of metric. The two metrics disagree at the top: Gemma-4-12B leads ROUGE-Lsum macro (0.347), while Mistral-7B leads BERTScore-F1 on both averages (macro 0.833, micro 0.835); both rank Qwen3-8B weakest. We base selection on BERTScore-F1 rather than ROUGE-Lsum, as embedding overlap tracks semantic adequacy more closely than surface n-gram overlap.

Robust to the choice of average. Mistral-7B's BERTScore-F1 lead holds under both averaging schemes, which matters because the test sets are themselves imbalanced (ICSI 6 meetings, MB 862): the micro average is dominated by MB, while macro gives every corpus equal weight. Leading

[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[8] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[9] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[10] https://huggingface.co/Qwen/Qwen3-8B

[11] https://huggingface.co/google/gemma-4-12B-it

on both means the choice is not an artifact of MB's size—the very majority-domination this paper analyses.

Robust across domains. Mistral-7B also ranks first on three of five corpora (ELITR, EPM, MB), separates from the field on the two largest (EPM, MB) with non-overlapping confidence intervals, and has the highest per-corpus floor of all six candidates: its weakest-domain BERTScore-F1 is 0.824, above every other model's worst domain. In a multi-domain setting this floor matters—it means Mistral does not sacrifice any single domain. Llama-3.1-8B edges it on the two smallest corpora (AMI, ICSI), but within confidence intervals (unsurprising at n=20 and n=6) and without matching Mistral's floor or its separation on the large corpora.

GPT-4o does not lead under our protocol. GPT-4o ranks last of the six on macro BERTScore-F1 (0.815, against 0.833 for Mistral-7B and 0.826 for the 3B Llama) and on every average we report. This does not so much contradict AutoMin's finding as expose a mismatch between output format and reference form. The precision/recall split makes the mechanism visible: on MB, GPT-4o attains the highest BERTScore recall of any candidate alongside the lowest precision. It locates the reference content and then pads heavily around it, emitting section headers, attendee blocks, horizontal rules, and closing formalities that our references-short, flat, procedural minutes do not contain. Every unmatched token depresses precision, and F1 with it. We report all systems raw, applying no output cleaning to GPT-4o that we do not apply to the open models, so the comparison measures the models as prompted rather than as post-processed. We did not prompt GPT-4o to suppress markdown, since the screening applies one prompt to all candidates; a format-controlled prompt or post-processing would likely raise its precision, but GPT-4o was not a fine-tuning candidate and the choice does not affect the main results. GPT-4o is therefore not a ceiling in our setting, and we do not treat it as one.

Reproducibility of the closed model. We use `gpt-4o-2024-11-20` with temperature 0, top-p 1.0, and a fixed seed. OpenAI's `seed` parameter is best-effort: the run returned eight distinct `system_fingerprint` values, so outputs are not bit-reproducible and we do not describe the setting as greedy decoding.

Model	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
ROUGE-1							
Llama-3.2-3B	0.462	0.360	0.369	0.318	0.211	0.344	0.243
Llama-3.1-8B	**0.468**	0.370	0.370	0.329	0.217	0.351	0.250
Qwen3-8B	0.402	0.398	0.344	0.294	0.179	0.324	0.213
Mistral-7B	0.454	0.343	0.401	**0.343**	**0.245**	0.357	**0.274**
Gemma-4-12B	0.443	**0.430**	**0.425**	0.315	0.215	**0.366**	0.248
GPT-4o	0.395	0.403	0.392	0.307	0.194	0.338	0.228
ROUGE-2							
Llama-3.2-3B	0.120	0.075	0.083	0.140	0.109	0.106	0.115
Llama-3.1-8B	**0.132**	0.079	0.088	**0.147**	0.115	0.112	0.121
Qwen3-8B	0.102	0.090	0.082	0.129	0.098	0.100	0.104
Mistral-7B	0.121	0.065	**0.101**	0.144	**0.131**	**0.112**	**0.132**
Gemma-4-12B	0.110	**0.098**	0.095	0.131	0.113	0.109	0.116
GPT-4o	0.095	0.087	0.087	0.129	0.101	0.100	0.106
ROUGE-L							
Llama-3.2-3B	0.219	**0.170**	0.178	0.207	0.159	0.187	0.171
Llama-3.1-8B	**0.220**	0.166	0.170	0.213	0.164	0.187	0.175
Qwen3-8B	0.180	0.152	0.150	0.187	0.136	0.161	0.148
Mistral-7B	0.207	0.145	0.181	**0.222**	**0.183**	**0.188**	**0.191**
Gemma-4-12B	0.191	0.153	**0.182**	0.201	0.158	0.177	0.168
GPT-4o	0.172	0.146	0.169	0.195	0.142	0.165	0.155
ROUGE-Lsum							
Llama-3.2-3B	0.439	0.345	0.359	0.307	0.188	0.328	0.223
Llama-3.1-8B	**0.444**	0.353	0.358	0.319	0.196	0.334	0.232
Qwen3-8B	0.385	0.383	0.334	0.285	0.164	0.310	0.199
Mistral-7B	0.435	0.325	0.389	**0.330**	**0.220**	0.340	**0.252**
Gemma-4-12B	0.420	**0.410**	**0.409**	0.303	0.194	**0.347**	0.229
GPT-4o	0.377	0.389	0.382	0.298	0.176	0.324	0.212
BERTScore-F1							
Llama-3.2-3B	0.841	0.821	0.831	0.816	0.819	0.826	0.819
Llama-3.1-8B	**0.844**	**0.828**	0.830	0.822	0.831	0.831	0.830
Qwen3-8B	0.816	0.818	0.824	0.815	0.821	0.819	0.820
Mistral-7B	0.838	0.824	**0.840**	**0.825**	**0.837**	**0.833**	**0.835**
Gemma-4-12B	0.818	0.816	0.828	0.818	0.824	0.821	0.823
GPT-4o	0.808	0.809	0.832	0.814	0.813	0.815	0.814

Table 13: Zero-shot summarisation quality across five meeting corpora. Macro is the unweighted mean over corpora; Micro is weighted by test-set size (counts in italics). ELITR scores take the maximum over its multiple references. Best per column in **bold**.

Llama-3.2-3B as a capacity control. We carry Llama-3.2-3B forward as a lower-capacity control for RQ5. Model size is not monotonic with quality in this screen (7B Mistral $>$ 12B Gemma and $>$ GPT-4o on BERTScore-F1; 3B Llama is competitive with the 8B models), so it is a genuine capacity probe rather than a uniformly weaker baseline.

G Per-Domain Fine-Tuned Baselines

This appendix reports the per-domain baselines for RQ6. One QLoRA adapter is trained per corpus on the full pruned training set of that corpus, with no token cap, mixing, or oversampling, so that every training meeting is seen exactly once per epoch. The development set is likewise restricted to the same corpus, so early stopping and checkpoint selection are driven by that domain alone. All other settings follow Table 18 with seed 42. At test time each adapter generates minutes only for the test meetings of its own domain, under the decoding settings of Section 3.5. The five outputs are then

concatenated into one 1,168-meeting system and scored with the procedure of Section 3.6, so that per-domain, macro, and micro results are directly comparable with Table 2.

G.1 Setup and Training Trajectories

The three minority corpora yield 0.58-0.94M tokens and 3-7 optimiser steps per epoch at the effective batch size of 16, against 130 and 324 steps for EPM and MB. Figure 6 shows the resulting trajectories. Two regimes are visible. The EPM and MB adapters behave like the mixture runs of Figure 7: development loss reaches its minimum at epoch 3 and 2 respectively, then rises, and patience-2 early stopping terminates training. The AMI, ICSI, and ELITR adapters never reach this regime. Their development loss decreases monotonically through epoch 5 and flattens only as the cosine schedule drives the learning rate to zero, so the final checkpoint is selected in all three cases. Under the fixed protocol these corpora receive 15-35 updates in total, an order of magnitude fewer than balanced allocation at 32M provides them through oversampling.

The minority adapters also generate unreliably. Where an adapter terminates, its minutes are of reference length, but on 7 of 20 AMI and 2 of 6 ICSI meetings it repeats a single sentence until the 2,048-token cap, and 2 of 38 ELITR outputs are empty; balanced allocation reaches the cap on none of these two domains and produces no empty output. The EPM and MB adapters show no such behaviour: their median lengths (247 and 73 tokens) and cap rates (17/242 and 47/862) match those of balanced allocation (249 and 74; 29/242 and 50/862). The ICSI outputs reproduce the reference section structure with near-identical text under every heading. These adapters have learned the form of the minutes but not their content, which accounts for scores below the untuned model. Decoding was not modified, so all systems are scored as generated.

G.2 Full Results

Table 14 reports the per-domain baselines across the complete metric family with bootstrap 95% confidence intervals, in the layout of Table 22. Against balanced allocation at 32M, the baseline is below the interval on ICSI and ELITR on every metric, and its AMI interval excludes the balanced point estimate on all ROUGE variants. On MB the baseline interval excludes the balanced value in the other direction, and on EPM the two overlap. The macro interval (ROUGE-Lsum 0.37-0.42) lies wholly below the balanced macro (0.500), the micro interval (0.63-0.66) wholly above it (0.615).

H Summary Lengths (Reference and Generated)

Table 15 reports the token-length distribution of the test-set reference summaries, measured with the Mistral-7B tokenizer (for ELITR, all annotator references are counted). Reference summaries are short to moderate in length across all domains, with medians between 87 and 664 tokens. The 2048-token generation budget covers every test reference in AMI, ICSI, ELITR, and MeetingBank in full, and 98.8% of test references overall. The remaining references belong to EuroParlMin and arise from a corpus formatting artefact: a small number of parliamentary minutes are provided as undivided full-chapter documents rather than per-session summaries. Such documents are not representative meeting minutes, and the most extreme of them occur only in the training and development splits. We therefore set the generation budget generously above the typical reference length rather than to these atypical outliers, which keeps decoding efficient without affecting the vast majority of references.

Table 16 reports the length of summaries generated by the six screened candidates on the **test set** of each corpus, complementing the reference-length distribution above (Table 15). Lengths are each model's own generated-token count, so they are measured with that model's tokeniser and are indicative rather than strictly commensurable across rows. ELITR is collapsed to one prediction per meeting (n=38).

The means sit well above the per-domain medians on EPM and MB, reflecting the same right-skewed, heavy-tailed length distributions seen in the references (Table 15); the large standard deviations on these two corpora are therefore a property of the heterogeneous inputs rather than unstable generation, and a small number of outputs approach the 2048-token cap. This is consistent with the cap being non-binding for over 98% of generations (Section 4.2).

The table also quantifies the length mismatch behind GPT-4o's ranking (Appendix F). On MeetingBank, whose references average 101 tokens, GPT-4o emits 575 on average, the largest ratio of gen-

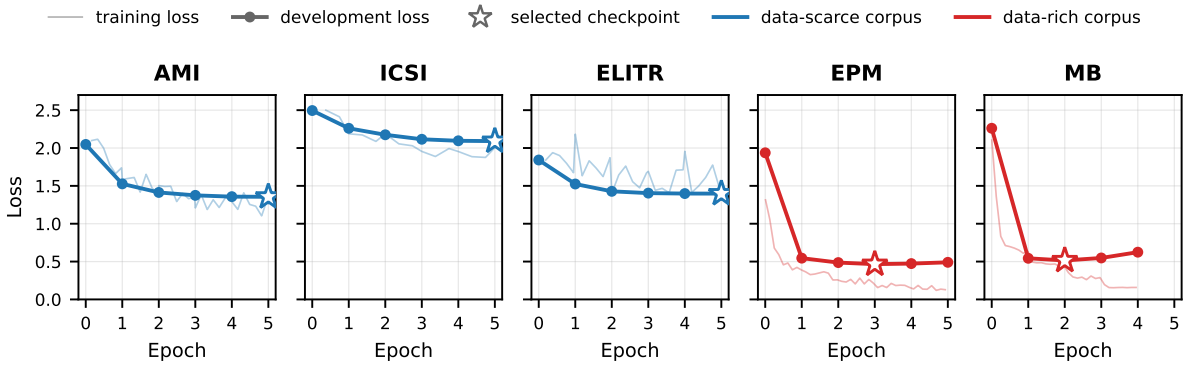


Figure 6: Per-domain baselines: per-epoch training loss (faint) and development loss (bold, with markers) for each adapter, measured on the development set of the same corpus (pruned, seed 42). The star marks the selected checkpoint. The data-rich adapters (EPM, MB) reach a development-loss minimum and are terminated by patience-2 early stopping, as in Figure 7; the data-scarce adapters (AMI, ICSI, ELITR) are still improving at the five-epoch ceiling.

Metric	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
ROUGE-1	$0.394_{[0.30,0.48]}$	$0.224_{[0.14,0.30]}$	$0.130_{[0.09,0.17]}$	$0.570_{[0.54,0.60]}$	$0.720_{[0.70,0.74]}$	$0.408_{[0.38,0.43]}$	$0.662_{[0.65,0.68]}$
ROUGE-2	$0.168_{[0.12,0.22]}$	$0.049_{[0.03,0.07]}$	$0.041_{[0.03,0.06]}$	$0.432_{[0.41,0.46]}$	$0.632_{[0.61,0.65]}$	$0.264_{[0.25,0.28]}$	$0.561_{[0.54,0.58]}$
ROUGE-L	$0.234_{[0.18,0.29]}$	$0.138_{[0.10,0.18]}$	$0.100_{[0.07,0.13]}$	$0.511_{[0.48,0.54]}$	$0.696_{[0.68,0.71]}$	$0.336_{[0.32,0.35]}$	$0.628_{[0.61,0.64]}$
ROUGE-Lsum	$0.382_{[0.29,0.47]}$	$0.210_{[0.13,0.29]}$	$0.126_{[0.09,0.17]}$	$0.564_{[0.53,0.59]}$	$0.703_{[0.69,0.72]}$	$0.397_{[0.37,0.42]}$	$0.647_{[0.63,0.66]}$
BERTScore	$0.868_{[0.86,0.88]}$	$0.817_{[0.79,0.84]}$	$0.772_{[0.70,0.82]}$	$0.897_{[0.89,0.90]}$	$0.939_{[0.94,0.94]}$	$0.858_{[0.84,0.87]}$	$0.923_{[0.92,0.93]}$

Table 14: Per-domain fine-tuned baselines (pruned, seed 42) across the complete metric family, with bootstrap 95% CIs in subscripts (10,000 resamples over test meetings; the macro CI resamples within each domain and averages the domain means). Each adapter is trained on its full training set: 0.66M (AMI), 0.58M (ICSI), 0.94M (ELITR), 19.7M (EPM) and 20.4M (MB) tokens per epoch. Column italics are test-set meeting counts. The balanced and natural blocks for comparison are in Table 22.

Corpus	N	Mean	P95	P99	Max
AMI	20	376	517	582	599
ICSI	6	628	918	938	944
MB	862	101	230	313	959
ELITR (test)	55	539	1122	1240	1272
ELITR (test2)	10	907	1618	1808	1856
ELITR (test2023)	12	769	1330	1381	1394
EPM	242	472	2400	4823	7253

Table 15: Test-set reference-summary lengths in tokens (Mistral-7B tokenizer). N is the number of test references; ELITR is split by its three test sets, with all annotator references counted. The 2048-token generation budget covers 98.8% of references in full; the longer ones are undivided full-chapter EuroParlMin documents rather than per-session minutes.

eration to reference of any candidate. The excess is structural rather than substantive: section headers, attendee blocks, and closing formalities that the reference minutes do not contain. Every such token is unmatched, which depresses precision and the F1 that follows from it, while leaving recall intact. (Reference lengths in Table 15 count all annotator references; here we report one generation per meeting.)

I Corpus Construction

This appendix gives the per-corpus construction, splitting, and unit details summarised in Section 4.1, and the verification of the ELITR reference-selection criterion of Section 3.5.

Schema and splits. All five corpora are cast into a uniform `(id, source, summary, split)` schema, where `source` is a speaker-labelled transcript and `summary` the reference minute. We adopt canonical splits where they exist-the AMI scenario-only partition, the ICSI six-meeting test set, and the official MB and EPM splits-and a seeded, released train/development division for ICSI (seed 42), whose canonical train/development membership is not standardised. Split sizes are given in Table 17.

Transcript and reference form. AMI and ICSI transcripts are reconstructed as speaker-labelled

Model	Dataset	n	Mean $\pm$ SD	95% CI
Llama-3.2-3B	AMI	20	457.7 $\pm$ 65.1	[427.2, 488.2]
	ICSI	6	420.5 $\pm$ 43.8	[374.5, 466.5]
	ELITR	38	567.1 $\pm$ 261.0	[481.3, 652.9]
	EPM	242	314.3 $\pm$ 157.4	[294.4, 334.3]
	MB	862	413.9 $\pm$ 284.4	[394.9, 433.0]
Llama-3.1-8B	AMI	20	443.2 $\pm$ 74.5	[408.4, 478.1]
	ICSI	6	551.5 $\pm$ 140.1	[404.4, 698.6]
	ELITR	38	733.7 $\pm$ 410.9	[598.6, 868.7]
	EPM	242	304.2 $\pm$ 177.6	[281.7, 326.7]
	MB	862	367.6 $\pm$ 159.1	[357.0, 378.3]
Mistral-7B	AMI	20	520.6 $\pm$ 162.6	[444.5, 596.7]
	ICSI	6	409.0 $\pm$ 126.0	[276.8, 541.2]
	ELITR	38	739.3 $\pm$ 386.3	[612.4, 866.3]
	EPM	242	313.9 $\pm$ 240.1	[283.5, 344.3]
	MB	862	374.9 $\pm$ 194.2	[361.9, 387.9]
Qwen3-8B	AMI	20	964.0 $\pm$ 183.9	[877.9, 1050.1]
	ICSI	6	941.7 $\pm$ 201.6	[730.1, 1153.2]
	ELITR	38	1317.6 $\pm$ 411.5	[1182.4, 1452.9]
	EPM	242	493.3 $\pm$ 345.8	[449.5, 537.1]
	MB	862	680.9 $\pm$ 379.3	[655.6, 706.3]
Gemma-4-12B	AMI	20	781.1 $\pm$ 148.2	[711.8, 850.4]
	ICSI	6	805.7 $\pm$ 79.2	[722.6, 888.8]
	ELITR	38	798.4 $\pm$ 108.9	[762.7, 834.2]
	EPM	242	361.7 $\pm$ 296.6	[324.1, 399.2]
	MB	862	507.6 $\pm$ 253.7	[490.6, 524.5]
GPT-4o	AMI	20	929.2 $\pm$ 152.4	[857.9, 1000.5]
	ICSI	6	965.0 $\pm$ 210.7	[743.9, 1186.1]
	ELITR	38	1002.5 $\pm$ 150.6	[953.0, 1052.0]
	EPM	242	431.1 $\pm$ 268.3	[397.1, 465.1]
	MB	862	575.4 $\pm$ 224.2	[560.4, 590.4]

Table 16: Generated meeting-minute length in tokens (mean $\pm$ standard deviation, with 95% confidence interval for the mean) per model and corpus. ELITR is collapsed to one prediction per meeting (n=38).

turns from their dialogue-act annotations with standard disfluency removal, and their structured, section-headered summaries are retained. MB, ELITR, and EPM are used as distributed, preserving their native speaker labelling and reference form. The unit of summarisation differs by corpus, reflecting each corpus's native structure: submeetings for AMI, whole meetings for ICSI and ELITR, and agenda-item segments for MB and EPM.

ELITR test sets and multiple references. ELITR supplies three distinct test sets (TEST/TEST2/TEST2023), which we keep separate following the corpus authors' intent; their sizes are summed in Table 17 (18+8+12 = 38 meetings). ELITR also provides multiple independent reference minutes per meeting. For training-mixture construction we consolidate each meeting to a single reference; for zero-shot evaluation we instead retain all references and score each prediction as the maximum over its references, so ELITR contributes 38 meetings to the evaluation.

Corpus	Domain	Train	Dev	Test
AMI	product meetings	98	20	20
ICSI	academic meetings	43	9	6
MB	council meetings	5169	861	862
ELITR	project meetings	84	10	38*
EPM	parliamentary sessions	2065	187	242
Total		7459	1087	1168

Table 17: Final train/development/test partitions. Splits are canonical where available, seeded-and-released for ICSI (seed 42). ELITR figures are for the English portion; its Test column sums three test sets (18+8+12). *ELITR test counts unique meetings (38), each carrying multiple reference minutes over which evaluation takes the maximum.

Verification of ELITR reference selection. The single-reference consolidation selects, per meeting, the reference with the highest mean ROUGE-Lsum recall against the other references. Because this rewards content coverage it correlates with length, so we verified its behaviour rather than assuming it. Across the 70 multi-reference meetings (all splits), the selected reference was the longest candidate in 68 cases; in the two exceptions the candidates were within three words of one another and selection was determined by content overlap. The criterion therefore favours the most comprehensive reference while remaining sensitive to content rather than length alone, and avoids the arbitrariness of a fixed-annotator or original-minutes heuristic.

J Fine-tuning Hyperparameters

For a frozen 4-bit weight matrix $W_0 \in \mathbb{R}^{d\times k}$, QLoRA learns a low-rank update $\Delta W = BA$, with $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d,k)$, so the forward pass is $h = W_0 x + \frac{\alpha}{r} BAx$ with scaling factor α. The loss is cross-entropy over summary positions $S = \{p{+}1, \ldots, n\}$ only, with prompt tokens masked:

$$\mathcal{L} = -\frac{1}{|S|} \sum_{i\in S} \log P(y_i \mid y_{<i}, x; \theta), \qquad (1)$$

where x is the transcript and θ the trainable adapter parameters. This trains ≈83.9M parameters, 1.14% of the model. Table 18 lists the full configuration, applied identically across all schemes, budgets, and conditions.

K Training Trajectories and Checkpoint Selection

Figure 7 shows per-epoch training and development loss for all twenty runs, and Table 19 reports the

Hyperparameter	Value
Quantisation	4-bit NF4, double-quant, bf16 compute
LoRA rank r	32
LoRA α	16
LoRA dropout	0.05
Adapted modules	q, k, v, o, gate, up, down
Optimiser	paged AdamW (8-bit)
Learning rate	1×10^{-4}
LR schedule	cosine, 3% warmup
Weight decay	0.01
Max gradient norm	0.3
Effective batch size	16 (1×16 accum.)
Max sequence length	16,384
Max epochs	5 (early stopping, patience 2)
Checkpoint selection	lowest dev loss
Seed	42 (main grid); 1, 2, 15† (robustness)

Table 18: QLoRA fine-tuning hyperparameters, applied identically across all schemes, budgets, and conditions. †Seeds 2 and 15 repeat both allocation schemes at 2M/8M/32M (Appendix E.1); seeds 1, 2, and 15 are used for the example-level baseline (Appendix E.2).

resulting checkpoint selection. Together they justify the early-stopping and best-checkpoint choices of Section 3.4. Development loss falls to a minimum and then rises as the model overfits, so the selected checkpoint (starred in the figure) is not the last. The minimum occurs at epoch 2 for the mid-range budgets and at epoch 3 at 2M; under balanced allocation the largest budget (32M) also selects epoch 3, the one mild departure from the trend of earlier convergence at larger budgets. Every run halts exactly two epochs after its minimum, confirming that patience-2 early stopping governed termination, while training loss continues to fall past the selected epoch. Pre-training development loss is ≈ 2.09 (unpruned) and ≈ 2.20 (pruned). These values determine only which checkpoint we evaluate; final quality is measured on the held-out test set (Section 3.6), and development loss is never used as a quality signal.

Scheme	Cond.	2M	4M	8M	16M	32M
Balanced	pruned	0.906 (3)	0.826 (2)	0.801 (2)	0.760 (2)	0.728 (3)
	unpruned	0.959 (3)	0.845 (2)	0.792 (2)	0.777 (2)	0.723 (3)
Natural	pruned	0.833 (3)	0.749 (2)	0.691 (2)	0.636 (2)	0.583 (2)
	unpruned	0.835 (3)	0.760 (2)	0.696 (2)	0.640 (2)	0.587 (2)

Table 19: Checkpoint selection for all runs. Each cell gives the best development loss and, in parentheses, the epoch at which it occurred (the selected checkpoint). Development loss is used only for checkpoint selection, never as a quality measure (Section 3.6).

L LLM Judge Protocol

ROUGE-Lsum and BERTScore measure surface and embedding overlap, but can diverge from content coverage and faithfulness in structured-output domains (Section 5). To assess summary quality at the level of individual claims, we adopt the bidirectional atomic fact-checking framework of (Zhou et al., 2026), adapting it to our {transcript, reference minute, generated minute} setting. Following their protocol, we decompose minutes into atomic facts and use an LLM judge to check factual entailment, scoring three dimensions: completeness (the fraction of reference-minute facts captured by the generated minute), faithfulness (the fraction of generated-minute facts supported by the transcript), and conciseness (the fraction of generated-minute facts that are salient with respect to the reference minute). Completeness and conciseness are the recall and precision of the generated minute against the reference; faithfulness is its precision against the source meeting.

We depart from (Zhou et al., 2026) in one respect where they extract a gold key-fact set from the source conversation via a dedicated human-LLM pipeline, we take the human-authored reference minute as the salience oracle, decomposing it directly into key facts. This suits our setting, where every meeting already carries an expert reference, and avoids a separate annotation stage; the trade-off is that any correct content absent from reference is neither credited (to completeness) nor considered salient (lowering conciseness). All judgments use a fixed rubric with explicit adjudication rules applied identically across mixtures, budgets, and domains, and a judge model isolated from the fine-tuned systems to avoid self-preference bias. The judge (Qwen2.5-72B-Instruct-AWQ) runs under greedy decoding on a single NVIDIA H100; all other inference and fine-tuning use a single A100 (Section 4.3)

M Human Validation of the LLM Judge

We validate the fact-level judge (Section 3.6) against two human annotators. The study measures how often the annotators agree with the judge's per-fact labels and with each other; the human–human figure is the ceiling that makes the human–judge figure interpretable, since a judge that agrees with humans as often as they agree with each other is performing at human level.

Design. Three choices follow. (i) Annotators adjudicate the judge's own atomic-fact lists rather than generating their own, so labels align fact-for-fact and a per-fact statistic is defined. (ii) We vali-

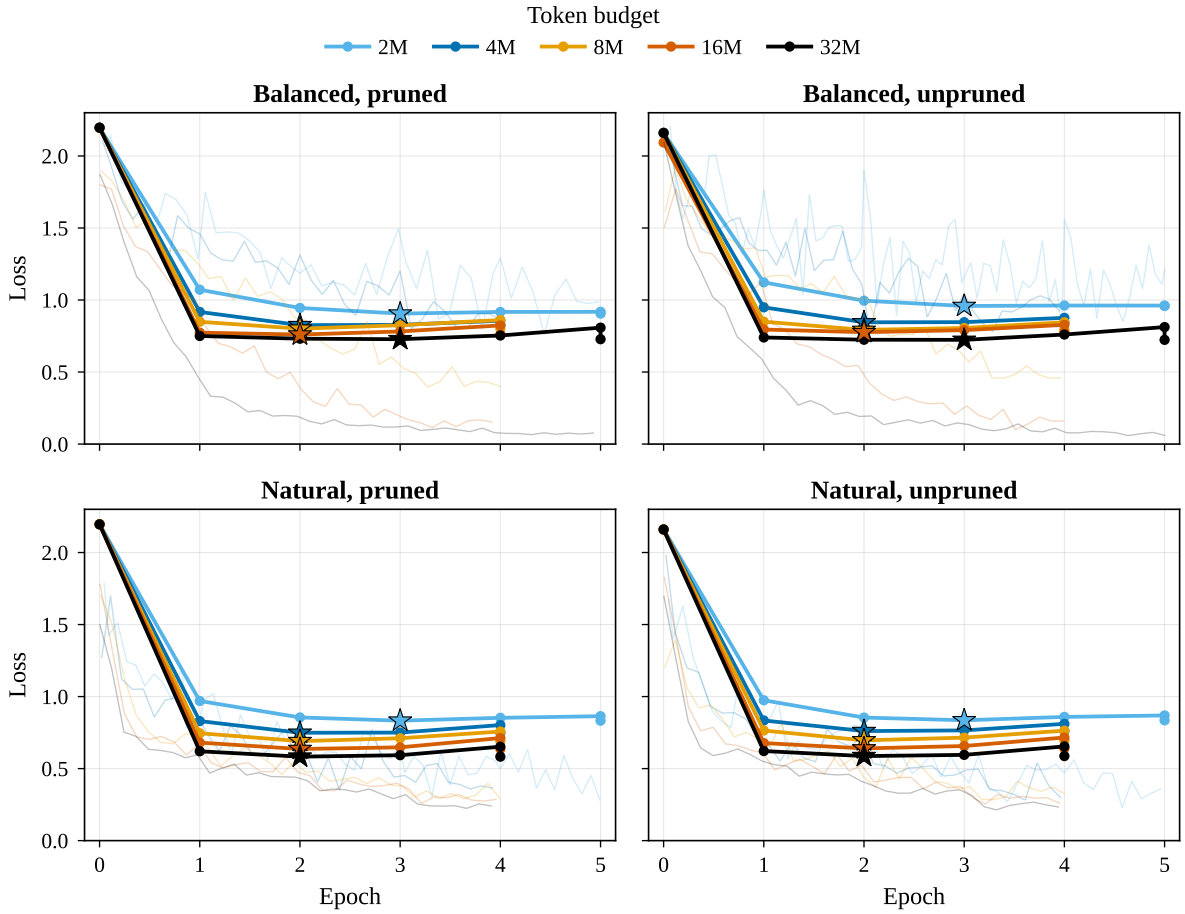


Figure 7: Per-epoch training loss (faint) and development loss (bold, with markers) for the balanced (top) and natural (bottom) schemes, pruned (left) and unpruned (right), on seed 42. The star marks the selected checkpoint, at the lowest development loss. Development loss bottoms out and then rises while training loss keeps falling; each run stops two epochs after its minimum under patience-2 early stopping.

date on `balanced-32M` only, as the object of study is the judge, not the allocation schemes. (iii) Facts are stratified by judge label: the judge's base rates are skewed (faithfulness is ∼79% `supported`), so under random sampling a constant-majority annotator would appear to agree ∼79% of the time, whereas an even split makes such agreement worth only ∼50%.

Sampling. From the `balanced-32M` run we drew 30 test meetings (six per domain, seed 42) and sampled facts within each (dimension × domain) stratum: up to 30 judge-positive and 30 judge-negative facts per domain for completeness and conciseness, and up to 25 each for faithfulness. Strata smaller than the target were taken in full; six of thirty cells were capped by MeetingBank, whose short minutes yield few facts, and further caps reflect genuine scarcity in the judge's output (only 12 `not_supported` facts exist across all six ICSI meetings, the model being ∼92% faithful there). This yields 741 items (257 completeness, 268 conciseness, 216 faithfulness). Because the caps bind hardest on the majority corpora, the sample is deliberately not weighted to the test distribution: the agreement figures are pooled across domains and stratified by label, the right estimand for validating an instrument that must work on every domain, not an estimate of agreement on the deployed test set.

Annotation. Both annotators labelled every item, as κ requires paired decisions. Each saw the fact and only the material its task permits (the generated minute for completeness, the transcript for faithfulness, the reference for conciseness), with the judge's label stripped and items shuffled within meeting. Faithfulness items were grouped by meeting so each transcript is read once. A one-line reason was required on every item; annotators did not confer. The judge labels compared here come

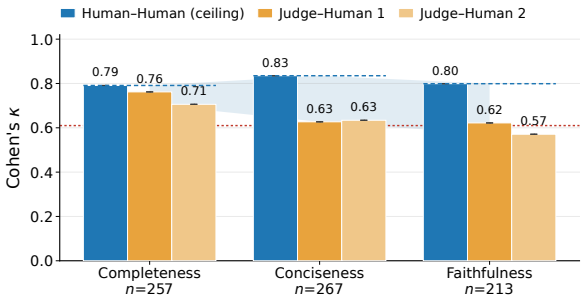


Figure 8: Inter-rater agreement on 741 judge-labelled facts from `balanced-32M` (Cohen's κ, bootstrap 95% CIs, 4,000 resamples). Human–human agreement is the ceiling (dashed rule); shading marks the judge's shortfall against it. The red dotted line is $\kappa = 0.61$, the threshold for substantial agreement. The judge reaches the human band on completeness but falls below it on conciseness and faithfulness.

	Compl.	**Concis.**	**Faith.**
n	257	267	213
Cohen's κ, pairwise (bootstrap 95% CI)			
H1–H2	0.791 [.71,.86]	0.835 [.77,.90]	0.799 [.71,.87]
Judge–H1	0.762 [.68,.84]	0.627 [.54,.71]	0.622 [.52,.72]
Judge–H2	0.706 [.61,.79]	0.634 [.54,.72]	0.571 [.47,.67]
Raw agreement, pairwise (%)			
H1–H2	91%	92%	91%
Judge–H1	89%	81%	81%
Judge–H2	86%	82%	78%
Judge error direction (vs. H1=H2 consensus, counts)			
over-accepts	18	34	32
over-rejects	2	5	2
% lenient	90%	87%	94%

Table 20: Agreement between the two annotators (H1, H2) and the LLM judge, per dimension. Top: chance-corrected agreement (Cohen's κ) for each rater pair; H1–H2 is the human ceiling. Middle: the same pairs as raw percentage agreement. Bottom: among items where the two annotators agree (a consensus ground truth), the number of judge labels that are too generous (over-accept: judge positive, humans negative) versus too strict (over-reject), showing the judge errs by over-accepting in 87-94% of cases.

from re-scoring each fact individually, matching the annotators' condition; the aggregate scores in Table 3 use the batched judge.

Results. Four items blank in one annotator are excluded pairwise ($n = 257/267/213$). Inter-annotator agreement is substantial on all three dimensions ($\kappa = 0.79/0.84/0.80$; 91–92% raw), establishing a reliable ground truth. Against it the judge reaches $\kappa = 0.76/0.71$ on completeness, within the human band, and $0.63/0.63$ and $0.62/0.57$ on conciseness and faithfulness, substantial but below the ceiling (Figure 8, Table 20). The errors are directional: restricted to items where both annotators agree, the judge over-accepts in 18/34/32 cases against 2/5/2 in the other direction, so 87–94% of its disagreements are over-acceptance. The absolute scores in Table 3 are therefore upper bounds; because the offset is applied by one judge under one rubric to every system and budget, the balanced-versus-natural and 2M-versus-32M comparisons are unaffected.

M.1 Annotation guidelines

The guidelines below were given to both annotators verbatim. Every item is a binary decision; annotators never write, summarise, or rate on a scale.

Materials. The transcript is the raw record of what was said. The reference minute is the summary a human wrote; it is treated as authoritative and defines what was worth recording. The generated minute is what the system produced. An atomic fact is a single indivisible claim extracted from a minute: "Sarah presented the Q3 budget and the team agreed to postpone the launch" becomes two facts. Annotators judge atomic facts, not whole minutes.

Task A: Completeness. Given one fact from the reference minute and the full generated minute: does the generated minute contain this fact? Label `captured` if the generated minute states the fact or states something from which it follows directly and unambiguously; different wording is fine. Label `not_captured` if the fact is absent, contradicted, or only vaguely gestured at. Losing the core value is `not_captured`: mentioning that a budget increase was agreed does not capture "the increase was 12%". Extra detail never hurts. Numbers, dates, and names must match. Use only the generated minute: do not consult the transcript, and do not use outside knowledge to fill a gap the generated minute left.

Task B: Faithfulness. Given one fact from the generated minute and the meeting transcript: does the transcript support this fact? Label `supported` if the fact is explicitly stated or follows directly and unambiguously from what was said; `not_supported` if it is not said, contradicted, or requires a leap the transcript does not license. Judge on the transcript alone—do not mark a fact supported because it is plausible. Over-specificity is

not_supported (transcript "around 25 euros", fact "exactly 25 euros"), whereas hedging a precise figure is supported. Attribution must itself be supported, not merely the content. Proposals are not decisions: a fact asserting a decision the transcript only floated is not_supported. Long transcripts are truncated; if a fact refers to something that would have occurred after the transcript ends, mark not_supported, because only what is visible can be judged.

Task C: Conciseness (salience). Given one fact from the generated minute and the reference minute: did the human who wrote the reference think this was worth recording? Label salient if the reference states the fact or states something from which it follows; non_salient if it is absent from the reference. This task is not a truth judgement. A fact may be perfectly true of the meeting and still be non_salient, because the minute-taker chose not to record it: if the reference does not mention how long the meeting ran, "the meeting lasted two hours" is non_salient even if it did. A fact more specific than the reference on a point the reference does make is still salient; a fact introducing content the reference does not touch at all is not. A fact that merely restates another fact already matched to the same reference content is non_salient.

Procedure. Each task is completed in full before the next is begun, because the rules are deliberately different and switching between them causes errors. For each item the annotator reads the fact, reads the permitted material, chooses a label, and writes a one-line reason; the reasons are how disagreements are understood afterwards. Annotators do not discuss items with each other, do not look anything up, and do not leave items blank; where genuinely torn, they choose the label they would defend and write "borderline" in the reason.

N Full Per-Domain Results

This appendix reports the complete per-domain grid over the full metric family (ROUGE-1/2/L/Lsum and BERTScore-F1), of which Section 5 presents the ROUGE-Lsum+BERTScore view that carries the finding.

Condition	Metric	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
Pruned	R-1	$0.339_{\pm.044}$	$0.310_{\pm.036}$	$0.237_{\pm.016}$	$0.277_{\pm.045}$	$0.532_{\pm.008}$	$0.339_{\pm.018}$	$0.465_{\pm.009}$
	R-2	$0.137_{\pm.016}$	$0.075_{\pm.010}$	$0.060_{\pm.007}$	$0.165_{\pm.036}$	$0.421_{\pm.007}$	$0.172_{\pm.011}$	$0.350_{\pm.007}$
	R-L	$0.206_{\pm.028}$	$0.170_{\pm.017}$	$0.137_{\pm.003}$	$0.223_{\pm.040}$	$0.498_{\pm.008}$	$0.247_{\pm.012}$	$0.423_{\pm.008}$
	R-Lsum	$0.328_{\pm.044}$	$0.292_{\pm.036}$	$0.229_{\pm.014}$	$0.271_{\pm.044}$	$0.505_{\pm.008}$	$0.325_{\pm.017}$	$0.443_{\pm.009}$
	BERT	$0.861_{\pm.004}$	$0.834_{\pm.003}$	$0.826_{\pm.006}$	$0.833_{\pm.013}$	$0.898_{\pm.000}$	$0.850_{\pm.004}$	$0.881_{\pm.003}$
Unpruned	R-1	$0.370_{\pm.056}$	$0.272_{\pm.031}$	$0.220_{\pm.011}$	$0.273_{\pm.041}$	$0.534_{\pm.008}$	$0.334_{\pm.014}$	$0.466_{\pm.009}$
	R-2	$0.155_{\pm.030}$	$0.060_{\pm.007}$	$0.058_{\pm.003}$	$0.164_{\pm.030}$	$0.426_{\pm.006}$	$0.173_{\pm.010}$	$0.353_{\pm.008}$
	R-L	$0.223_{\pm.032}$	$0.157_{\pm.016}$	$0.129_{\pm.004}$	$0.222_{\pm.034}$	$0.502_{\pm.006}$	$0.247_{\pm.012}$	$0.425_{\pm.009}$
	R-Lsum	$0.357_{\pm.056}$	$0.256_{\pm.028}$	$0.213_{\pm.009}$	$0.267_{\pm.041}$	$0.508_{\pm.007}$	$0.320_{\pm.016}$	$0.445_{\pm.009}$
	BERTScore	$0.863_{\pm.004}$	$0.830_{\pm.005}$	$0.824_{\pm.006}$	$0.829_{\pm.010}$	$0.899_{\pm.001}$	$0.849_{\pm.004}$	$0.881_{\pm.003}$

Table 21: Example-level results across all five metrics: mean$_{\pm\text{std}}$ over seeds 1/2/15, per condition. Column italics are test-set meeting counts. Seed variation is small for aggregate measures (Macro/Micro std $\leq$0.018); minority domains with few test meetings (ICSI, AMI) show larger per-seed spread. R-1/2/L = ROUGE-1/2/L.

Metric	Budget	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
Balanced (equal-token) allocation								
ROUGE-1	2M	$0.389_{[0.30,0.47]}$	$0.272_{[0.24,0.30]}$	$0.159_{[0.12,0.20]}$	$0.227_{[0.20,0.25]}$	$0.585_{[0.57,0.60]}$	$0.326_{[0.31,0.35]}$	$0.492_{[0.48,0.51]}$
	4M	$0.359_{[0.27,0.45]}$	$0.224_{[0.12,0.33]}$	$0.166_{[0.13,0.21]}$	$0.326_{[0.29,0.36]}$	$0.587_{[0.57,0.61]}$	$0.333_{[0.30,0.36]}$	$0.514_{[0.50,0.53]}$
	8M	$0.465_{[0.40,0.52]}$	$0.443_{[0.40,0.49]}$	$0.251_{[0.20,0.30]}$	$0.443_{[0.41,0.47]}$	$0.626_{[0.61,0.64]}$	$0.446_{[0.43,0.46]}$	$0.572_{[0.56,0.59]}$
	16M	$0.522_{[0.50,0.55]}$	$0.440_{[0.41,0.47]}$	$0.383_{[0.34,0.42]}$	$0.392_{[0.36,0.42]}$	$0.649_{[0.63,0.67]}$	$0.477_{[0.46,0.49]}$	$0.584_{[0.57,0.60]}$
	32M	$0.516_{[0.49,0.54]}$	$0.465_{[0.42,0.50]}$	$0.379_{[0.35,0.41]}$	$0.554_{[0.52,0.58]}$	$0.670_{[0.65,0.69]}$	$0.517_{[0.50,0.53]}$	$0.633_{[0.62,0.65]}$
ROUGE-2	2M	$0.162_{[0.12,0.21]}$	$0.054_{[0.04,0.06]}$	$0.039_{[0.03,0.05]}$	$0.145_{[0.13,0.16]}$	$0.480_{[0.46,0.50]}$	$0.176_{[0.17,0.19]}$	$0.389_{[0.37,0.40]}$
	4M	$0.150_{[0.10,0.20]}$	$0.048_{[0.03,0.07]}$	$0.039_{[0.03,0.05]}$	$0.220_{[0.19,0.25]}$	$0.480_{[0.46,0.50]}$	$0.188_{[0.18,0.20]}$	$0.404_{[0.39,0.42]}$
	8M	$0.179_{[0.14,0.22]}$	$0.121_{[0.10,0.14]}$	$0.064_{[0.05,0.08]}$	$0.315_{[0.29,0.34]}$	$0.516_{[0.50,0.53]}$	$0.239_{[0.23,0.25]}$	$0.452_{[0.44,0.47]}$
	16M	$0.193_{[0.17,0.22]}$	$0.100_{[0.08,0.11]}$	$0.112_{[0.09,0.13]}$	$0.271_{[0.25,0.30]}$	$0.553_{[0.53,0.57]}$	$0.245_{[0.24,0.26]}$	$0.471_{[0.46,0.49]}$
	32M	$0.195_{[0.17,0.22]}$	$0.115_{[0.10,0.14]}$	$0.099_{[0.08,0.11]}$	$0.407_{[0.38,0.43]}$	$0.572_{[0.55,0.59]}$	$0.278_{[0.27,0.29]}$	$0.514_{[0.50,0.53]}$
ROUGE-L	2M	$0.243_{[0.19,0.30]}$	$0.158_{[0.15,0.17]}$	$0.101_{[0.08,0.12]}$	$0.194_{[0.17,0.22]}$	$0.553_{[0.53,0.57]}$	$0.250_{[0.24,0.26]}$	$0.457_{[0.44,0.47]}$
	4M	$0.211_{[0.16,0.26]}$	$0.118_{[0.08,0.16]}$	$0.106_{[0.09,0.13]}$	$0.281_{[0.25,0.31]}$	$0.553_{[0.53,0.57]}$	$0.254_{[0.24,0.27]}$	$0.474_{[0.46,0.49]}$
	8M	$0.256_{[0.22,0.29]}$	$0.198_{[0.18,0.22]}$	$0.143_{[0.12,0.17]}$	$0.395_{[0.37,0.42]}$	$0.590_{[0.57,0.61]}$	$0.316_{[0.30,0.33]}$	$0.527_{[0.51,0.54]}$
	16M	$0.289_{[0.27,0.31]}$	$0.184_{[0.17,0.20]}$	$0.194_{[0.17,0.21]}$	$0.340_{[0.31,0.37]}$	$0.623_{[0.60,0.64]}$	$0.326_{[0.32,0.34]}$	$0.543_{[0.53,0.56]}$
	32M	$0.274_{[0.26,0.29]}$	$0.192_{[0.18,0.21]}$	$0.184_{[0.17,0.20]}$	$0.486_{[0.46,0.51]}$	$0.641_{[0.62,0.66]}$	$0.355_{[0.35,0.36]}$	$0.586_{[0.57,0.60]}$
ROUGE-Lsum	2M	$0.376_{[0.29,0.45]}$	$0.258_{[0.22,0.29]}$	$0.152_{[0.12,0.19]}$	$0.222_{[0.20,0.25]}$	$0.562_{[0.54,0.58]}$	$0.314_{[0.29,0.33]}$	$0.473_{[0.46,0.49]}$
	4M	$0.347_{[0.26,0.43]}$	$0.211_{[0.12,0.32]}$	$0.159_{[0.12,0.20]}$	$0.320_{[0.29,0.35]}$	$0.561_{[0.54,0.58]}$	$0.320_{[0.29,0.35]}$	$0.492_{[0.48,0.51]}$
	8M	$0.449_{[0.38,0.51]}$	$0.419_{[0.38,0.47]}$	$0.244_{[0.20,0.29]}$	$0.438_{[0.41,0.47]}$	$0.598_{[0.58,0.62]}$	$0.429_{[0.41,0.45]}$	$0.550_{[0.54,0.56]}$
	16M	$0.502_{[0.48,0.53]}$	$0.420_{[0.40,0.45]}$	$0.373_{[0.33,0.41]}$	$0.386_{[0.36,0.42]}$	$0.630_{[0.61,0.65]}$	$0.462_{[0.45,0.48]}$	$0.568_{[0.55,0.58]}$
	32M	$0.495_{[0.47,0.52]}$	$0.445_{[0.40,0.49]}$	$0.367_{[0.33,0.40]}$	$0.547_{[0.52,0.57]}$	$0.648_{[0.63,0.67]}$	$0.500_{[0.49,0.51]}$	$0.615_{[0.60,0.63]}$
BERTScore	2M	$0.861_{[0.85,0.87]}$	$0.827_{[0.82,0.84]}$	$0.806_{[0.80,0.82]}$	$0.832_{[0.83,0.84]}$	$0.910_{[0.91,0.91]}$	$0.847_{[0.84,0.85]}$	$0.890_{[0.89,0.89]}$
	4M	$0.863_{[0.85,0.87]}$	$0.818_{[0.80,0.84]}$	$0.818_{[0.81,0.83]}$	$0.838_{[0.83,0.85]}$	$0.912_{[0.91,0.92]}$	$0.850_{[0.84,0.85]}$	$0.892_{[0.89,0.90]}$
	8M	$0.870_{[0.86,0.88]}$	$0.851_{[0.84,0.86]}$	$0.835_{[0.83,0.84]}$	$0.865_{[0.86,0.87]}$	$0.918_{[0.91,0.92]}$	$0.868_{[0.86,0.87]}$	$0.903_{[0.90,0.91]}$
	16M	$0.878_{[0.87,0.88]}$	$0.847_{[0.84,0.86]}$	$0.847_{[0.84,0.85]}$	$0.855_{[0.85,0.86]}$	$0.925_{[0.92,0.93]}$	$0.870_{[0.87,0.87]}$	$0.907_{[0.90,0.91]}$
	32M	$0.875_{[0.87,0.88]}$	$0.852_{[0.85,0.86]}$	$0.846_{[0.84,0.85]}$	$0.887_{[0.88,0.89]}$	$0.928_{[0.92,0.93]}$	$0.877_{[0.87,0.88]}$	$0.915_{[0.91,0.92]}$
Natural (proportional) allocation								
ROUGE-1	2M	$0.344_{[0.28,0.40]}$	$0.137_{[0.08,0.20]}$	$0.189_{[0.16,0.22]}$	$0.280_{[0.25,0.31]}$	$0.613_{[0.59,0.63]}$	$0.312_{[0.29,0.33]}$	$0.523_{[0.51,0.54]}$
	4M	$0.275_{[0.21,0.35]}$	$0.087_{[0.06,0.11]}$	$0.130_{[0.10,0.16]}$	$0.454_{[0.42,0.48]}$	$0.641_{[0.62,0.66]}$	$0.317_{[0.30,0.34]}$	$0.577_{[0.56,0.59]}$
	8M	$0.242_{[0.17,0.32]}$	$0.116_{[0.08,0.14]}$	$0.116_{[0.09,0.15]}$	$0.524_{[0.50,0.55]}$	$0.670_{[0.65,0.69]}$	$0.333_{[0.31,0.35]}$	$0.611_{[0.60,0.62]}$
	16M	$0.298_{[0.21,0.39]}$	$0.171_{[0.08,0.27]}$	$0.159_{[0.12,0.20]}$	$0.530_{[0.50,0.56]}$	$0.702_{[0.69,0.72]}$	$0.372_{[0.34,0.40]}$	$0.639_{[0.62,0.65]}$
	32M	$0.410_{[0.33,0.48]}$	$0.297_{[0.23,0.36]}$	$0.212_{[0.17,0.26]}$	$0.539_{[0.51,0.57]}$	$0.713_{[0.70,0.73]}$	$0.434_{[0.41,0.46]}$	$0.653_{[0.64,0.67]}$
ROUGE-2	2M	$0.110_{[0.08,0.14]}$	$0.027_{[0.01,0.04]}$	$0.040_{[0.03,0.05]}$	$0.179_{[0.16,0.20]}$	$0.507_{[0.49,0.53]}$	$0.172_{[0.16,0.18]}$	$0.414_{[0.40,0.43]}$
	4M	$0.094_{[0.06,0.13]}$	$0.017_{[0.01,0.02]}$	$0.028_{[0.02,0.04]}$	$0.319_{[0.29,0.34]}$	$0.538_{[0.52,0.56]}$	$0.199_{[0.19,0.21]}$	$0.466_{[0.45,0.48]}$
	8M	$0.085_{[0.05,0.12]}$	$0.018_{[0.02,0.02]}$	$0.023_{[0.02,0.03]}$	$0.375_{[0.35,0.40]}$	$0.569_{[0.55,0.59]}$	$0.214_{[0.20,0.22]}$	$0.500_{[0.48,0.51]}$
	16M	$0.123_{[0.08,0.17]}$	$0.033_{[0.01,0.06]}$	$0.046_{[0.03,0.06]}$	$0.393_{[0.36,0.42]}$	$0.600_{[0.58,0.62]}$	$0.239_{[0.23,0.25]}$	$0.528_{[0.51,0.54]}$
	32M	$0.162_{[0.12,0.20]}$	$0.065_{[0.05,0.08]}$	$0.058_{[0.04,0.08]}$	$0.402_{[0.37,0.43]}$	$0.621_{[0.60,0.64]}$	$0.262_{[0.25,0.27]}$	$0.547_{[0.53,0.56]}$
ROUGE-L	2M	$0.202_{[0.17,0.23]}$	$0.094_{[0.06,0.13]}$	$0.115_{[0.10,0.13]}$	$0.232_{[0.21,0.26]}$	$0.579_{[0.56,0.60]}$	$0.244_{[0.23,0.26]}$	$0.483_{[0.47,0.50]}$
	4M	$0.172_{[0.13,0.22]}$	$0.075_{[0.05,0.10]}$	$0.085_{[0.07,0.10]}$	$0.398_{[0.37,0.43]}$	$0.611_{[0.59,0.63]}$	$0.268_{[0.26,0.28]}$	$0.539_{[0.52,0.55]}$
	8M	$0.160_{[0.11,0.21]}$	$0.081_{[0.06,0.09]}$	$0.075_{[0.06,0.10]}$	$0.459_{[0.43,0.48]}$	$0.639_{[0.62,0.66]}$	$0.283_{[0.27,0.30]}$	$0.572_{[0.56,0.59]}$
	16M	$0.186_{[0.13,0.24]}$	$0.114_{[0.07,0.16]}$	$0.105_{[0.08,0.13]}$	$0.471_{[0.44,0.50]}$	$0.670_{[0.65,0.69]}$	$0.309_{[0.29,0.33]}$	$0.600_{[0.59,0.61]}$
	32M	$0.247_{[0.20,0.29]}$	$0.160_{[0.13,0.19]}$	$0.134_{[0.11,0.16]}$	$0.483_{[0.45,0.51]}$	$0.688_{[0.67,0.71]}$	$0.342_{[0.33,0.36]}$	$0.617_{[0.60,0.63]}$
ROUGE-Lsum	2M	$0.331_{[0.27,0.39]}$	$0.129_{[0.08,0.18]}$	$0.184_{[0.15,0.22]}$	$0.274_{[0.25,0.30]}$	$0.586_{[0.57,0.60]}$	$0.301_{[0.28,0.32]}$	$0.502_{[0.49,0.52]}$
	4M	$0.263_{[0.20,0.33]}$	$0.080_{[0.06,0.11]}$	$0.124_{[0.09,0.16]}$	$0.448_{[0.42,0.48]}$	$0.618_{[0.60,0.64]}$	$0.306_{[0.29,0.32]}$	$0.558_{[0.54,0.57]}$
	8M	$0.230_{[0.16,0.31]}$	$0.107_{[0.08,0.13]}$	$0.107_{[0.08,0.14]}$	$0.516_{[0.49,0.54]}$	$0.646_{[0.63,0.66]}$	$0.321_{[0.30,0.34]}$	$0.592_{[0.58,0.61]}$
	16M	$0.287_{[0.20,0.37]}$	$0.158_{[0.07,0.25]}$	$0.153_{[0.11,0.19]}$	$0.523_{[0.49,0.55]}$	$0.678_{[0.66,0.69]}$	$0.360_{[0.33,0.39]}$	$0.620_{[0.61,0.63]}$
	32M	$0.393_{[0.32,0.46]}$	$0.281_{[0.21,0.35]}$	$0.204_{[0.16,0.25]}$	$0.534_{[0.50,0.56]}$	$0.694_{[0.68,0.71]}$	$0.421_{[0.40,0.44]}$	$0.637_{[0.62,0.65]}$
BERTScore	2M	$0.851_{[0.84,0.86]}$	$0.814_{[0.79,0.83]}$	$0.806_{[0.80,0.81]}$	$0.831_{[0.82,0.84]}$	$0.915_{[0.91,0.92]}$	$0.843_{[0.84,0.85]}$	$0.892_{[0.89,0.90]}$
	4M	$0.843_{[0.83,0.85]}$	$0.779_{[0.77,0.79]}$	$0.796_{[0.79,0.81]}$	$0.872_{[0.86,0.88]}$	$0.922_{[0.92,0.93]}$	$0.843_{[0.84,0.85]}$	$0.905_{[0.90,0.91]}$
	8M	$0.835_{[0.82,0.85]}$	$0.801_{[0.79,0.81]}$	$0.815_{[0.80,0.83]}$	$0.882_{[0.88,0.89]}$	$0.925_{[0.92,0.93]}$	$0.852_{[0.85,0.86]}$	$0.910_{[0.91,0.91]}$
	16M	$0.852_{[0.84,0.87]}$	$0.806_{[0.78,0.82]}$	$0.819_{[0.81,0.83]}$	$0.884_{[0.88,0.89]}$	$0.935_{[0.93,0.94]}$	$0.859_{[0.85,0.86]}$	$0.919_{[0.92,0.92]}$
	32M	$0.865_{[0.85,0.88]}$	$0.833_{[0.83,0.84]}$	$0.837_{[0.83,0.85]}$	$0.889_{[0.88,0.90]}$	$0.938_{[0.93,0.94]}$	$0.872_{[0.87,0.88]}$	$0.923_{[0.92,0.93]}$

Table 22: Full per-domain results (**pruned**, seed 42) across the complete metric family (ROUGE-1/2/L/Lsum and BERTScore-F1), with bootstrap 95% CIs in subscripts. Rows are grouped by allocation scheme (Balanced vs. Natural) and, within each scheme, by metric; each metric block reports the five-budget ladder (2-32M). Column italics are test-set meeting counts. The per-domain crossover of RQ1 is visible in every metric: Balanced leads on the minority domains (AMI/ICSI/ELITR) while Natural leads on the majority domains (EPM/MB), with the gap widest at small budgets.

Metric	Budget	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
				Balanced (equal-token) allocation				
ROUGE-1	2M	$0.281_{[0.20,0.37]}$	$0.113_{[0.08,0.14]}$	$0.143_{[0.11,0.18]}$	$0.332_{[0.31,0.35]}$	$0.559_{[0.54,0.58]}$	$0.285_{[0.27,0.31]}$	$0.491_{[0.48,0.51]}$
	4M	$0.318_{[0.23,0.41]}$	$0.260_{[0.17,0.34]}$	$0.174_{[0.13,0.22]}$	$0.189_{[0.17,0.21]}$	$0.615_{[0.60,0.63]}$	$0.311_{[0.28,0.34]}$	$0.505_{[0.49,0.52]}$
	8M	$0.393_{[0.31,0.47]}$	$0.367_{[0.27,0.44]}$	$0.241_{[0.20,0.28]}$	$0.406_{[0.37,0.44]}$	$0.637_{[0.62,0.65]}$	$0.409_{[0.38,0.43]}$	$0.571_{[0.56,0.58]}$
	16M	$0.508_{[0.47,0.54]}$	$0.449_{[0.40,0.50]}$	$0.306_{[0.26,0.35]}$	$0.354_{[0.32,0.38]}$	$0.664_{[0.65,0.68]}$	$0.456_{[0.44,0.47]}$	$0.585_{[0.57,0.60]}$
	32M	$0.521_{[0.50,0.54]}$	$0.459_{[0.41,0.50]}$	$0.382_{[0.34,0.42]}$	$0.541_{[0.51,0.57]}$	$0.678_{[0.66,0.69]}$	$0.516_{[0.50,0.53]}$	$0.636_{[0.62,0.65]}$
ROUGE-2	2M	$0.120_{[0.08,0.17]}$	$0.020_{[0.01,0.03]}$	$0.029_{[0.02,0.04]}$	$0.196_{[0.18,0.21]}$	$0.453_{[0.43,0.47]}$	$0.164_{[0.15,0.17]}$	$0.378_{[0.36,0.39]}$
	4M	$0.144_{[0.09,0.20]}$	$0.065_{[0.04,0.09]}$	$0.045_{[0.03,0.06]}$	$0.120_{[0.11,0.14]}$	$0.513_{[0.49,0.53]}$	$0.177_{[0.16,0.19]}$	$0.408_{[0.39,0.42]}$
	8M	$0.158_{[0.12,0.20]}$	$0.084_{[0.05,0.11]}$	$0.062_{[0.05,0.08]}$	$0.286_{[0.26,0.31]}$	$0.533_{[0.51,0.55]}$	$0.225_{[0.21,0.24]}$	$0.458_{[0.44,0.47]}$
	16M	$0.189_{[0.16,0.21]}$	$0.101_{[0.09,0.12]}$	$0.081_{[0.07,0.10]}$	$0.249_{[0.22,0.27]}$	$0.563_{[0.54,0.58]}$	$0.237_{[0.23,0.25]}$	$0.473_{[0.46,0.49]}$
	32M	$0.196_{[0.17,0.22]}$	$0.109_{[0.10,0.12]}$	$0.106_{[0.09,0.12]}$	$0.389_{[0.36,0.41]}$	$0.578_{[0.56,0.60]}$	$0.276_{[0.27,0.28]}$	$0.515_{[0.50,0.53]}$
ROUGE-L	2M	$0.176_{[0.12,0.24]}$	$0.081_{[0.06,0.10]}$	$0.095_{[0.08,0.12]}$	$0.263_{[0.25,0.28]}$	$0.526_{[0.51,0.55]}$	$0.228_{[0.21,0.24]}$	$0.449_{[0.43,0.46]}$
	4M	$0.203_{[0.14,0.26]}$	$0.148_{[0.10,0.19]}$	$0.112_{[0.09,0.14]}$	$0.159_{[0.14,0.18]}$	$0.587_{[0.57,0.61]}$	$0.242_{[0.23,0.26]}$	$0.474_{[0.46,0.49]}$
	8M	$0.231_{[0.18,0.28]}$	$0.183_{[0.14,0.21]}$	$0.133_{[0.11,0.15]}$	$0.359_{[0.33,0.39]}$	$0.606_{[0.59,0.62]}$	$0.302_{[0.29,0.32]}$	$0.531_{[0.52,0.55]}$
	16M	$0.274_{[0.25,0.30]}$	$0.188_{[0.18,0.20]}$	$0.161_{[0.14,0.18]}$	$0.309_{[0.28,0.34]}$	$0.635_{[0.62,0.65]}$	$0.313_{[0.30,0.32]}$	$0.544_{[0.53,0.56]}$
	32M	$0.289_{[0.27,0.31]}$	$0.200_{[0.19,0.21]}$	$0.186_{[0.17,0.20]}$	$0.473_{[0.45,0.50]}$	$0.648_{[0.63,0.67]}$	$0.359_{[0.35,0.37]}$	$0.589_{[0.57,0.60]}$
ROUGE-Lsum	2M	$0.272_{[0.19,0.36]}$	$0.104_{[0.08,0.13]}$	$0.135_{[0.11,0.17]}$	$0.324_{[0.30,0.34]}$	$0.531_{[0.51,0.55]}$	$0.273_{[0.25,0.29]}$	$0.469_{[0.45,0.48]}$
	4M	$0.306_{[0.22,0.40]}$	$0.247_{[0.16,0.32]}$	$0.166_{[0.13,0.21]}$	$0.185_{[0.16,0.21]}$	$0.594_{[0.58,0.61]}$	$0.300_{[0.27,0.33]}$	$0.488_{[0.47,0.50]}$
	8M	$0.378_{[0.30,0.46]}$	$0.352_{[0.25,0.42]}$	$0.235_{[0.20,0.28]}$	$0.400_{[0.37,0.43]}$	$0.614_{[0.60,0.63]}$	$0.396_{[0.37,0.42]}$	$0.551_{[0.54,0.57]}$
	16M	$0.486_{[0.45,0.52]}$	$0.423_{[0.38,0.47]}$	$0.296_{[0.25,0.34]}$	$0.349_{[0.32,0.38]}$	$0.642_{[0.63,0.66]}$	$0.439_{[0.42,0.46]}$	$0.566_{[0.55,0.58]}$
	32M	$0.499_{[0.48,0.52]}$	$0.436_{[0.39,0.48]}$	$0.369_{[0.33,0.40]}$	$0.535_{[0.51,0.56]}$	$0.655_{[0.64,0.67]}$	$0.499_{[0.48,0.51]}$	$0.617_{[0.60,0.63]}$
BERTScore	2M	$0.855_{[0.84,0.87]}$	$0.804_{[0.80,0.81]}$	$0.805_{[0.80,0.81]}$	$0.837_{[0.83,0.84]}$	$0.904_{[0.90,0.91]}$	$0.841_{[0.84,0.84]}$	$0.886_{[0.88,0.89]}$
	4M	$0.857_{[0.84,0.87]}$	$0.829_{[0.81,0.85]}$	$0.823_{[0.81,0.83]}$	$0.822_{[0.82,0.83]}$	$0.917_{[0.91,0.92]}$	$0.850_{[0.84,0.85]}$	$0.893_{[0.89,0.90]}$
	8M	$0.870_{[0.86,0.88]}$	$0.839_{[0.83,0.85]}$	$0.835_{[0.83,0.84]}$	$0.860_{[0.85,0.87]}$	$0.921_{[0.92,0.92]}$	$0.865_{[0.86,0.87]}$	$0.904_{[0.90,0.91]}$
	16M	$0.874_{[0.87,0.88]}$	$0.851_{[0.85,0.86]}$	$0.835_{[0.83,0.84]}$	$0.846_{[0.84,0.85]}$	$0.926_{[0.92,0.93]}$	$0.866_{[0.86,0.87]}$	$0.905_{[0.90,0.91]}$
	32M	$0.877_{[0.87,0.88]}$	$0.849_{[0.84,0.86]}$	$0.844_{[0.84,0.85]}$	$0.883_{[0.88,0.89]}$	$0.928_{[0.92,0.93]}$	$0.876_{[0.87,0.88]}$	$0.915_{[0.91,0.92]}$
				Natural (proportional) allocation				
ROUGE-1	2M	$0.177_{[0.13,0.22]}$	$0.087_{[0.03,0.15]}$	$0.150_{[0.12,0.19]}$	$0.438_{[0.41,0.46]}$	$0.620_{[0.60,0.64]}$	$0.294_{[0.28,0.31]}$	$0.557_{[0.54,0.57]}$
	4M	$0.243_{[0.16,0.33]}$	$0.186_{[0.09,0.28]}$	$0.129_{[0.10,0.16]}$	$0.370_{[0.34,0.40]}$	$0.625_{[0.61,0.64]}$	$0.311_{[0.28,0.34]}$	$0.547_{[0.53,0.56]}$
	8M	$0.266_{[0.18,0.35]}$	$0.223_{[0.13,0.32]}$	$0.139_{[0.10,0.18]}$	$0.373_{[0.34,0.41]}$	$0.667_{[0.65,0.68]}$	$0.334_{[0.31,0.36]}$	$0.580_{[0.56,0.59]}$
	16M	$0.280_{[0.20,0.36]}$	$0.135_{[0.08,0.21]}$	$0.158_{[0.12,0.19]}$	$0.502_{[0.47,0.54]}$	$0.697_{[0.68,0.71]}$	$0.354_{[0.33,0.38]}$	$0.629_{[0.61,0.64]}$
	32M	$0.341_{[0.25,0.43]}$	$0.235_{[0.14,0.33]}$	$0.183_{[0.15,0.22]}$	$0.500_{[0.47,0.53]}$	$0.719_{[0.70,0.74]}$	$0.396_{[0.37,0.42]}$	$0.647_{[0.63,0.66]}$
ROUGE-2	2M	$0.050_{[0.03,0.07]}$	$0.016_{[0.00,0.03]}$	$0.034_{[0.02,0.05]}$	$0.287_{[0.27,0.31]}$	$0.502_{[0.48,0.52]}$	$0.178_{[0.17,0.19]}$	$0.432_{[0.42,0.45]}$
	4M	$0.088_{[0.05,0.13]}$	$0.028_{[0.01,0.04]}$	$0.027_{[0.02,0.04]}$	$0.253_{[0.23,0.28]}$	$0.521_{[0.50,0.54]}$	$0.183_{[0.17,0.19]}$	$0.439_{[0.42,0.45]}$
	8M	$0.104_{[0.06,0.15]}$	$0.046_{[0.02,0.07]}$	$0.026_{[0.02,0.04]}$	$0.272_{[0.24,0.30]}$	$0.564_{[0.54,0.58]}$	$0.202_{[0.19,0.21]}$	$0.475_{[0.46,0.49]}$
	16M	$0.110_{[0.07,0.15]}$	$0.028_{[0.01,0.05]}$	$0.046_{[0.04,0.06]}$	$0.378_{[0.35,0.41]}$	$0.606_{[0.59,0.63]}$	$0.234_{[0.22,0.25]}$	$0.529_{[0.51,0.54]}$
	32M	$0.147_{[0.10,0.20]}$	$0.053_{[0.03,0.08]}$	$0.042_{[0.03,0.05]}$	$0.380_{[0.35,0.41]}$	$0.625_{[0.61,0.64]}$	$0.249_{[0.24,0.26]}$	$0.544_{[0.53,0.56]}$
ROUGE-L	2M	$0.111_{[0.08,0.14]}$	$0.059_{[0.02,0.10]}$	$0.097_{[0.08,0.12]}$	$0.371_{[0.35,0.40]}$	$0.581_{[0.56,0.60]}$	$0.244_{[0.23,0.26]}$	$0.511_{[0.50,0.53]}$
	4M	$0.161_{[0.12,0.21]}$	$0.119_{[0.07,0.16]}$	$0.094_{[0.08,0.11]}$	$0.322_{[0.29,0.35]}$	$0.590_{[0.57,0.61]}$	$0.257_{[0.24,0.27]}$	$0.508_{[0.49,0.52]}$
	8M	$0.169_{[0.12,0.22]}$	$0.137_{[0.09,0.18]}$	$0.084_{[0.07,0.10]}$	$0.334_{[0.30,0.37]}$	$0.637_{[0.62,0.65]}$	$0.272_{[0.26,0.29]}$	$0.545_{[0.53,0.56]}$
	16M	$0.171_{[0.12,0.22]}$	$0.094_{[0.06,0.13]}$	$0.113_{[0.09,0.14]}$	$0.457_{[0.43,0.49]}$	$0.675_{[0.66,0.69]}$	$0.302_{[0.29,0.32]}$	$0.600_{[0.58,0.61]}$
	32M	$0.215_{[0.16,0.28]}$	$0.143_{[0.10,0.18]}$	$0.114_{[0.09,0.13]}$	$0.459_{[0.43,0.49]}$	$0.692_{[0.67,0.71]}$	$0.325_{[0.31,0.34]}$	$0.614_{[0.60,0.63]}$
ROUGE-Lsum	2M	$0.171_{[0.13,0.21]}$	$0.083_{[0.02,0.15]}$	$0.141_{[0.11,0.18]}$	$0.430_{[0.40,0.46]}$	$0.591_{[0.57,0.61]}$	$0.283_{[0.26,0.30]}$	$0.533_{[0.52,0.55]}$
	4M	$0.233_{[0.16,0.31]}$	$0.178_{[0.08,0.27]}$	$0.122_{[0.10,0.15]}$	$0.364_{[0.34,0.39]}$	$0.597_{[0.58,0.62]}$	$0.299_{[0.27,0.32]}$	$0.525_{[0.51,0.54]}$
	8M	$0.254_{[0.17,0.34]}$	$0.204_{[0.11,0.29]}$	$0.134_{[0.10,0.17]}$	$0.368_{[0.33,0.40]}$	$0.644_{[0.63,0.66]}$	$0.321_{[0.29,0.35]}$	$0.562_{[0.55,0.58]}$
	16M	$0.267_{[0.19,0.35]}$	$0.123_{[0.07,0.19]}$	$0.147_{[0.12,0.18]}$	$0.497_{[0.46,0.53]}$	$0.680_{[0.66,0.70]}$	$0.343_{[0.32,0.37]}$	$0.615_{[0.60,0.63]}$
	32M	$0.328_{[0.24,0.42]}$	$0.219_{[0.13,0.31]}$	$0.177_{[0.14,0.21]}$	$0.495_{[0.46,0.53]}$	$0.698_{[0.68,0.72]}$	$0.383_{[0.36,0.41]}$	$0.630_{[0.62,0.64]}$
BERTScore	2M	$0.817_{[0.81,0.82]}$	$0.769_{[0.75,0.79]}$	$0.823_{[0.81,0.83]}$	$0.866_{[0.86,0.87]}$	$0.916_{[0.91,0.92]}$	$0.838_{[0.83,0.84]}$	$0.900_{[0.90,0.90]}$
	4M	$0.832_{[0.82,0.85]}$	$0.800_{[0.78,0.82]}$	$0.788_{[0.78,0.80]}$	$0.853_{[0.85,0.86]}$	$0.918_{[0.91,0.92]}$	$0.838_{[0.83,0.84]}$	$0.898_{[0.89,0.90]}$
	8M	$0.839_{[0.82,0.85]}$	$0.817_{[0.79,0.84]}$	$0.790_{[0.78,0.80]}$	$0.850_{[0.84,0.86]}$	$0.928_{[0.92,0.93]}$	$0.845_{[0.84,0.85]}$	$0.905_{[0.90,0.91]}$
	16M	$0.851_{[0.84,0.86]}$	$0.805_{[0.79,0.82]}$	$0.834_{[0.82,0.85]}$	$0.878_{[0.87,0.89]}$	$0.935_{[0.93,0.94]}$	$0.861_{[0.86,0.87]}$	$0.918_{[0.92,0.92]}$
	32M	$0.862_{[0.85,0.87]}$	$0.813_{[0.79,0.84]}$	$0.832_{[0.82,0.84]}$	$0.879_{[0.87,0.89]}$	$0.939_{[0.94,0.94]}$	$0.865_{[0.86,0.87]}$	$0.921_{[0.92,0.92]}$

Table 23: Full per-domain results (**unpruned**, seed 42) across the complete metric family (ROUGE-1/2/L/Lsum and BERTScore-F1), with bootstrap 95% CIs in subscripts. Rows are grouped by allocation scheme (Balanced vs. Natural) and, within each scheme, by metric; each metric block reports the five-budget ladder (2–32M). Column italics are test-set meeting counts. This is the unpruned companion to the pruned grid; comparing the two at matched (scheme, budget, metric) cells isolates the effect of transcript pruning (RQ3).

Metric	Budget	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
Balanced (equal-token) allocation								
ROUGE-1	2M	0.105	0.090	0.096	0.219	0.530	0.208	0.442
	8M	0.324	0.185	0.128	0.256	0.587	0.296	0.497
	32M	0.497	0.424	0.238	0.224	0.613	0.399	0.518
ROUGE-2	2M	0.033	0.015	0.018	0.123	0.416	0.121	0.334
	8M	0.137	0.040	0.028	0.170	0.474	0.170	0.388
	32M	0.174	0.102	0.060	0.155	0.505	0.199	0.410
ROUGE-L	2M	0.077	0.077	0.072	0.172	0.496	0.178	0.405
	8M	0.198	0.111	0.083	0.213	0.550	0.231	0.457
	32M	0.265	0.182	0.127	0.195	0.578	0.269	0.477
ROUGE-Lsum	2M	0.098	0.080	0.091	0.213	0.502	0.197	0.420
	8M	0.312	0.174	0.122	0.252	0.560	0.284	0.476
	32M	0.476	0.402	0.229	0.220	0.587	0.383	0.496
BERTScore-F1	2M	0.821	0.788	0.786	0.813	0.897	0.821	0.874
	8M	0.861	0.823	0.817	0.838	0.910	0.850	0.891
	32M	0.872	0.844	0.836	0.842	0.915	0.862	0.896
Natural (proportional) allocation								
ROUGE-1	2M	0.161	0.031	0.042	0.268	0.563	0.213	0.475
	8M	0.121	0.086	0.087	0.432	0.642	0.274	0.569
	32M	0.163	0.131	0.111	0.538	0.699	0.328	0.634
ROUGE-2	2M	0.039	0.004	0.007	0.169	0.458	0.135	0.374
	8M	0.035	0.014	0.016	0.296	0.536	0.179	0.458
	32M	0.064	0.026	0.019	0.392	0.599	0.220	0.525
ROUGE-L	2M	0.101	0.027	0.030	0.218	0.533	0.182	0.442
	8M	0.087	0.076	0.067	0.374	0.609	0.243	0.531
	32M	0.110	0.101	0.071	0.474	0.672	0.286	0.599
ROUGE-Lsum	2M	0.152	0.029	0.040	0.263	0.539	0.204	0.456
	8M	0.113	0.079	0.082	0.426	0.615	0.263	0.547
	32M	0.156	0.124	0.105	0.531	0.679	0.319	0.617
BERTScore-F1	2M	0.800	0.750	0.759	0.834	0.906	0.810	0.884
	8M	0.821	0.784	0.779	0.867	0.922	0.835	0.904
	32M	0.831	0.800	0.805	0.885	0.935	0.851	0.918

Table 24: Llama-3.2-3B (RQ5): full per-domain results across the complete metric family, for both allocation schemes at the three shared budgets (2M/8M/32M; pruned, seed 42). Column italics are test-set meeting counts. The per-domain crossover of RQ1 reproduces here: balanced leads the minority domains (AMI/ICSI/ELITR) while natural leads the majority domains (EPM/MB).

System	Dim.	AMI	ICSI	ELITR	EPM	MB	Macro	Micro
balanced 2M	Comple.	$0.370_{[0.29,0.46]}$	$0.147_{[0.07,0.22]}$	$0.077_{[0.05,0.11]}$	$0.310_{[0.28,0.34]}$	$0.478_{[0.45,0.50]}$	$0.276_{[0.25,0.30]}$	$0.427_{[0.41,0.45]}$
	Faithf.	$0.884_{[0.83,0.93]}$	$0.906_{[0.87,0.94]}$	$0.784_{[0.68,0.88]}$	$0.502_{[0.46,0.55]}$	$0.749_{[0.73,0.77]}$	$0.765_{[0.74,0.79]}$	$0.707_{[0.69,0.73]}$
	Concis.	$0.541_{[0.48,0.61]}$	$0.549_{[0.44,0.65]}$	$0.443_{[0.36,0.52]}$	$0.427_{[0.39,0.46]}$	$0.615_{[0.59,0.64]}$	$0.515_{[0.48,0.54]}$	$0.573_{[0.55,0.59]}$
balanced 32M	Comple.	$0.496_{[0.42,0.57]}$	$0.209_{[0.15,0.28]}$	$0.201_{[0.15,0.25]}$	$0.411_{[0.38,0.44]}$	$0.550_{[0.53,0.57]}$	$0.373_{[0.35,0.40]}$	$0.507_{[0.49,0.53]}$
	Faithf.	$0.892_{[0.86,0.92]}$	$0.916_{[0.87,0.96]}$	$0.791_{[0.74,0.83]}$	$0.418_{[0.38,0.45]}$	$0.691_{[0.67,0.71]}$	$0.742_{[0.73,0.76]}$	$0.642_{[0.62,0.66]}$
	Concis.	$0.481_{[0.42,0.54]}$	$0.457_{[0.39,0.51]}$	$0.313_{[0.26,0.36]}$	$0.469_{[0.44,0.50]}$	$0.639_{[0.62,0.66]}$	$0.472_{[0.45,0.49]}$	$0.589_{[0.57,0.61]}$
natural 2M	Comple.	$0.308_{[0.26,0.36]}$	$0.107_{[0.05,0.17]}$	$0.077_{[0.04,0.11]}$	$0.298_{[0.27,0.33]}$	$0.505_{[0.48,0.53]}$	$0.259_{[0.24,0.28]}$	$0.443_{[0.42,0.46]}$
	Faithf.	$0.886_{[0.84,0.93]}$	$0.854_{[0.69,0.99]}$	$0.774_{[0.67,0.86]}$	$0.438_{[0.39,0.49]}$	$0.744_{[0.72,0.76]}$	$0.739_{[0.70,0.78]}$	$0.685_{[0.67,0.70]}$
	Concis.	$0.498_{[0.43,0.57]}$	$0.361_{[0.22,0.53]}$	$0.478_{[0.38,0.58]}$	$0.351_{[0.31,0.39]}$	$0.645_{[0.62,0.67]}$	$0.467_{[0.43,0.51]}$	$0.575_{[0.55,0.60]}$
natural 32M	Comple.	$0.428_{[0.34,0.52]}$	$0.136_{[0.07,0.22]}$	$0.125_{[0.09,0.17]}$	$0.392_{[0.36,0.42]}$	$0.580_{[0.56,0.60]}$	$0.332_{[0.31,0.36]}$	$0.522_{[0.50,0.54]}$
	Faithf.	$0.875_{[0.85,0.91]}$	$0.916_{[0.82,0.99]}$	$0.786_{[0.70,0.86]}$	$0.359_{[0.32,0.39]}$	$0.701_{[0.68,0.72]}$	$0.728_{[0.70,0.75]}$	$0.636_{[0.62,0.66]}$
	Concis.	$0.559_{[0.50,0.62]}$	$0.499_{[0.26,0.72]}$	$0.347_{[0.28,0.42]}$	$0.434_{[0.40,0.47]}$	$0.694_{[0.67,0.72]}$	$0.506_{[0.46,0.56]}$	$0.625_{[0.61,0.64]}$

Table 25: Per-domain fact-level LLM-judge scores (Qwen2.5-72B-Instruct-AWQ) for the balanced and natural schemes at the 2M and 32M budgets, with bootstrap 95% CIs in subscripts. Comple. (completeness) is the fraction of reference-minute facts recovered by the generated minute; Faithf. (faithfulness) is the fraction of generated-minute facts supported by the transcript; Concis. (conciseness) is the fraction of generated-minute facts that are salient with respect to the reference. Macro weights the five domains equally, micro weights meetings. This is the per-domain breakdown of the aggregate scores in Table 3; the RQ1 crossover on completeness (balanced leads the minority domains, natural the majority) is visible per domain, while faithfulness falls from 2M to 32M under both schemes, most sharply on EPM.

		Mistral-7B			Llama-3.2-3B		
Dataset	**Budget**	Bal.	Nat.	Δ	Bal.	Nat.	Δ
ROUGE-Lsum							
AMI	2M	0.376	0.331	+0.045	0.098	0.152	−0.054
	8M	0.449	0.230	+0.219	0.312	0.113	+0.199
	32M	0.495	0.393	+0.102	0.476	0.156	+0.320
ICSI	2M	0.258	0.129	+0.129	0.080	0.029	+0.051
	8M	0.419	0.107	+0.312	0.174	0.079	+0.095
	32M	0.445	0.281	+0.164	0.402	0.124	+0.278
ELITR	2M	0.152	0.184	−0.032	0.091	0.040	+0.051
	8M	0.244	0.107	+0.137	0.122	0.082	+0.040
	32M	0.367	0.204	+0.163	0.229	0.105	+0.124
EPM	2M	0.222	0.274	−0.052	0.213	0.263	−0.050
	8M	0.438	0.516	−0.078	0.252	0.426	−0.174
	32M	0.547	0.534	+0.013	0.220	0.531	−0.311
MB	2M	0.562	0.586	−0.024	0.502	0.539	−0.037
	8M	0.598	0.646	−0.048	0.560	0.615	−0.055
	32M	0.648	0.694	−0.046	0.587	0.679	−0.092
BERTScore-F1							
AMI	2M	0.861	0.851	+0.010	0.821	0.800	+0.021
	8M	0.870	0.835	+0.035	0.861	0.821	+0.040
	32M	0.875	0.865	+0.010	0.872	0.831	+0.041
ICSI	2M	0.827	0.814	+0.013	0.788	0.750	+0.038
	8M	0.851	0.801	+0.050	0.823	0.784	+0.039
	32M	0.852	0.833	+0.019	0.844	0.800	+0.044
ELITR	2M	0.806	0.806	+0.000	0.786	0.759	+0.027
	8M	0.835	0.815	+0.020	0.817	0.779	+0.038
	32M	0.846	0.837	+0.009	0.836	0.805	+0.031
EPM	2M	0.832	0.831	+0.001	0.813	0.834	−0.021
	8M	0.865	0.882	−0.017	0.838	0.867	−0.029
	32M	0.887	0.889	−0.002	0.842	0.885	−0.043
MB	2M	0.910	0.915	−0.005	0.897	0.906	−0.009
	8M	0.918	0.925	−0.007	0.910	0.922	−0.012
	32M	0.928	0.938	−0.010	0.915	0.935	−0.020

Table 26: RQ5 full numeric detail backing Table 5. For each dataset, budget, and model, Bal. and Nat. are the raw balanced and natural scores and Δ = balanced − natural (from the displayed rounded values); all runs are pruned, single seed 42. The sign of Δ agrees across the two models for every dataset on both metrics—positive on AMI/ICSI/ELITR, negative on EPM/MB—the only per-cell exceptions being at 2M (AMI and ELITR on ROUGE-Lsum, EPM on BERTScore-F1).

Metric	Scheme	Budget	AMI *20*	ICSI *6*	ELITR *38*	EPM *242*	MB *862*	Macro	Micro
ROUGE-1	Balanced	2M	$0.380_{\pm 0.057}$	$0.271_{\pm 0.022}$	$0.161_{\pm 0.005}$	$0.253_{\pm 0.048}$	$0.576_{\pm 0.029}$	$0.328_{\pm 0.003}$	$0.491_{\pm 0.028}$
	Balanced	8M	$0.455_{\pm 0.022}$	$0.391_{\pm 0.046}$	$0.235_{\pm 0.024}$	$0.365_{\pm 0.087}$	$0.630_{\pm 0.005}$	$0.415_{\pm 0.031}$	$0.558_{\pm 0.016}$
	Balanced	32M	$0.527_{\pm 0.011}$	$0.452_{\pm 0.016}$	$0.370_{\pm 0.008}$	$0.524_{\pm 0.027}$	$0.675_{\pm 0.007}$	$0.510_{\pm 0.006}$	$0.630_{\pm 0.006}$
	Natural	2M	$0.316_{\pm 0.031}$	$0.176_{\pm 0.079}$	$0.176_{\pm 0.012}$	$0.355_{\pm 0.065}$	$0.611_{\pm 0.017}$	$0.327_{\pm 0.026}$	$0.536_{\pm 0.018}$
	Natural	8M	$0.241_{\pm 0.004}$	$0.131_{\pm 0.042}$	$0.105_{\pm 0.010}$	$0.481_{\pm 0.046}$	$0.669_{\pm 0.007}$	$0.325_{\pm 0.007}$	$0.602_{\pm 0.013}$
	Natural	32M	$0.351_{\pm 0.061}$	$0.217_{\pm 0.072}$	$0.178_{\pm 0.030}$	$0.548_{\pm 0.025}$	$0.715_{\pm 0.002}$	$0.402_{\pm 0.028}$	$0.654_{\pm 0.005}$
ROUGE-2	Balanced	2M	$0.155_{\pm 0.026}$	$0.059_{\pm 0.007}$	$0.037_{\pm 0.003}$	$0.163_{\pm 0.035}$	$0.473_{\pm 0.033}$	$0.177_{\pm 0.006}$	$0.387_{\pm 0.030}$
	Balanced	8M	$0.179_{\pm 0.003}$	$0.098_{\pm 0.020}$	$0.061_{\pm 0.006}$	$0.251_{\pm 0.070}$	$0.522_{\pm 0.005}$	$0.222_{\pm 0.018}$	$0.443_{\pm 0.011}$
	Balanced	32M	$0.201_{\pm 0.007}$	$0.110_{\pm 0.005}$	$0.092_{\pm 0.005}$	$0.386_{\pm 0.020}$	$0.577_{\pm 0.007}$	$0.273_{\pm 0.004}$	$0.513_{\pm 0.006}$
	Natural	2M	$0.108_{\pm 0.007}$	$0.034_{\pm 0.015}$	$0.037_{\pm 0.003}$	$0.234_{\pm 0.048}$	$0.499_{\pm 0.018}$	$0.183_{\pm 0.013}$	$0.420_{\pm 0.015}$
	Natural	8M	$0.087_{\pm 0.006}$	$0.020_{\pm 0.005}$	$0.021_{\pm 0.002}$	$0.348_{\pm 0.032}$	$0.569_{\pm 0.004}$	$0.209_{\pm 0.007}$	$0.495_{\pm 0.009}$
	Natural	32M	$0.139_{\pm 0.023}$	$0.044_{\pm 0.018}$	$0.043_{\pm 0.013}$	$0.416_{\pm 0.021}$	$0.624_{\pm 0.003}$	$0.253_{\pm 0.007}$	$0.551_{\pm 0.005}$
ROUGE-L	Balanced	2M	$0.233_{\pm 0.036}$	$0.158_{\pm 0.018}$	$0.103_{\pm 0.005}$	$0.214_{\pm 0.042}$	$0.547_{\pm 0.032}$	$0.251_{\pm 0.003}$	$0.456_{\pm 0.030}$
	Balanced	8M	$0.260_{\pm 0.007}$	$0.189_{\pm 0.008}$	$0.132_{\pm 0.014}$	$0.318_{\pm 0.084}$	$0.596_{\pm 0.005}$	$0.299_{\pm 0.020}$	$0.515_{\pm 0.014}$
	Balanced	32M	$0.284_{\pm 0.009}$	$0.191_{\pm 0.006}$	$0.174_{\pm 0.010}$	$0.463_{\pm 0.022}$	$0.647_{\pm 0.007}$	$0.352_{\pm 0.003}$	$0.585_{\pm 0.006}$
	Natural	2M	$0.190_{\pm 0.016}$	$0.112_{\pm 0.025}$	$0.108_{\pm 0.007}$	$0.297_{\pm 0.056}$	$0.573_{\pm 0.018}$	$0.256_{\pm 0.016}$	$0.492_{\pm 0.017}$
	Natural	8M	$0.156_{\pm 0.006}$	$0.091_{\pm 0.021}$	$0.072_{\pm 0.003}$	$0.426_{\pm 0.037}$	$0.639_{\pm 0.006}$	$0.277_{\pm 0.006}$	$0.566_{\pm 0.011}$
	Natural	32M	$0.214_{\pm 0.035}$	$0.125_{\pm 0.031}$	$0.114_{\pm 0.018}$	$0.495_{\pm 0.021}$	$0.690_{\pm 0.002}$	$0.328_{\pm 0.013}$	$0.620_{\pm 0.005}$
ROUGE-Lsum	Balanced	2M	$0.368_{\pm 0.054}$	$0.256_{\pm 0.020}$	$0.154_{\pm 0.006}$	$0.248_{\pm 0.048}$	$0.553_{\pm 0.033}$	$0.316_{\pm 0.002}$	$0.472_{\pm 0.031}$
	Balanced	8M	$0.440_{\pm 0.022}$	$0.371_{\pm 0.042}$	$0.228_{\pm 0.024}$	$0.360_{\pm 0.087}$	$0.604_{\pm 0.005}$	$0.400_{\pm 0.030}$	$0.537_{\pm 0.016}$
	Balanced	32M	$0.506_{\pm 0.010}$	$0.430_{\pm 0.018}$	$0.359_{\pm 0.008}$	$0.518_{\pm 0.027}$	$0.654_{\pm 0.007}$	$0.493_{\pm 0.006}$	$0.612_{\pm 0.006}$
	Natural	2M	$0.302_{\pm 0.032}$	$0.165_{\pm 0.076}$	$0.169_{\pm 0.013}$	$0.349_{\pm 0.065}$	$0.582_{\pm 0.018}$	$0.314_{\pm 0.025}$	$0.513_{\pm 0.018}$
	Natural	8M	$0.230_{\pm 0.005}$	$0.120_{\pm 0.037}$	$0.097_{\pm 0.008}$	$0.474_{\pm 0.045}$	$0.647_{\pm 0.006}$	$0.314_{\pm 0.006}$	$0.583_{\pm 0.012}$
	Natural	32M	$0.338_{\pm 0.058}$	$0.203_{\pm 0.069}$	$0.170_{\pm 0.030}$	$0.542_{\pm 0.025}$	$0.696_{\pm 0.002}$	$0.390_{\pm 0.027}$	$0.638_{\pm 0.005}$
BERTScore-F1	Balanced	2M	$0.862_{\pm 0.005}$	$0.825_{\pm 0.003}$	$0.801_{\pm 0.007}$	$0.831_{\pm 0.009}$	$0.909_{\pm 0.007}$	$0.846_{\pm 0.002}$	$0.888_{\pm 0.007}$
	Balanced	8M	$0.870_{\pm 0.002}$	$0.847_{\pm 0.004}$	$0.826_{\pm 0.012}$	$0.853_{\pm 0.014}$	$0.919_{\pm 0.001}$	$0.863_{\pm 0.006}$	$0.901_{\pm 0.003}$
	Balanced	32M	$0.877_{\pm 0.002}$	$0.850_{\pm 0.002}$	$0.842_{\pm 0.003}$	$0.882_{\pm 0.005}$	$0.930_{\pm 0.002}$	$0.876_{\pm 0.002}$	$0.916_{\pm 0.001}$
	Natural	2M	$0.849_{\pm 0.006}$	$0.815_{\pm 0.007}$	$0.803_{\pm 0.007}$	$0.847_{\pm 0.014}$	$0.914_{\pm 0.004}$	$0.846_{\pm 0.003}$	$0.895_{\pm 0.004}$
	Natural	8M	$0.838_{\pm 0.006}$	$0.799_{\pm 0.003}$	$0.810_{\pm 0.008}$	$0.874_{\pm 0.008}$	$0.926_{\pm 0.002}$	$0.850_{\pm 0.003}$	$0.909_{\pm 0.002}$
	Natural	32M	$0.854_{\pm 0.009}$	$0.813_{\pm 0.017}$	$0.821_{\pm 0.014}$	$0.892_{\pm 0.005}$	$0.938_{\pm 0.001}$	$0.864_{\pm 0.008}$	$0.922_{\pm 0.001}$

Table 27: Multi-seed robustness (RQ1): per-domain scores as mean $\pm$ std over seeds $\{2, 15, 42\}$ for the balanced and natural schemes, pruned condition, at the 2M, 8M, and 32M budgets. Column italics are test-set meeting counts. The per-domain crossover of Section 5.1 reproduces across seeds: at 8M and 32M, balanced leads all three minority domains (AMI, ICSI, ELITR) and natural leads the majority domains (EPM, MB) on every metric, so macro favours balanced while micro favours natural. At 2M the minority advantage is smaller and ELITR falls within seed variation.

O Prompts

O.1 Pruning Prompt

The verbatim prompt supplied to the pruning model is given below, comprising a system prompt and a user-prompt template. It is applied identically to all five corpora. Each transcript is processed in chunks; `transcript_chunk` in the user template is replaced with the numbered lines of the current chunk, and the global line numbering is preserved across chunks.

Pruning Prompt

[System]

```
You clean meeting transcripts by removing ONLY conversational noise. This is NOT summarization.
You KEEP the large majority of lines. You are removing filler, not compressing the meeting.

Each line starts with a number in square brackets: "[n] SPEAKER: text".

DELETE a line ONLY if it is pure conversational noise with NO standalone content:
- bare backchannels / fillers: "okay", "right", "yeah", "mm-hmm", "uh", "great", "cool"
- greetings, goodbyes, thanks, social pleasantries
- connection/logistics chatter: "can you hear me?", "you're muted"
- pure acknowledgements that do not answer a question or state anything
- meaningless fragments / false starts that carry no information ("Thin", "other")

KEEP a line if it contains ANY of the following, even briefly:
- any topic, fact, opinion, reason, problem, suggestion, or description
- decisions, agreements, action items, deadlines, numbers, prices, names, dates
- questions, and answers to questions (even short ones like "Friday", "Yes", "25 Euro")
- introductions, role assignments, anything about what the meeting is about

Rules:
- KEEP is the default. Only delete a line you are confident is pure noise.
- Most transcripts have only a MINORITY of lines deleted (roughly 15-35%), never the majority.
- If unsure, KEEP it (leave it out of the delete list).
- Do NOT rewrite, merge, renumber, or alter any line. Only decide deletions.
- Use the line number exactly as it appears in the brackets.

Output ONLY a JSON object with one key "delete": a list of the line numbers (integers) to
delete. Nothing else. Example: {"delete": [1, 2, 4, 7]}
```

[User]

```
Below is a meeting transcript with numbered lines.

Remove ONLY non-essential conversational noise (bare "okay"/"yeah"/"right" acknowledgements,
greetings, goodbyes, thanks, "can you hear me?", meaningless fragments). KEEP everything that
carries any content: decisions, numbers, questions, answers, opinions, reasons, descriptions.

This is NOT summarization. Keep the majority of lines. When unsure, keep the line.

Transcript:
<<<
{transcript_chunk}
>>>

Return ONLY the JSON object with the "delete" list.
```

O.2 Summarisation Prompt

The instruction used during zero-shot settings and fine-tuning. It is applied identically to all candidate models; `{transcript}` is replaced with the raw (unpruned) test transcript.

Summarisation Prompt

[System]

```
You are a meeting summarizer. Given a meeting transcript, write the meeting minutes
that faithfully capture the substantive content of the meeting. Base the minutes only
on what is stated in the transcript; do not introduce information that is not present.
```

[User]

```
Here is the meeting transcript. Write the meeting minutes.

{transcript}
```

O.3 LLM-as-a-Judge Evaluation Prompts

The prompts below are used in our fact-level evaluation (Section 3.6), applied under greedy decoding with a judge model isolated from the fine-tuned systems. Each enforces a strict JSON schema and a fixed set of adjudication rules applied identically across all mixtures, budgets, and domains; the same rules serve as annotation guidelines for the human-validation slice. Atomic facts are produced by the decomposition prompt, applied once to the reference minute and once to the generated minute.

Atomic Fact Decomposition

[System]

```
You are an expert at decomposing meeting minutes into atomic facts. This is NOT
summarization: you re-express existing content as a complete list of atomic facts,
adding and omitting nothing.

Follow four criteria for every atomic fact:
- Losslessness: together the facts capture ALL information in the summary.
- Atomicity: each fact is the smallest indivisible unit -- exactly one claim.
  Split every compound statement into separate facts.
- Independence: each fact is self-contained and understandable in isolation.
  Resolve pronouns and references (write "Mark", not "he").
- Declarativeness: each fact is one concise, objective declarative sentence.

Decompose ONLY what the summary states; do not infer beyond it. Do not treat
section headers or formatting as facts unless they carry standalone content.

Output ONLY a JSON object with one key "atomic_facts": a list of strings.
Example: {"atomic_facts": ["fact_1", "fact_2", "...", "fact_n"]}
```

[User]

```
Decompose the following meeting minutes into atomic facts.

[Summary Start]
{summary}
[Summary End]

Return ONLY the JSON object.
```

Completeness: Key-Fact Matching

[System]

```
You are an expert evaluator. Given KEY FACTS from a reference set of meeting minutes
and a candidate SUMMARY, decide for each key fact whether the candidate captures it.

Labels:
- captured: the candidate explicitly states the key fact, or states something from
  which it follows directly and unambiguously. Paraphrase and different wording
  count, provided meaning is preserved.
- not_captured: the key fact is absent, contradicted, or only vaguely alluded to
  without its substantive content.

Adjudication rules (apply strictly and identically):
- Meaning over wording: a faithful paraphrase captures the fact.
- Specificity, downward: if the candidate is LESS specific than the key fact, it is
  captured ONLY IF it still conveys the fact's essential content. Omitting the core
  value (number, name, decision) is not_captured, even if the topic is mentioned.
- Specificity, upward: extra detail in the candidate never harms capture.
- Numbers/dates/names: captured only if the specific value appears or is
  unambiguously entailed. A different value is not_captured.
- Base your judgment SOLELY on the candidate summary. Do NOT use the transcript or
  outside knowledge to fill gaps the candidate leaves.

For a captured fact, give the candidate sentence index that best evidences it;
for not_captured, set it to null.

Output ONLY a JSON list of objects with keys: fact_idx (int),
conclusion ("captured" or "not_captured"), evidence_sentence_idx (int or null),
reason (one sentence).
```

[User]

```
Summary:
{indexed_generated_summary_sentences}

Key Facts:
{indexed_reference_atomic_facts}

Return ONLY the JSON list.
```

Faithfulness: Fact Verification

[System]

```
You are an expert faithfulness checker. Given a meeting TRANSCRIPT and atomic FACTS
from a candidate summary of it, decide for each fact whether the transcript
supports it.

Labels:
- supported: the fact is explicitly stated in the transcript, or follows directly
  and unambiguously from what is stated.
- not_supported: the fact is missing, contradicted, or requires an inference the
  transcript does not license.

Adjudication rules (apply strictly and identically):
- Base your judgment SOLELY on the transcript. Do NOT use outside knowledge, and do
  NOT assume a claim is true merely because it is plausible.
- Meaning over wording: a faithful paraphrase of transcript content is supported.
- Specificity, upward: if the fact is MORE specific than the transcript supports, it
  is not_supported. (Transcript "around 25 euros" / fact "exactly 25 euros" ->
  not_supported; transcript "25 euros" / fact "around 25 euros" -> supported.)
- Approximation: rounding or hedging is acceptable ONLY when the transcript itself
  signals approximation.
- Numbers/dates/names: supported only if the specific value appears in or is
  unambiguously entailed by the transcript. Any mismatch is not_supported.
- Attribution: if the fact attributes a statement or position to a speaker, the
  transcript must support that attribution, not just the content.
- Modality: distinguish decided vs. proposed / considered / rejected. A fact
  asserting a decision the transcript only floated is not_supported.

Output ONLY a JSON list of objects with keys: fact_idx (int),
conclusion ("supported" or "not_supported"), reason (one sentence).
```

[User]

```
[Transcript Start]
{transcript}
[Transcript End]

Facts:
{indexed_generated_summary_atomic_facts}

Return ONLY the JSON list.
```

Conciseness: Salience Check

[System]

```
You are an expert evaluator. Given a candidate SUMMARY's atomic FACTS and a
reference set of meeting minutes, decide for each candidate fact whether it
corresponds to salient content present in the reference (i.e. whether it is salient
rather than padding).

Labels:
- salient: the reference states this fact, or states something from which it follows
  directly. Paraphrase and different wording count.
- non_salient: the fact is absent from the reference -- extra content the reference
  did not deem minute-worthy, or a redundant restatement.

Adjudication rules (apply strictly and identically):
- Meaning over wording: a faithful paraphrase of reference content is salient.
- Judge salience only against the reference. Do NOT use the transcript or outside
  knowledge; a fact may be true of the meeting yet non_salient if the reference did
  not record it.
- Specificity: if the candidate fact is more specific than the reference on the same
  point, still mark salient. If it introduces a value or claim the reference does
  not contain at all, mark non_salient.
- Redundancy: if a candidate fact merely restates another candidate fact already
  matched to the same reference content, mark it non_salient.

Output ONLY a JSON list of objects with keys: fact_idx (int),
conclusion ("salient" or "non_salient"), evidence_sentence_idx (int or null),
reason (one sentence).
```

[User]

```
Reference:
{indexed_reference_sentences}

Facts:
{indexed_generated_summary_atomic_facts}

Return ONLY the JSON list.
```