Paired by the Teacher: Turning Unpaired Data into High-Fidelity Pairs for Low-Resource Text Generation

Yen-Ju Lu, Thomas Thebaud, Laureano Moro-Velazquez,
Najim Dehak, Jesus Villalba
Center for Language and Speech Processing, Johns Hopkins University
`{ylu125, tthebau1, laureano, ndehak3, jvillal7}@jhu.edu`

Abstract

We present *Paired by the Teacher* (PbT), a two-stage teacher–student pipeline that synthesizes accurate input–output pairs without human labels or parallel data. In many low-resource natural language generation (NLG) scenarios, practitioners may have only raw outputs, like highlights, recaps, or questions, or only raw inputs, such as articles, dialogues, or paragraphs, but seldom both. This mismatch forces small models to learn from very few examples or rely on costly, broad-scope synthetic examples produced by large LLMs. PbT addresses this by asking a teacher LLM to compress each unpaired example into a concise intermediate representation (IR), and training a student to reconstruct inputs from IRs. This enables outputs to be paired with student-generated inputs, yielding high-quality synthetic data. We evaluate PbT on five benchmarks—document summarization (XSum, CNNDM), dialogue summarization (SAMSum, DialogSum), and question generation (SQuAD)—as well as an unpaired setting on SwitchBoard (paired with DialogSum summaries). An 8B student trained only on PbT data outperforms models trained on 70 B teacher-generated corpora and other unsupervised baselines, coming within 1.2 ROUGE-L of human-annotated pairs and closing 82% of the oracle gap at one-third the annotation cost of direct synthesis. Human evaluation on SwitchBoard further confirms that only PbT produces concise, faithful summaries aligned with the target style, highlighting its advantage of generating in-domain sources that avoid the mismatch, limiting direct synthesis.

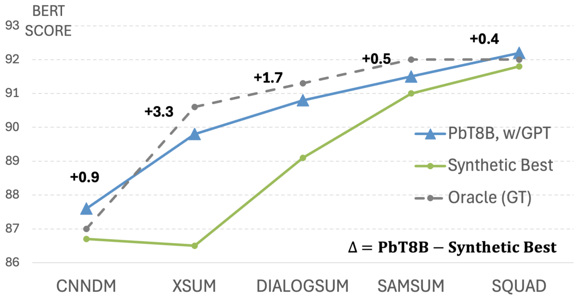


Figure 1: PbT consistently outperforms the strongest synthetic baselines from GPT and LLAMA70B (by up to +3.3 BERTScore points) and closes the gap to oracle (real) data across five summarization and QG tasks.

1 Introduction

Many NLG tasks, such as summarization and question generation, lack readily available input–output pairs, often referred to as source–target pairs. Creating these pairs requires experts to read each document or dialogue and write its summary or questions. This process is slow, costly, and must be repeated for every new domain. It also raises privacy concerns when conversations contain sensitive information. To solve this problem, we introduce *Paired by the Teacher* (PbT). Given only unpaired in-domain texts and a small set of example summaries or questions, PbT generates high-quality synthetic source–target pairs without any human pairing. This pipeline turns raw data into ready training material that matches the target style and speeds up adaptation across domains.

Existing work shows that synthesis models can be trained using only unpaired data, such as back-translation (Lample et al., 2018), GAN-based, or VAE-based summarization (Wang and LEE, 2020; Bražinskas et al., 2020). More recent techniques leverage instruction-tuned teacher LLMs to generate training pairs, such as DiaSynth with GPT-4o (Suresh et al., 2025), MixSumm with LLaMA-70B (Sahu and Laradji, 2024), and ChatGPT self-chat corpora (Xu et al.). In few-shot scenarios, smaller teachers like GPT-3.5 or in-model synthesis adapters can also guide the pairing process (Mishra et al., 2023; Lu et al., 2025). However, these methods rely heavily on the teacher's generation, refinement, and filtering abilities. This drives up inference costs and makes deployment in resource-constrained settings challenging, or requiring some

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20401–20423
November 4-9, 2025 ©2025 Association for Computational Linguistics

ground-truth labeled data to ensure quality.

Paired by the Teacher (PbT) is a two-phase teacher–student pipeline that transforms unpaired inputs and outputs into high-quality source–target pairs without any manual annotation. In the first phase, a teacher model extracts a concise IR from each raw source example, and a smaller student model is fine-tuned via parameter-efficient methods to invert this mapping and reconstruct the source. In the second stage, the teacher model annotates IR for each unpaired target with the source-side IR example as in-context demonstrations. Each generated IR is then passed through the trained student model to synthesize an aligned source example. Because PbT asks the teacher to output only short IRs, the heavy text generation is offloaded to the student. The teacher's knowledge still guides the process, but its compute cost is a fraction of end-to-end generation. At the same time, the student has been trained on real in-domain documents and dialogues. As a result, the synthetic sources it produces remain in-domain, unlike direct synthesis where mismatched sources hurt downstream performance. This in-domain alignment is the key reason PbT generalizes better at test time.

We evaluate PbT on five benchmarks—document summarization (XSum, CNNDM), dialogue summarization (SAMSum, DialogSum), and question generation (SQuAD). An 8B student trained only on PbT data outperforms models trained on 70B teacher-generated or self-generated data, and comes on average within 1.2 ROUGE-L points of human-annotated pairs, recovering 5.6 of the 6.8 ROUGE-L gap (82% closure) compared with the best synthetic-only baseline. We further build an unpaired setting on SwitchBoard dialogue-only dataset, paired with summaries from DialogSum and train the summarization model on purely cross-domain data. Human evaluation on SwitchBoard shows that only PbT adheres to the target summary lengths with concise, fluent, and faithful outputs. In contrast, all baselines produce overly long summaries due to a domain mismatch between the synthetic training data and the real test inputs. This highlights PbT's core advantage: by generating sources through a student trained on real inputs, PbT keeps training data aligned with test-time conditions.

Major contributions:

- **Teacher–Student Pairing Pipeline.** PbT transforms unpaired sources and targets into high-fidelity training pairs, enabling a student model to learn from realistic source examples without any human annotations. Unlike direct synthesis, PbT's student-generated sources remain in-domain, ensuring alignment between training and test distributions.

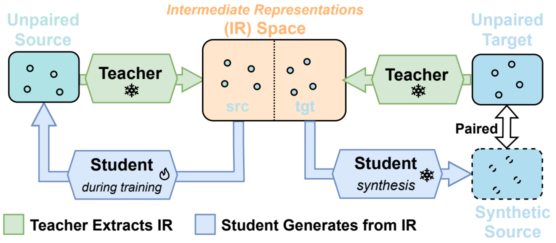


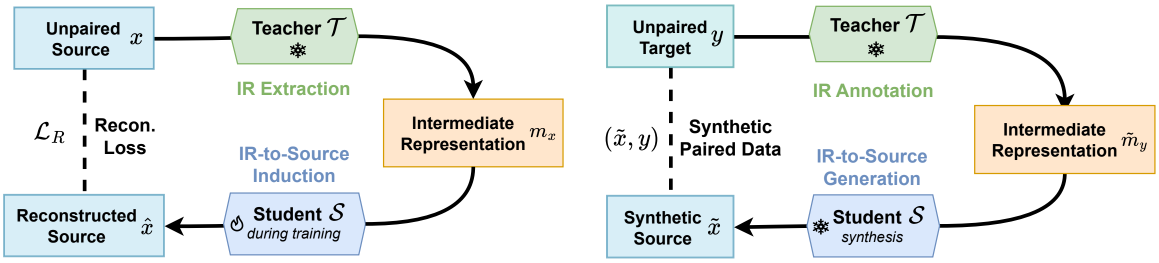
Figure 2: High-level overview of Paired-by-the-Teacher (PbT). The teacher LLM extracts intermediate representations (IR) from both unpaired sources (left) and unpaired targets (right). A student model is then trained to invert IR→source (blue loop) and to synthesize new sources from IR (blue arrow), producing synthetic pairs $(\tilde{x}, y)$ for downstream NLG fine-tuning.

- **Robust Intermediate Representation.** We conduct the experiments with various IR formats and annotated by larger, same-sized, or smaller teachers, reducing dependence on model scale while maintaining data quality. Also, the cost for IR annotation is 3× smaller than direct synthesis from the teacher model.

- **Intra-and Cross-Dataset Experiments:** PbT not only surpasses teacher/student-only baselines across all five benchmarks in automatic metrics and model-based evaluation, but also yields a functional summarizer on the label-free SwitchBoard corpus by borrowing DialogSum summaries—adhering to their style and outperforming baseline systems in both human and GPT evaluation.

2 Related Work

Unpaired and Low-Resource NLG with Model Selection Before the emergence of LLMs, unsupervised text generation used cycle objectives or auto-encoding on smaller models. Back-translation let dual translators learn without parallel data (Lample et al., 2018). Similar cycle or reconstruction ideas powered unsupervised summarization, sentence compression, and opinion summarization (Wang and LEE, 2020; Baziotis et al., 2019; Bražinskas et al., 2020), as well as large-scale pseudo-pre-



(a) **Source-side IR Learning (Student Reconstruction).** (b) **Target-side IR Annotation (Synthetic Pair Generation).**

Figure 3: Overview of the *Paired by the Teacher* (PbT) pipeline. (a) The teacher model $\mathcal{T}$ extracts an intermediate representation m_x from each source x, and the student $\mathcal{S}$ learns to reconstruct x from m_x. (b) The teacher annotates each target y with a plausible IR $\tilde{m}_y$, which the student expands into a synthetic source $\tilde{x}$, forming synthetic pairs $(\tilde{x}, y)$ for downstream training without human-written labels.

training from news leads (Yang et al., 2020). Cycle consistency was later applied to data-to-text generation (Guo et al., 2020). Because no gold validation exists, these studies chose checkpoints with round-trip BLEU, reconstruction loss, or discriminator scores, which correlate well with true quality. We follow the same principle, using IR-to-document reconstruction to guide model selection for LLM-scale training.

LLM-Driven Pseudo-Labeling and In-Context NLG With instruction-tuned LLMs, generating synthetic corpora for low-resource tasks has become widespread. MixSumm (Sahu and Laradji, 2024) and PPSL (Sahu et al., 2025) prompt LLaMA-3-70B to synthesize documents and pseudo-summaries, then fine-tune a 100–300 M summarizer to match fully-supervised ROUGE with only 5 % gold pairs (Sahu et al., 2025). Mishra et al. (2023) use GPT-3.5 to pseudo-label customer-service chats; with just 10 % human data, their BERT-based model recovers 95 % of full-data performance. MRDS (Lu et al., 2025) further intertwines dialogue synthesis and summarization LoRA adapters inside a single LLM with pseudo-pair generation. In parallel, self-chat methods like Baize (Xu et al.) collect $\sim$100 k ChatGPT-generated dialogues to LoRA-tune LLaMA, approaching proprietary chatbot quality, and task-oriented work (Surana et al., 2025) turns product reviews into synthetic recommendation dialogues for ReDial (Li et al., 2018) and INSPIRED (Hayati et al., 2020) benchmarks. Finally, zero-shot and in-context prompting with GPT-3 can produce news summaries preferred by humans over gold references (Goyal et al., 2022), though such methods lack a lightweight deployable model, motivating hybrid approaches that combine in-context learning with pseudo-fine-tuning.

3 Methodology

Paired by the Teacher (PbT) is a unified teacher–student pipeline that converts unpaired inputs and outputs into high-fidelity source–target pairs without any human labeling.

PbT Overview Figure 2 provides a high-level illustration of the PbT pipeline. The process unfolds in three phases. In the Source IR Induction phase, the teacher LLM compresses each unpaired source into a concise intermediate representation (IR), and the student model is trained to reconstruct the original input from this IR. In the Target IR Annotation phase, the teacher annotates each unpaired target with a plausible IR, which the trained student then expands into a synthetic source, forming paired examples $(\tilde{x}, y)$. Finally, in the *Downstream Fine-tuning* phase, a summarization or question-generation model is fine-tuned on these synthetic pairs. By shifting heavy generation to the student while keeping the teacher's guidance at the IR level, PbT achieves both efficiency and alignment (see Appendix A.3 for concrete IR examples).

3.1 Source-side IR Learning

Our pipeline begins by teaching the student model to reconstruct in-domain source data from the IR. We first use a large teacher model $\mathcal{T}$ to extract a structured IR from each raw source example, and then fine-tune a smaller student model $\mathcal{S}$ to invert that mapping.

IR Extraction (Teacher) As illustrated in Figure 3a, for every unpaired source instance x (dialogue, article, paragraph, ...), the teacher $\mathcal{T}$ produces a concise IR m_x (e.g., a topical outline, chain-of-thought, or related facts). This yields a semi-supervised dataset of (x_i, m_{x_i}) pairs:

$$\{(x_i, m_{x_i})\}_{i=1}^{N}. \tag{1}$$

The form of m_{x_i} is task-dependent. For dialogue and document summarization, the teacher is instructed to split the input into paragraph groups and write a topic sentence with key entities. For question generation, the teacher extracts the answer-containing sentence and writes supporting facts as bullets. These task-specific IRs provide the information needed to reconstruct the source x_i (see Table 12).

IR-to-Source Induction (Student) We then fine-tune the student model $\mathcal{S}$ with parameter-efficient methods (PEFT) on these source-side annotations to learn the inverse mapping

$$\mathcal{S}_{m \to x} : m \mapsto x. \tag{2}$$

By optimizing on (x_i, m_{x_i}), the student internalizes the source domain's structural patterns and can expand IRs into realistic source texts $\hat{x} = \mathcal{S}_\theta(m)$. The model is trained with the reconstruction loss

$$\theta = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}_R(\mathcal{S}_\theta(m_{x_i}), x_i), \tag{3}$$

where $\mathcal{L}_R$ is the cross-entropy loss. These learned synthesis skills later allow $\mathcal{S}$ to generate candidate sources for unpaired targets.

3.2 Target-side IR Annotation

IR Annotation (Teacher) As shown in Figure 3b, we prompt the teacher LLM $\mathcal{T}$ to annotate each unpaired target example y with a plausible IR $\tilde{m}_y$. To ensure consistent formatting, $\mathcal{T}$ is provided with a few in-context examples of (x, m_x) pairs from the source-side training phase. This yields a set of target annotations:

$$\{(y_j, \tilde{m}_{y_j})\}_{j=1}^{N_y}. \tag{1}$$

Because the teacher does not observe the corresponding source x_j, the generated $\tilde{m}_{y_j}$ are not aligned with the actual sources. Such explicit alignment is unnecessary, as the student generates synthetic sources directly from these IRs.

IR-to-Source Generation (Student) Once the student model $\mathcal{S}_{m \to x}$ has learned to expand IRs into full source texts, we apply it to each target-side representation $\tilde{m}_{y_j}$. The student then produces a synthetic source $\tilde{x}_j$ for each target y_j, resulting in a synthetic paired corpus

$$\{(\tilde{x}_j,\, y_j)\}_{j=1}^{N_y}, \tag{2}$$

where every target is matched with a candidate source generated solely from its corresponding IR.

3.3 Target Task Training

Finally, we fine-tune a downstream student model $\mathcal{S}_{x \to y}$ on the synthetic pairs $\{(\tilde{x}_j,\, y_j)\}_{j=1}^{N_y}$. The model learns to map synthetic sources to targets:

$$\mathcal{S}_{x \to y} \;:\; \tilde{x} \;\mapsto\; y. \tag{3}$$

The training objective is

$$\phi = \arg\min_{\phi} \sum_{j=1}^{N_y} \mathcal{L}_S\big(\mathcal{S}_\phi(\tilde{x}_j),\, y_j\big), \tag{4}$$

yielding predictions $\hat{y} = \mathcal{S}_\phi(\tilde{x})$.

PbT mimics the effect of having true training pairs for two reasons. First, it generates only synthetic source inputs $\tilde{x}$ and pairs them with authentic targets y, avoiding propagation of synthesis artifacts into outputs. Second, by training the student on real source data, PbT ensures that synthetic sources remain in-domain. As a result, at test time, PbT-generated pairs better match the distribution of real inputs, unlike direct synthesis, which suffers from mismatched sources.

4 Experiments

Intra-Datasets Experiment We evaluate PbT on three tasks with five benchmarks: dialogue-summarization (DialogSum (Chen et al., 2021) and SAMSum (Gliwa et al., 2019)), document summarization (XSum (Narayan et al., 2018) and CNNDM (Hermann et al., 2015)), and question generation (SQuAD (Rajpurkar et al., 2016)). For each dataset, we perform intra-dataset experiments by partitioning the data into "source-only" (dialogues, documents, and paragraphs) and "target-only" (summaries and questions) subsets, training the model without any paired examples.

Cross-Datasets Experiment Additionally, we assess the cross-domain transfer ability of PbT by training it on SwitchBoard (Godfrey et al., 1992), a conversation-style dialogue-only dataset, containing 1155 dialogues without any available summary labels. We train the PbT model with the dialogues

Table 1: **XSum (200-shot unpaired documents)**: Comparison of methods on small-source settings.

Synthesis model	Ref-based			G-Eval				
	R-2	R-L	BERTScore	Coh.	Con.	Rel.	Flu.	Ave.
Zero shot	5.3	12.5	86.2	N/A†				
ICL (summary-only)	7.9	18.2	87.6	N/A†				
Unpaired Sources — Direct synthesis (Summ$\rightarrow$(*Teacher/Student*)$\rightarrow$Document)								
LLAMA8B$_{\text{direct}}$ ($\mathcal{S}$)	6.5	18.4	87.0	2.29	3.83	2.09	2.95	2.79
LLAMA70B$_{\text{direct}}$ ($\mathcal{T}$)	5.0	16.5	86.5	2.33	4.03	2.09	2.94	2.84
GPT$_{\text{direct}}$ ($\mathcal{T}$)	5.0	16.4	86.2	2.22	3.78	2.09	2.91	2.75
GPT$_{\text{direct, BoN-5}}$ ($\mathcal{T}$)	5.4	16.9	86.5	2.38	3.95	2.25	2.92	2.88
Unpaired Sources — PbT synthesis (Summ$\rightarrow$*Teacher*$\rightarrow$IR$\rightarrow$*Student*$\rightarrow$Document)								
PbT$_{\text{8B, w/ llama70B}}$	16.0	29.3	89.7	2.95	4.14	2.79	**3.00**	3.22
PbT$_{\text{8B, w/ GPT}}$	**16.1**	**29.4**	**89.8**	2.99	4.08	2.85	2.98	3.22
PbT$_{\text{8B, w/ GPT, BoN-5}}$	15.4	28.6	89.6	**3.18**	**4.43**	**3.10**	**3.00**	**3.43**
Oracle Paired								
Real Data	18.7	32.6	90.6	2.74	3.71	2.59	**3.00**	3.01

† N/A indicates that G-Eval results are omitted for pretrained LLM outputs due to known evaluation bias.

from SwitchBoard and summaries from DialogSum, aiming to train a summarization model that can summarize the dialogues from SwitchBoard into DialogSum-style summaries.

4.1 Alternative Methods

We compare our PbT method with several unpaired summarization alternative approaches, divided into two categories: methods using the pre-trained Llama3 model without fine-tuning and methods fine-tuned on real or synthetic data.

Pre-trained Methods

- **Zero-shot:** Zero-shot summarization performance of LLAMA8B.
- **ICL:** Summarization performance of LLAMA8B using in-context learning with $k = 5$ summaries examples.

Fine-tuned Methods

- **Direct Synthesis**: Using a fixed teacher or student model to generate source from the summaries with in-context source examples. Then we fine-tune the student summarization model from those synthetic source-target pairs.
- **Best-of-5 Filtering (BoN-5)**: Applied to both baselines and PbT, this strategy generates five candidate pairs and filter as in Genie (Yehudai et al.). GPT-4o-mini is first used to filter inconsistent document–summary pairs, and a QRM-Gemma-2-27B ranker (Dorka, 2024) selects the top candidate.
- **Oracle Paired**: Fine-tuning the student summarization model with the oracle summaries for the source shots used in other experiments.
- **DiaSynth**: Fine-tuning the student summarization model on synthetic data from (Suresh et al., 2025), which leverages LLM CoT reasoning. The amount of synthetic training data is aligned with the other experiments.

4.2 Implementation Details

Teacher/Student Selection Our primary experiments use Llama3.1-8B-Instruct (Dubey et al., 2024) as the student model (denoted LLAMA8B). We train LoRA adapters of rank 16 and $\alpha = 32$ on top of the frozen base weights. To generate IRs, we rely on two large teachers: the Llama3.3-70B-Instruct model (LLAMA70B) and GPT-4o mini (GPT), and then fine-tune LLAMA8B on the synthetic data they produce.

Training Settings (Summarization, QG, and Source-synthesis). Both the summarization / question-generation models and the IR$\rightarrow$ source synthesizer are trained with the same optimization recipe from (Lu et al., 2025). During synthesis training, the student sees only the IR without summaries, forcing them to reconstruct the source from

Table 2: **CNNDM (200-shot unpaired Documents)**: Comparison of methods on small-source settings.

Synthesis model	Ref-based			G-Eval				
	R-2	R-L	BERTScore	Coh.	Con.	Rel.	Flu.	Ave.
Zero shot	13.1	21.0	86.8	N/A†				
ICL (summary-only)	15.2	24.6	87.5	N/A†				
Unpaired Sources								
GPT$_{\text{direct}}$ ($\mathcal{T}$)	16.0	22.7	86.4	4.51	4.92	4.47	2.94	4.21
LLAMA8B$_{\text{direct}}$ ($\mathcal{S}$)	16.6	23.9	86.7	4.47	**4.95**	4.29	2.94	4.16
PbT$_{\text{8B, w/ GPT}}$	**17.5**	**27.3**	**87.6**	**4.61**	4.91	**4.53**	**2.98**	**4.26**
Oracle Paired	16.8	25.7	87.0	4.41	4.90	4.21	2.93	4.11

† N/A indicates that G-Eval results are omitted for pretrained LLM outputs due to known evaluation bias.

the IR alone. For the proposed methods in the main experiments, we run the synthesis and training process three times and compute the average. The hyperparameters and prompts for IR extraction, source generation, and downstream training are listed in Appendix A.1 A.3.

5 Results

5.1 Unpaired Document Summarization

For document summarization, we conduct experiments at two different scales—200-shot and 1,000-shot source training—on the XSum and CNNDM datasets. For validation and testing, we randomly sample 500 and 1,000 examples, respectively, from each split. In addition to ROUGE and BERTScore, we also report G-Eval (Liu et al., 2023), a rigorously designed model-based evaluation framework that uses fixed prompts and the GPT-4 model, and has shown high correlation with human judgments.

XSum As shown in Table 1, XSum is challenging dataset, as it requires generating a one-line summary from a long news article. Both zero-shot and ICL experiments yield ROUGE-2 scores below 10%, and direct synthesis performs even worse than ICL (7.9 vs. 5.0 and 6.5 R-2), falling far short of the oracle paired-data baseline (18.7 R-2). In contrast, the proposed PbT approach achieves a ROUGE-2 score of 16.1, closing more than 80% of the gap to the oracle performance. Incorporating the Best-of-5 (BoN-5) filtering strategy does not further improve ROUGE-based scores, suggesting limited gains under reference-based evaluation. Results for the medium-scale XSum setting are provided in Appendix A.2.

XSum G-Eval Table 1 also presents G-Eval metrics on XSum, including Coherence (Coh.), Consistency (Con.), Relevance (Rel.), and Fluency (Flu.). We compute over the first 200 test examples across 20 runs. We find that the direct-synthesis baselines fall short in coherence and relevance. In contrast, PbT outperforms the real paired data on three metrics, achieving an overall average of 3.22 compared to 3.01. With BoN-5 filtering, although no further improvements are observed in Ref-based scores, PbT achieves a new best overall G-Eval score of 3.43, showing consistent gains in coherence, consistency, and relevance. This indicates that BoN-5 effectively refines the quality of synthetic data under unpaired scenarios when evaluated by human-aligned metrics.

CNNDM We also conduct experiments on CNNDM, a dataset with higher risk of pretraining contamination (Jiang et al., 2024). As a result, the gap between pretrained models and oracle paired data is relatively small—only 2–4% in ROUGE-L (21.0, 24.6 vs. 25.7)—compared to the much larger gap on XSum (12.5, 18.2 vs. 32.6). In this setting, direct synthesis offers only modest ROUGE-2 gains, whereas PbT outperforms both ICL and oracle paired data. Moreover, G-Eval shows that PbT achieves the best overall average of 4.26 (vs. 4.11 for oracle), with notable improvements in relevance (4.53 vs. 4.21) and coherence (4.61 vs. 4.51), highlighting that PbT yields more human-aligned quality gains even when reference-based improvements are limited.

5.2 Unpaired Dialogue Summarization

We evaluate PbT on DialogSum and SAMSum under two resource settings: small (1% of dialogues,

Table 3: **DialogSum (124/623-shot unpaired dialogues)**: Comparison of methods on small- and medium-source settings.

Synthesis model	Small (124-shot)			Medium (623-shot)		
	R-2	R-L	BERTScore	R-2	R-L	BERTScore
Zero shot	9.0	19.6	87.3	≡	≡	≡
ICL (summary-only)	13.8	28.7	89.9	≡	≡	≡
Unpaired Sources — Direct synthesis (Summ→(*Teacher/Student*)→Dialogue)						
DiaSynth	11.9	25.0	88.8	≡	≡	≡
GPT$_{\text{direct}}$ ($\mathcal{T}$)	14.7	27.7	89.1	15.4	29.2	89.5
LLAMA70B$_{\text{direct}}$ ($\mathcal{T}$)	15.2	28.7	89.1	15.1	28.9	89.3
LLAMA8B$_{\text{direct}}$ ($\mathcal{S}$)	14.4	29.2	89.4	13.7	27.3	89.0
Unpaired Sources — PbT synthesis (Summ→*Teacher*→IR→*Student*→Dialogue)						
PbT$_{\text{8B, w/ GPT}}$	**17.7**	**35.0**	**90.8**	**18.4**	**35.3**	**91.2**
PbT$_{\text{8B, w/ llama70B}}$	**17.8**	**34.1**	**90.4**	**18.7**	**35.1**	**90.9**
Oracle Paired						
Real Data	19.6	36.9	91.3	20.6	37.4	91.3

≡ indicates the result is the same as in the Small setting.

Table 4: **SAMSum (147-shot unpaired dilaogues)**: Comparison of methods on the small-size source data.

Synthesis model	R-2	R-L	B-S
Zero-shot	12.9	25.1	88.3
ICL (summary only)	18.5	33.8	90.4
Unpaired Sources			
DiaSynth	16.0	29.8	89.5
GPT$_{\text{direct}}$	23.6	37.7	90.9
LLAMA8B$_{\text{direct}}$	22.6	37.2	91.0
PbT$_{\text{8B, w/ GPT}}$	**24.6**	**40.3**	**91.5**
Oracle Paired	27.9	43.7	92.0

124/147-shot) and medium (5%, 623/736-shot) isolated source data. For fair comparison across datasets, we fix 1,000 unpaired targets in all experiments. The evaluation includes ROUGE-2 (R-2), ROUGE-L (R-L) (Lin, 2004), and BERTScore (Zhang et al., 2020).

DialogSum We observe in Table 3 that in-context learning (ICL) delivers a substantial boost over the zero-shot baseline, raising the R-L score from 19.6 to 28.7. Directly synthesizing source dialogues from target summaries with a large teacher model or a smaller student model yields performance slightly better or close to ICL. In addition, training on the DiaSynth (Suresh et al., 2025) shows performance that is worse than the ICL results. This suggests that while applying strong LLM summarizers, directly generating synthetic data does not offer additional gains beyond what can be achieved by simply prompting the model with example summaries. The experiments emphasize the difficulty of using LLMs to synthesize data without actual paired source-target examples. By contrast, our PbT methods deliver substantially greater gains than either ICL or direct synthesis. In the low-resource setting, they raise R-L from 29.2 to 35.3, achieving 1.6 points below the oracle paired-data upper bound of 36.9. Moreover, this improvement holds consistently for both teacher models (GPT and LLAMA70B) and across different data scales.

SAMSum We conduct the same set of experiments on SAMSum. Table 4 shows results for the small training setup (147-shot), and Appendix A.2 for the medium (736-shot). Overall, SAMSum yields higher ROUGE scores in the zero-shot and ICL settings compared to DialogSum. Direct synthesis already provides gains over the ICL baselines. However, the PbT approach further improves R-L scores by 3% absolute.

5.3 Question Generation

We evaluate PbT on the SQuAD dataset using the same training settings. Here, the source is a paragraph with its answer span, and the target is the corresponding question. The IR is defined as the

Table 5: **SQuAD (200-shot unpaired paragraphs)**: Comparison of question-generation pipelines. Metrics on the left (ROUGE-2, ROUGE-L, BERTScore, BLEU, METEOR) measure surface similarity to the reference question; columns on the right evaluate the usefulness of each generated question by running a QA model on the original passage (Exact-Match / F1).

Synthesis model	R-2	R-L	BERTScore	BLEU	METEOR	QA EM.	QA F1.
Zero shot	25.1	41.6	91.2	12.8	41.6	**53.0**	**72.2**
ICL (summary-only)	24.1	40.8	91.0	13.0	39.4	47.3	65.3
Unpaired Sources							
GPT$_{\text{direct}}$ ($\mathcal{T}$)	27.8	45.5	91.8	18.8	39.3	47.7	66.8
LLAMA8B$_{\text{direct}}$ ($\mathcal{S}$)	27.0	44.5	91.6	17.7	37.1	46.7	64.2
PbT$_{\text{8B, w/ GPT}}$	**30.3**	**48.0**	**92.2**	**20.3**	**41.1**	**52.3**	**71.1**
Oracle Paired	29.5	47.2	92.0	19.4	39.7	51.4	70.8

Table 6: **SwitchBoard → DialogSum (1155 unpaired dialogues)**: Cross-dataset transfer results for dialogue summarization.

	Fai.	**Flu.**	**Inf.**	**Red.**	**Ave.**
Zero	1.53	1.83	**1.94**	1.44	1.69
GPT$_{\text{dir}}$	**1.90**	1.44	1.28	1.47	1.52
PbT	1.87	**1.94**	1.14	**2.00**	**1.74**
Real$_{\text{DSum}}$	1.87	1.47	1.47	1.47	1.57

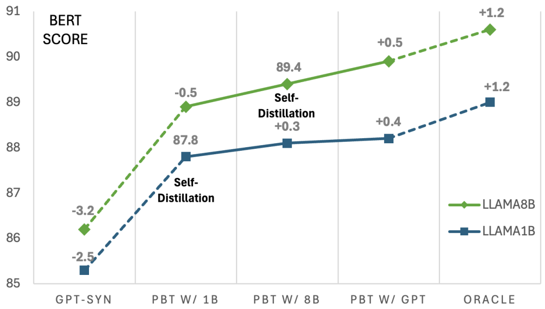


Figure 4: Direct synthesis from GPT, PbT w/ various teacher models, and real data comparison on XSum with LLAMA8B and LLAMA1B student models.

sentence containing the answer, together with key facts and the answer span. Examples and prompts are provided in Appendix A.3.

PbT outperforms both ICL and direct synthesis across all metrics, even surpassing oracle-paired training, mirroring the trend on CNNDM. Although zero-shot LLAMA8B achieves the best downstream QA scores, it lags far behind in BLEU/ROUGE and generates much longer questions than ground-truth. Direct synthesis recovers some lexical overlap (+2 ROUGE over the zero-shot baseline) but still trails IR-based PbT by 4 ROUGE-L and 2 BLEU, confirming the value of explicit IRs in question generation.

5.4 Cross Dataset Experiments

We performed unpaired training combining SwitchBoard dialogues with DialogSum summaries, since SwitchBoard contains only raw conversations without summary annotations. We compare PbT against three baselines: zero-shot pretrained model, GPT-based direct synthesis, and a model trained on true DialogSum pairs and applied directly to SwitchBoard. Six human evaluators assessed the summaries in four dimensions–informativeness, faithfulness, fluency, and redundancy– using a 0-2 scale, following the evaluation protocol from (Lu et al., 2025; Xie et al., 2024).

Table 6 presents human evaluation on SwitchBoard datset. In faithfulness, all fine-tuned methods score similarly (1.87–1.90), outperforming the zero-shot model (1.53). PbT achieves the highest fluency (1.94), likely due to its training data being more closely aligned with SwitchBoard than the GPT$_{\text{dir}}$ outputs or the Real$_{\text{DSum}}$ pairs. For informativeness and redundancy, we measure each model's adherence to DialogSum's concise two-to-three-sentence style. Summarizing long SwitchBoard dialogs naturally lowers informativeness (1.14) and reduces redundancy (2.00). PbT produces the most compact, on-target summaries, whereas other methods yield longer, more redundant outputs (1.44–1.47). Overall, PbT ranks among the top performers on three of the four metrics, demonstrating robust, on-style summarization. Example summaries are provided in the Appendix A.6.

5.5 Ablation Study

Teacher Analysis Figure 4 tests PbT with 8B and 1B students across six teacher–student pairings. In every setting, PbT surpasses the GPT-synthesis

Table 7: Teacher-side cost on XSum (LLAMA8B), measured on A100 GPUs. PbT requires only IR-level annotation ($m_x + m_y$), substantially reducing teacher effort while improving G-Eval scores.

Pipeline	**Teacher-side Cost (min)**	**Rel. Cost**	**Saving (min)**	**G-Eval (Ave.)**
Direct (full doc)	224	1×	0	2.75
Direct w/ BoN-5	1120	5×	0	2.88
PbT (IR: $m_x + m_y$)	83 (22+61)	**0.37×**	141	3.22
PbT w/ BoN-5	415	1.85×	705	**3.43**

Table 8: Different IR settings for PbT on XSum dataset. Teacher model: LLAMA8B, Student model: LLAMA1B

	Fact.	**Coh.**	**Cont.**	**B-S**
PbT(IR$_{\text{section}}$)	**2.51**	**3.30**	**3.29**	**88.1**
PbT(IR$_{\text{hier}}$)	2.33	2.84	3.10	87.6
PbT(IR$_{\text{CoT}}$)	2.43	2.95	3.12	87.9

baseline. Even with the small 1B teacher guiding an 8B student, BERTScore improves by 2.7 points. For the 1B student, upgrading the teacher from 8B to GPT yields only a 0.1-point gain, indicating that oversized teachers offer limited additional benefit. These findings show that PbT remains effective even with modest teacher models, highlighting the framework's robustness to teacher capacity.

Efficiency Analysis Table 7 compares teacher-side cost and quality across different pipelines on XSum-small. Direct synthesis requires 224 minutes of full-document annotation and achieves a G-Eval score of 2.75. Adding BoN-5 increases the cost fivefold (1120 minutes) while providing only a marginal gain (2.88). In contrast, PbT reduces annotation time to just 83 minutes by operating at the IR level ($m_x + m_y$), achieving a higher score of 3.22. When combined with BoN-5, PbT incurs 415 minutes and reaches the best quality of 3.43. These results demonstrate that PbT reduces teacher effort while improving summary quality.

IR types Analysis Under the LLAMA8B/1B setting, we compared three IR encoding methods—sectioned IR (our default), hierarchical IR, and chain-of-thought IR (Table 8). We used GPT-4o-mini to assess factual consistency, coherence, and content coverage of the generated documents. Results show that sectioned IR provides the highest overall quality, while all three variants remain the summarization scores within a half-point range in B-S, indicating stable performance across designs.

6 Conclusion

Paired by the Teacher (PbT) addresses the shortage of aligned input–output pairs in natural language generation by combining teacher-annotated intermediate representations with student-generated sources. This design converts unpaired corpora into usable supervision without human labeling, while shifting most text generation from the teacher to the student. Across five benchmarks covering document summarization, dialogue summarization, and question generation, PbT consistently surpasses teacher- and student-only baselines. An 8B student trained with PbT data comes within 1.2 ROUGE-L points of human-labeled systems, closing 82% of the gap over the strongest synthetic baseline at only one-third the annotation cost of direct synthesis. On the unlabeled SwitchBoard corpus, PbT further shows better style adherence and fluency than competing approaches. This advantage comes from PbT's in-domain alignment: a student trained on real inputs generates sources close to the true distribution, avoiding the mismatch that limits direct synthesis. PbT thus offers a fast, cost-effective way to train compact models for specialized or privacy-sensitive applications, even when labeled input–output pairs are scarce.

Limitations

While synthesizing data, we train the student to reconstruct the source text in an open-ended manner. PbT generally introduces fewer hallucinations than directly synthesizing from the target. However, because the IR inevitably omits some details, the generated sources and summaries still require careful human review in high-fidelity applications. All current experiments use English corpora only. We have not yet studied other languages or cross-lingual scenarios such as summarizing documents or dialogues written in a different language.

Ethical Considerations

As we extend the frontiers of NLP, we remain attentive to the ethical dimensions of our work. Generating synthetic data with large-scale models reduces our reliance on vast real-world datasets that could include personal or sensitive information. To uphold privacy and meet regulatory standards, we draw exclusively on openly available corpora under permissive terms (CC BY-NC-ND 4.0 and CC BY-NC-SA 4.0) and enforce stringent anonymization measures. When PbT is applied to copyrighted web or news material, the synthetic sources can echo portions of the originals. Practitioners should verify redistribution rights or ensure that outputs are sufficiently transformed. For conversational data, intermediate representations may reveal private entities or sensitive details, so names should be anonymized before teacher annotation. Finally, because PbT makes it easier to create large paired datasets, the same pipeline could be exploited to mass-produce malicious content. To limit misuse, we will release the code under a research-only license.

References

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. Seq3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of NAACL-HLT*, pages 673–681.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. Dialogsum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *CoRR*.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Qipeng Guo, Zhijing Jin, Xipeng Qiu, Weinan Zhang, David Wipf, and Zheng Zhang. 2020. Cyclegt: Unsupervised graph-to-text and text-to-graph generation via cycle training. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 77–88.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 8142–8152. Association for Computational Linguistics (ACL).

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yen-Ju Lu, Ting-Yao Hu, Hema Swetha Koppula, Hadi Pouransari, Jen-Hao Rick Chang, Yin Xia, Xiang Kong, Qi Zhu, Xiaoming Simon Wang, Oncel Tuzel, and 1 others. 2025. Mutual reinforcement of llm dialogue synthesis and summarization capabilities for few-shot dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7237–7256.

Nishant Mishra, Gaurav Sahu, Iacer Calixto, Ameen Abu-Hanna, and Issam H Laradji. 2023. Llm aided semi-supervision for extractive dialog summarization. *arXiv preprint arXiv:2311.11462*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Gaurav Sahu and Issam H Laradji. 2024. Mixsumm: Topic-based data augmentation using llms for low-resource extractive text summarization. *arXiv preprint arXiv:2407.07341*.

Gaurav Sahu, Olga Vechtomova, and Issam H Laradji. 2025. A guide to effectively leveraging llms for low-resource text summarization: Data augmentation and semi-supervised approaches. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1584–1603.

Rohan Surana, Junda Wu, Zhouhang Xie, Yu Xia, Harald Steck, Dawen Liang, Nathan Kallus, and Julian McAuley. 2025. From reviews to dialogues: Active synthesis for zero-shot llm-based conversational recommender system. *arXiv preprint arXiv:2504.15476*.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and EngSiong Chng. 2025. Diasynth: Synthetic dialogue generation framework for low resource dialogue applications. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 673–690.

Y-S Wang and HUNG-YI LEE. 2020. Learning to encode text as human-readable summaries using generative adversarial networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*.

Kaige Xie, Tong Yu, Haoliang Wang, Junda Wu, Handong Zhao, Ruiyi Zhang, Kanak Mahadik, Ani Nenkova, and Mark Riedl. 2024. Few-shot dialogue summarization via skeleton-assisted prompt transfer in prompt tuning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2408–2421.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. Ted: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Eyal Shnarch, and Leshem Choshen. Achieving human parity in content-grounded datasets generation. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Training Settings

Our experiments mirror the setup in (Lu et al., 2025) for both data synthesis and summarization. We train with mini-batches of 10 and linearly ramp the learning rate up to 2.0×10^{-4} over the first 50 batches. A ReduceLROnPlateau scheduler (patience = 5, factor = 0.7) handles learning-rate adjustments, and we invoke early stopping if the loss plateaus for 100 steps. Since the paired data are not available, we also generate the source data from the validation summaries to form the synthetic validation set for learning rate adjustment and model selection for summarization and question generation.

A.2 Medium Resource Experiments

We include the medium resource (5% dialogues and 1,000-shot summaries) experiments of XSum (Table 9), SAMSum (Table 10), and CNNDM (Table 11). Consistent improvement compared to the ICL and direct synthesis baselines can be observed, as in the small resource experiments.

A.3 Prompts for Intermediate Representation Extraction

We provide all the prompts for IR extraction from both sources and targets in Table 12 and Table 13, including dialogue summarization, document summarization, and question generation. Also, the

Table 9: **XSum** (Medium, 1,000-shot unpaired documents): Comparison of methods on medium-source setting

Synthesis model	R-2	R-L	BERTScore
Zero shot	5.3	12.5	86.2
ICL (summary only)	7.9	18.2	87.6
Unpaired Sources			
GPT$_{\text{direct}}$ ($\mathcal{T}$)	5.3	16.6	86.3
LLAMA8B$_{\text{direct}}$ ($\mathcal{S}$)	6.6	18.6	87.0
PbT$_{\text{8B, w/ GPT}}$	**15.7**	**29.0**	**89.6**
Oracle Paired			
Real Data	18.9	32.7	90.6

Table 10: **SAMSum (736-shot unpaired dialogues)**: Comparison of methods on medium-source setting.

Synthesis model	R-2	R-L	B-S
Zero-shot	12.9	25.1	88.3
ICL (summary only)	18.5	33.8	90.4
Unpaired Sources			
GPT$_{\text{direct}}$	23.7	38.3	91.0
LLAMA8B$_{\text{direct}}$	21.5	36.5	90.8
DiaSynth	16.0	29.8	89.5
PbT$_{\text{8B, w/ GPT}}$	**25.9**	**41.6**	**91.7**
Oracle Paired	28.0	43.3	92.0

prompts for synthesizing source data from IR are shown in Table 14.

A.4 Human Evaluation Details

We recruited six graduate students to evaluate our dialogue summaries, with all participants providing informed consent for their contributions to be reported. Following the evaluation protocols established by (Lu et al., 2025; Xie et al., 2024), annotators assessed each summary along four dimensions: faithfulness, fluency, informativeness, and redundancy. The guidelines for evaluation are outlined in Table 15.

A.5 Summarization and Question Generation Examples

In this section, we show the source and generated target examples for each of the three tasks: dialogue summarization, document summarization, and question generation. The results are in Table 16, Table 17, and Table 18. We also show the SwitchBoard dialogues summarization results with the target summaries examples in Table 19 and Table 20.

Table 11: **CNNDM (Medium, 1,000-shot unpaired documents)**: Comparison of methods on medium-source setting.

Synthesis model	R-2	R-L	B-S
Zero-shot	13.1	21.0	86.8
ICL (summary-only)	15.2	24.7	87.5
Unpaired Sources			
GPT$_{\text{direct}}$	16.6	23.6	86.6
LLAMA8B$_{\text{direct}}$	17.3	25.1	86.9
PbT$_{\text{8B, w/ GPT}}$	**17.3**	**27.3**	**87.6**
Oracle Paired	**17.5**	26.2	87.2

A.6 Document Generation Examples

In this section, we show the real document and synthetic examples of XSum in Table 22 with IR-based PbT generation and directed synthesis baseline from the teacher. We found the direct synthesis baseline generates text like a polished news article: it relies heavily on third-person narration, employs a wide range of complex sentence structures, and features frequent shifts in subjects and verb tenses. In contrast, PbT-generated text is straightforward, declarative sentences that mirror the style of real XSum data, using simple syntax and consistent phrasing to convey information directly.

Dialogue IR: You are an expert at segmenting dialogues and writing concise summaries. Given a raw dialogue, divide it into segments and write one sentence of summary for roughly every 2–3 dialogue turns. The number of segments should adjust to the dialogue length. Guidelines 1. Order: Follow the original flow of the conversation. 2. Density: Aim for 2–3 turns → 1 summary sentence. 3. Clarity: Each summary starts with “Segment N:” and is a single, clear sentence. 4. Content: Capture who does what, key facts, and any notable reactions or tone. – Mention speaker names when it aids clarity. – Do not quote full lines. 5. Output only the segment summaries: no additional commentary. Output Format Segment 1: … Segment 2: … Segment 3: …
Document IR: You are an expert in document segmentation and summarization. Given an entire document, divide it into 3–7 coherent sections (introduction, key topics, conclusion, etc.). If any section is under 80 words, merge it with the previous one. For each section, output exactly: Section n Summary: <1–2 sentences (20–30 words) capturing who/what, where, main action/result> Entities: <comma-separated list of key people, places, organisations> (blank line)
Question IR: You are an expert information extractor. You will be given a paragraph and an answer span. Produce only the following 1) A single masked sentence with the answer replaced by [BLANK]. 2) Write 3–5 bullet points (•) capturing related facts (entities, actions, dates, numbers). Do NOT include or repeat any other text from the paragraph. Output format: <masked sentence> • bullet 1 • bullet 2 • bullet 3

Table 12: Prompts for IR extraction from source

Dialogue IR: You are an expert in dialogue reconstruction. Given a global summary of a dialogue and the approximate number of original dialogue lines, produce a plausible set of segment summaries that would let a reader rebuild the dialogue flow.
Guidelines:
1. Segment density: 1 summary sentence for every 2–3 dialogue lines.
2. Order & flow: Follow the chronology implied by the global summary.
3. Speaker cues: Mention speaker names when helpful, but do not quote full lines.
4. Content: Capture key actions, questions, answers, emotions, and shifts.
5. Format: Output only the segment summaries—no extra commentary.
Format:
Segment 1: . . .
Segment 2: . . .
. . .
6. Output only the segment summaries — no extra commentary or restating the global summary.

Document IR: You are an expert news editor. Given a global summary of an article and the article's approximate total length in words Produce a set of segment summaries that partition the article logically Guidelines:
1. Density: Aim for 1 segment summary per 80-120 words
2. Order & chronology: Preserve the timeline of events as implied by the global summary.
3. Content: Each segment summary (1-2 sentences) must cover key facts, actors, actions, and transitions; avoid speculative or external information.
4. Entities line: After each summary, list the main named entities (people, organisations, locations) separated by commas.
5. Format strictly as:
Section 1
Summary: <concise sentence(s)>
Entities: <Entity A>, <Entity B>, . . .
Section 2 . . .
6. Output only the segment summaries — no extra commentary or restating the global summary.

Question IR: You are an expert content planner.
Given a Question and the Answer, produce only the following:
1) One masked sentence that answers the question, with the answer replaced by [BLANK].
2) Exactly 3–5 bullet points that would help a writer expand the idea into a full paragraph.
Guidelines:
• The masked sentence must be grammatically correct and fully answer the question when [BLANK] is filled.
• Each bullet <= 20 words and begins with •.
• Bullets should cover key entities, events, dates, figures, causes, etc.
• Do NOT repeat or rephrase any part of the question text.
• Output only the masked sentence and bullets—no extra commentary.

Table 13: Prompts for IR extraction from target

Dialogue: You are a dialogue-generation model. Please follow these rules:
1. You will be given segment summaries.
2. Based solely on these summaries, reconstruct the conversation.
3. Each turn starts with a speaker’s name exactly as provided, followed by a colon.
4. Target length: about {m} turns and {N} words.
5. Do not output the segment summaries again or invent content beyond what is implied.
6. Output must include a single section titled ’=== Dialogue Begins ===
7. Ensure the dialogue reads naturally and fluently.

Document: You are a professional news-style writer. Follow these rules:
1. You will receive segment summaries describing an article.
2. Rely only on the information in those summaries—do not add facts that are not implied.
3. Write a coherent, well-structured document in clear journalistic prose.
4. Organise the text into logical paragraphs; you may merge or split ideas as long as all summary details are covered.
5. Target length: around {N} words
6. Do not repeat the segment-summary bullets in the output.
7. Start your answer with exactly the line: === Document Begins ===
8. End with a newline; no extra commentary after the article.

Question: You are a professional writer. Follow these rules:
1. You will receive 3-5 bullet points (IR) and the answer span.
2. Rely only on the information in those bullet points—do not add facts that are not implied.
3. Write ONE coherent paragraph that uses all the bullet facts and naturally includes the answer.
4. Target length: around{N} words.
5. Do not repeat the bullets in the output.
6. Start your answer with exactly the line: === Paragraph Begins ===
7. End with a newline; no extra commentary after the paragraph.

Table 14: Prompts for student IR inversion in both training and testing.

Faithfulness
Whether the information presented in the summary is factually incorrect or unmentioned according to the dialogue: *0: lots of the information presented in the summary is factually incorrect or unmentioned; *1: a small amount of the information presented in the summary is factually incorrect or unmentioned; *2: no information presented in the summary is factually incorrect or unmentioned.
Fluency
Whether the sentences in the summary are ungrammatical or ill-formed: *0: lots of the sentences in the summary are ungrammatical or ill-formed; *1: a small amount of the sentences in the summary are ungrammatical or ill-formed; *2: no sentence in the summary is ungrammatical or ill-formed.
Informativeness
Whether the critical information in the dialogue is missed in the summary: *0: lots of the critical information in the dialogue is missed; *1: a small amount of the critical information in the dialogue is missed; *2: no critical information in the dialogue is missed.
Redundancy
Whether the expressions of the summary can be simplified: *0: lots of the expressions of the summary can be simplified; *1: a small amount of the expressions of the summary can be simplified; *2: no expression of the summary can be simplified.

Table 15: Human evaluation metrics for Faithfulness, Fluency, Infomativeness, and Redundancy, and their corresponding instructions

Input Dialogue
Person1: What are you working on? Person2: I'm figuring out my budget. Person1: What budget? Person2: I'm making a shopping budget, so that I don't spend too much money. Person1: How much money can you spend? Person2: I can only spend $ 300 a month. Person1: Why only $ 300? Person2: I need to save the rest. Person1: For what? Person2: I need to pay my bills. Person1: Your budget is a good idea. Person2: I know. It's going to save me a lot of money, I hope.
Ground Truth Summary
Person2 is making a shopping budget to save the rest of the money. Person1 thinks it's a good idea.
Generated Summary (Zero Shot)
Person 1 asks Person 2 what they are working on, and Person 2 responds that they are creating a shopping budget to avoid overspending. Person 1 asks how much they can spend, and Person 2 says $300 per month. Person 1 asks why they can only spend $300, and Person 2 explains that they need to save the rest to pay their bills. Person 1 thinks the budget is a good idea, and Person 2 hopes it will help them save money.
Generated Summary (In-context Learning)
Person1 asks Person2 what they are working on, and Person2 explains that they are creating a budget to control their spending. Person2 sets a limit of $300 per month and plans to save the rest for paying bills. Person1 praises Person2's budgeting idea, which Person2 hopes will help them save money.
Generated Summary (Direct Synthesis, GPT)
Mary is making a shopping budget so that she doesn't spend too much money. Person1 asks her how much money she can spend and why only $ 300. Mary says she needs to save the rest to pay her bills.
Generated Summary (PbT, Ours)
Person2 is making a budget to save money. Person1 thinks it's a good idea.
Generated Summary (Real Data)
Person2 is making a shopping budget so that Person2 doesn't spend too much money.

Table 16: DialogSum Summarization results from PbT and baselines.

Input document
Their recent run is in contrast to the way they started the season with six victories from their first seven games.Since the end of March, the Vikings' only wins have been in the Challenge Cup against lower-league sides."We've got the personnel and we've got the people to spark us back into life," Betts told BBC Radio Merseyside."When we get rolling again I'm sure, or I'm positive, that we can really turn this year around for ourselves."Widnes lie eighth after their poor run and face Salford at Magic Weekend on Saturday."The players are hurting and we've got to win," added England assistant coach Betts. "Whether it be pretty, whether it be the last second of the game, we've got to win."It's alright being a good group and it's alright talking about team spirit and being in it together, but not winning really hurts."We've got to find that win and we've got to find it some way and that's within the team and it's in those 80 minutes out there."
Ground Truth Summary
Widnes head coach Denis Betts is confident his squad will find their form after a seven-match losing streak in Super League.
Generated Summary (Zero Shot)
Widnes Vikings, an English rugby league team, have struggled since the end of March, winning only against lower-league teams in the Challenge Cup. Despite having the talent and team spirit, they are currently in 8th place and need to turn their season around. The team's assistant coach, Betts, emphasizes the importance of winning, stating that it's not just about being a good team, but about getting results. They face Salford at Magic Weekend on Saturday and are determined to find a way to win.
Generated Summary (In-context Learning)
Widnes Vikings are struggling in the season, having only won in the Challenge Cup against lower-league teams since the end of March. Assistant coach Betts is confident that the team can turn their season around, citing their personnel and team spirit as reasons for optimism. They face Salford at Magic Weekend on Saturday.
Generated Summary (Direct Synthesis, GPT)
Their run of six wins from their first seven games is the opposite of the way they started the season with six from their first seven.
Generated Summary (PbT, Ours)
Widnes Vikings coach Denis Betts says his team are "hurting" after a run of 10 Super League defeats in a row.
Generated Summary (Real Data)
Widnes Vikings head coach Denis Betts says he is confident his side can turn their season around after a run of 10 Super League defeats.

Table 17: XSum Summarization results from PbT and baselines.

Input paragraph and Answer Span
Paragraph: In October 2010, the open-access scientific journal PLoS Pathogens published a paper by a multinational team who undertook a new investigation into the role of Yersinia pestis in the Black Death following the disputed identification by Drancourt and Raoult in 1998. They assessed the presence of DNA/RNA with Polymerase Chain Reaction (PCR) techniques for Y. pestis **[from the tooth sockets in human skeletons]** from mass graves in northern, central and southern Europe that were associated archaeologically with the Black Death and subsequent resurgences. The authors concluded that this new research, together with prior analyses from the south of France and Germany, ". . . ends the debate about the etiology of the Black Death, and unambiguously demonstrates that Y. pestis was the causative agent of the epidemic plague that devastated Europe during the Middle Ages". **Answer:** from the tooth sockets in human skeletons
Ground Truth Question
Where did scientists find their Y. pestis sample?
Generated Question (Zero Shot)
Where did a multinational team use Polymerase Chain Reaction (PCR) techniques to assess the presence of Y. pestis DNA/RNA?
Generated Question (In-context Learning)
Where did a multinational team assess the presence of DNA/RNA with Polymerase Chain Reaction (PCR) techniques?
Generated Question (Direct Synthesis, GPT)
Where did the new investigation of Yersinia pestis take place?
Generated Question (PbT, Ours)
Where did the researchers assess the presence of DNA/RNA with PCR techniques for Y. pestis?
Generated Question (Real Data)
Where did the researchers look for Yersinia pestis?

Table 18: SQuAD Question Generation results from PbT and baselines.

Input Dialogue (SwitchBoard, Unpaired)
A: Hi, my name is Donna, and Im calling, calling from Plano, Texas.
B: Im Jay in Dallas.
A: Oh, okay, um, I, um, find exercise kind of a task, I guess Id say. I go to aerobics two days a week, but I also have three kids at home so they keep me pretty busy.
B: That keeps you plenty of exercise .
A: Yeah, that gives me plenty of exercise, but I enjoy, um, walking outside and bicycling once in a while with the family. Excuse me . Thats fine. And yourself.
B: Well, I, uh, I have, - I walked and run ev-, um, almost every morning. Ive sort of stopped here with the weather change, but, uh, its about a thirty minute exercise every morning.
A: Uh-huh, every morning. Well, my husband swims every day at lunch, which he loves doing –
B: Oh, that would, - thats fantastic exercise.
A: – yeah, he does it every day like a ritual. Once in a while or if he has a business meeting or if he, um, someone invites him to lunch, hell, hell not go, but he goes, um, every, every day, and then he ru-, hes a runner, too. He runs three mornings a week he gets up about five and, and –
B: Um.
A: – does his, his deal, and some days he doesnt, but then maybe hell catch up on the weekend and go for a run.
B: All right.
A: And, um, we . -
B: Id rather get a bicycle, but I have an ex-, I had an exercise bike in the house, but didnt seem to use it enough, so somebody sold it in a garage sale, and so, uh –
A: Oh.
B: – but Id like to get a bike and just do, uh, do biking outside.
A: The, the, the bicycling is nice especially with the weather around here, and I know, um, uh, we, we both of, my husband and I both do it to just keep the pounds off. Were getting up in age, I guess, in our late thirties so we need to –
B: Yeah.
A: – keep exercising just to be able to, uh. -
B: It keeps going, because Im fifty-five, and it, uh, it just gets worse. You still want to eat, and, uh –
A: Oh, I know.
B: – but you just, - I think you feel better, too, if you get on some program like that, but, uh. -
A: Yeah, I go to this aerobics class two mornings a week, and I take my three year old with me. Theres a nursery there. And its kind of another day out for her –
B: Uh-huh.
A: – to, um, have little friends, and, um, I, I enjoy the socialization of it, you know, getting to see the same women every –
B: Right.
A: – morning, but um, there are some parts during that exercise that I think, I cannot do another –
B: Oh.
A: – second of this.
B: I know it, it, it, some of it is so tough –
A: It is.
B: – and, uh, especially when somebodys, uh, planning lunch for you and they know what, whats good for you, I think when we do it ourselves, we take the easy way sometimes.
A: Thats true, thats true.

Table 19: Input dialogue from SwitchBoard used for summarization evaluation.

Example Summaries (DialogSum, Unpaired)
B sleeps through a movie and gets woken up by A. B thinks the movie is unrealistic while A thinks it was great. A doubts how B knows. B doesn't answer and suggests going to bed.
Generated Summary (Zero Shot)
The conversation is between Donna from Plano, Texas, and Jay from Dallas, Texas, who are participating in a phone survey about their exercise habits. Here's a summary of their conversation: * Donna mentions that she finds exercise a task, but enjoys walking and bicycling with her family. * Jay says he used to walk and run every morning, but has stopped due to the weather change. * Donna's husband swims every day at lunch and runs three mornings a week. * Jay wants to get a bicycle to do outdoor biking, but his exercise bike was sold in a garage sale. * Both Donna and Jay mention that they exercise to keep the pounds off and stay healthy, especially as they get older. * Donna takes her three-year-old daughter to an aerobics class with a nursery, which provides socialization for her child. * Both Donna and Jay agree that some exercises can be tough, and it's easy to take the easy way out when planning meals. * They wrap up the conversation by thanking each other for participating in the survey.
Generated Summary (In-context Learning)
Donna from Plano, Texas, and Jay from Dallas discuss their exercise routines. Donna mentions that she goes to aerobics classes two days a week and enjoys walking and bicycling with her family. Jay says he used to walk and run every morning but has stopped due to the weather change. Donna's husband swims every day at lunch and runs three mornings a week. They both agree that regular exercise is important, especially as they get older, to keep the pounds off and feel better.
Generated Summary (Direct Synthesis, GPT)
Donna calls from Plano, Texas. Donna finds exercise a task. She goes to aerobics two days a week and has three kids at home so they give her plenty of exercise. Donna enjoys walking outside and bicycling once in a while with the family. B says B has walked and run almost every morning. B's husband swims every day at lunch which is fantastic exercise. B's husband is a runner and runs three mornings a week. B wants to get a bicycle but B had an exercise bike in the house which somebody sold in a garage sale.
Generated Summary (PbT, Ours)
Donna and Jay talk about their exercise habits. Donna goes to aerobics two times a week and enjoys walking and bicycling with her family. Jay runs every morning and wants to get a bicycle.
Generated Summary (DialogSum Real Data)
Donna from Plano, Texas calls Jay in Dallas. Donna has three kids at home so they keep her pretty busy. She enjoys walking outside and bicycling once in a while with the family. Her husband swims every day at lunch and he runs three mornings a week. They both do bicycling to keep the pounds off. Jay has an exercise bike in the house but didn't seem to use it enough. Jay likes to get a bike and just do biking outside.

Table 20: Summarization results from PbT and baselines on the SwitchBoard dialogue in Table 19.

Target summary
US actor Randy Quaid is due to appear in court in Vermont on Monday over a California vandalism case.
Real Document (A news-reporting style with a neutral tone, using precise descriptions of locations, times, and amounts)
Quaid, 65, and his wife, Evi, were taken into custody on Friday night by Vermont police while trying to enter the US from Canada. They fled to Canada in 2010 after facing charges of vandalism and squatting in a Santa Barbara property. But his request for refugee status, in which he claimed his life was in danger from organised crime, had been denied. Quaid was recently informed he would be sent back to the US to face criminal charges. His wife has dual US-Canadian citizenship. The Quaids are alleged to have caused more than $5,000 (£3,253) in damage to a property in Santa Barbara, which they claim belongs to them. Santa Barbara officials have previously said they would arrest Quaid if he returned to the US. At a hearing in 2010, he told a Canadian court he was being targeted by "Hollywood star whackers", and that eight of his fellow actors had been "murdered in odd, strange manners", including Heath Ledger and David Carradine. Quaid, the older brother of fellow actor Dennis, is best known for his roles in such films as Kingpin, National Lampoon's Vacation and The Last Detail. He and his wife Evi have been posting bizarre expletive-ridden videos on their YouTube channel, denouncing the US authorities' attempts to force them back.
Generated Documents (Direct Synthesis, GPT) (Adopts a commentary-and-interpretation tone, inserting embellishments like “media attention” and “Fans and critics,” which clash with the original's investigative news style.)
US actor Randy Quaid is set to appear in a Vermont court on Monday as part of ongoing legal troubles stemming from a vandalism case in California. The 73-year-old actor, known for his roles in films such as "National Lampoon's Vacation" and "Independence Day," was arrested in 2020 after allegedly causing damage to a home in California. The incident sparked significant **media attention**, particularly due to Quaid's erratic behavior and public outbursts in recent years. In a series of social media posts, Quaid has maintained his innocence, claiming he was targeted unfairly. His legal team has argued that the vandalism charges are exaggerated and stem from a personal dispute. The actor has previously faced several legal issues, including arrests related to trespassing and other charges. Quaid's upcoming court appearance in Vermont relates to a separate matter, but his past incidents continue to cast a shadow over his career. **Fans and critics** alike have expressed concern over his well-being, with some calling for him to seek help for mental health issues. As the court date approaches, many are watching closely to see how this high-profile case will unfold. Legal experts suggest that Quaid's history may affect the outcome of the proceedings. Regardless of the court's decision, the actor's reputation and future in Hollywood remain uncertain.

Table 21: XSum Document Synthesis examples results from PbT and baselines (Part 1).

Generated Documents (PbT, Ours) (Direct stating tone, similar to the real XSum document.)
Randy Quaid, best known for playing the evil brother Lyle in National Lampoon's Vacation, has also made headlines recently in California where his wife Evi is charged with looting two hotels while she was facing foreclosure. Quaid and his wife face charges over \$5m (£3.15m) damages at several hotels they stayed in. According to Vermont officials, Randy Quaid would fly up to attend a court hearing if needed. The court date concerns a charge of vandalism by a tenant who had been locked out or excluded from a rented building. The allegations come under Vermont state law but were originally made in California. Vermont Attorney General William Sorrell said it did not matter where Mr Quaid lived, as he was being sued there because of events alleged to have happened there. Prosecutors allege the couple broke windows and doors at several homes including in Marin County and in San Francisco. Mr Quaid is also accused of leaving a threatening note saying "you're a victim", authorities added. The couple lost their home in Tiburon last year. In December, the sheriff's office took Mrs Quaid and her husband into custody and booked them at jail on suspicion of burglary and vandalism - the same charges prosecutors said they planned to pursue. Their children and grandchildren still live in the properties the couple are accused of damaging, according to media reports. They had moved to Canada to work out a possible deal allowing them to stay temporarily. But California lawyers say they failed to complete what they promised and return home.

Table 22: XSum Document Synthesis examples results from PbT and baselines (Part 2).