No Audio, No Transcripts, No Problem: LLM-based ASR Adaptation Using Only Domain Documents

Sergio Burdisso[1], Dairazalia Sánchez-Cortés[1], Esaú Villatoro-Tello[1], Thibault Bañeras-Roux[1],
Manjunath K E[2], Kadri Hacioglu[2], Petr Motlicek[1,3], Andreas Stolcke[2]
[1]*Idiap Research Institute, Martigny, Switzerland* [2]*Uniphore, U.S.A.* [3]*Brno University of Technology, Brno, Czech Republic*

***Abstract*—Can a speech recognizer be adapted to a new domain with no target audio and no transcripts, using only written documentation? This is the realistic onboarding setting: a new customer hands over raw documentation, not the transcripts that text-only adaptation normally assumes. We turn a domain's documents into adaptation text in two ways, directly as written sentences or rendered by a large language model (LLM) into synthetic spoken dialogs, and feed them to a recently proposed denoising-based text-only adaptation method. We show that, across in-domain, out-of-domain, and cross-domain settings on two corpora, document-only adaptation can match or closely approach real transcriptions; whether written or conversational rendering is best depends on speaking style and the amount of text generated. To our knowledge, this is the first study of document-only adaptation of LLM-based ASR. As no existing corpus pairs audio, transcripts, and documents, we also built and released a document-grounded benchmark.**

***Index Terms*—text-only adaptation, domain adaptation, LLM-based ASR, synthetic data, document-grounded generation.**

I. Introduction

LLM-based ASR systems, which connect a pretrained speech encoder to a frozen large language model (LLM) through a lightweight learnable projector, have become a powerful approach to transcription [1]–[6]. However, these systems still degrade under domain shift and transcribing target-domain audio for model adaptation is expensive. *Text-only adaptation*, which specializes the system using text from the target domain, is therefore an appealing alternative given the abundance of text relative to transcribed speech [7].

A recurring and largely unexamined assumption underlies this line of work: the available target-domain text consists of *ground-truth transcriptions* of spoken conversations. Prior text-only adaptation methods for conventional end-to-end ASR [8]–[10] and for LLM-based ASR [11]–[15] all adapt on transcripts. Yet in the most realistic onboarding scenario (a new client hands over their materials so an ASR can be delivered for their domain), transcripts of in-domain calls are exactly what is *not* available. What is available is *raw documentation*: company websites, product and service descriptions, manuals, policies, etc. Obtaining transcripts instead presupposes the very audio collection and annotation that text-only adaptation is meant to avoid. Beyond practicality, document-based adaptation suits privacy-aware deployment: adapting from a company's existing documentation avoids recording and transcribing sensitive customer conversations.

All this motivates the central question of this paper: *can an ASR system be adapted to a new domain using only raw documents, with no target-domain audio and no target-domain transcripts?* We answer this question in the affirmative. Our key insight is that documents and transcripts differ not only in content but in *form*: a document states facts in written prose, whereas a transcript realizes those facts as spontaneous, often disfluent, turn-based speech. We therefore turn the domain's documentation into adaptation text in two ways: used directly as written sentences or rendered by an LLM into synthetic spoken dialogs (a conversational grounded in the document), and feed them to the multi-view denoising adaptation method of [13]. We find that document-only adaptation is competitive with ground-truth transcriptions across domains and regimes, and that whether written or conversational rendering is preferable depends on the target's speaking style (*form*) and the amount of generated text (*quantity*). To expose the *form* dependence, we evaluate on two deliberately contrasting corpus types, spontaneous customer–agent conversations and presentation-style talks and lectures: their registers sit at opposite ends of the speaking-style spectrum, with the conversational corpus matching the synthetic dialogs and the presentation corpus being closer in style to the original documents. Our contributions are: (a) We open up a *new regime for text-only adaptation*, framed by the question above: adapting from a domain's written documentation alone. This is the most challenging case, as well as the most realistic, based on what a new client can provide; to our knowledge this scenario has not been studied before; (b) We show this *regime is viable*: rendering documents into adaptation text, as their own written sentences or as zero-shot LLM-generated dialogs, recovers most or all of the real-transcription gain across in-, out-of-, and cross-domain transfer, using two distinct speech genres; (c) We *characterize when each rendering is best* across the different settings, showing the choice turns not on speaking style (*form*) alone, but also on the amount of adaptation text (*quantity*); (d) We release a *document-grounded benchmark*: as no public corpus pairs audio, transcripts, and documents, we reconstruct the document side from transcripts under a coverage constraint and release aligned audio–transcript–document triples for SlideSpeech [16], along with the code.[1]

II. Related Work

Text-only adaptation of ASR. For conventional end-to-end ASR, text-only adaptation has been approached via

[1]https://github.com/idiap/llm-asr-document-only-adaptation

intermediate-CTC adaptation [8], external language-model fusion [17]–[19], text-to-spectrogram generation [9], and decoder adaptation for Whisper [10]. For LLM-based ASR, fine-tuning the LLM on text alone disrupts the speech–text alignment learned by the projector, causing catastrophic forgetting [11]–[13]. Fang et al. [11] mitigate this by monitoring alignment and stopping before collapse, while Ma et al. [12] learn soft-prompt embeddings to replace the speech input. The multi-view denoising approach of [13], our adaptation backbone, preserves alignment by mixing source audio with synthetically noised text per batch. Crucially, all of these methods adapt on *ground-truth transcriptions*; none consider raw documents as the only available signal.

Synthetic conversational data generation. A growing body of work generates synthetic conversations to overcome the data scarcity and privacy constraints of real call-center corpora: eliciting natural customer-support dialogs [20], converting raw recordings into structured datasets [21], recognizing entities in phone-call transcripts [22], and synthesizing dialogs for downstream systems [23], with further work on adaptive prompting [24], multi-document grounding [25], and user-diversity simulation [26], [27]. These methods aim to build or augment *dialog systems*, and their grounded variants typically rely on retrieval and passage segmentation [25], [28], [29]. In contrast, our purpose is *ASR domain adaptation*: we generate spoken-style transcripts directly from raw documentation, with no retrieval or passage segmentation. We find that simple zero-shot generation over the whole document is highly effective for this task, as it preserves discourse coherence and avoids retrieval-stage errors.

Synthetic data for ASR. A long line of work reduces the audio burden in ASR by synthesizing data, e.g., text-to-speech augmentation and text-conditioned generation of acoustic representations [9], [30]–[33]. Our setting differs in both *objective* and *constraint*: we do not augment an existing in-domain text distribution, since no target-domain spoken text exists at all, and we must *induce* a spoken distribution from non-spoken documents, evaluating the result by its *downstream value* for text-only ASR adaptation rather than by generation quality alone.

III. Method: Adapting an ASR from Documents

A. Problem setting

Let $\mathcal{D}_{\text{src}} = \{(a_i, t_i)\}$ be a source-domain corpus of paired audio a_i and transcripts t_i, used to train the base LLM-based ASR. For the target domain we are given *only* a collection of raw documents $\mathcal{C}_{\text{tgt}} = \{d_1, \ldots, d_M\}$, with no target audio and no target transcripts. The goal is to adapt the system to the target domain using $\mathcal{D}_{\text{src}}$ and $\mathcal{C}_{\text{tgt}}$ alone. This is strictly harder than the standard text-only setting, which assumes a target transcript set $\mathcal{D}_{\text{tgt}} = \{t_j\}$.

B. From documents to synthetic transcripts

We instantiate the missing transcripts $\mathcal{D}_{\text{tgt}}$ by generating synthetic conversational transcripts that are faithful to the documents' content but realized as spontaneous speech. To produce one transcript, we sample a document from the collection $\mathcal{C}_{\text{tgt}}$ and prompt an instruction-tuned LLM (Gemma 4:31B) to generate a telephone call between an Agent and a Customer for a call center operating in the domain described by that document. The generation prompt enforces four properties that make the output usable as ASR adaptation text (Fig. 1):

1) **Document grounding.** All factual content (product and service names, plans, policies, fees, procedures) must be drawn exclusively from the document, preferring the same terminology; The model is encouraged to avoid introducing new facts beyond those in the document.
2) **Turn-taking-like segmentation.** Output is segmented into short speech units rather than full sentences, with irregular, non-alternating speaker sequences and standalone backchannels, reflecting the turn-taking structure of conversational speech as produced by ASR segmentation.
3) **Spontaneous speaking style.** The text contains restarts,

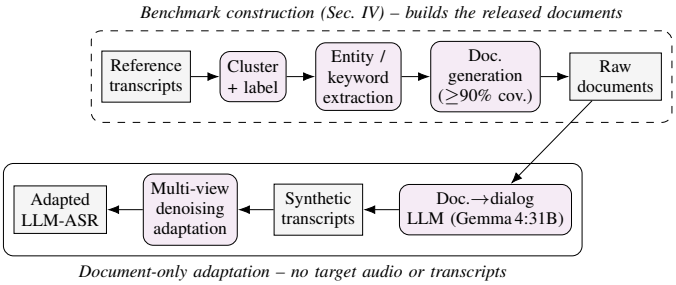

Fig. 1. Overview. *Top (dashed), benchmark construction:* no corpus pairs audio, transcripts, and documents, so we synthesize the document side offline from reference transcripts and release it (Sec. IV); this only builds the benchmark, and the adaptation method never sees these reference transcripts. *Bottom, document-only adaptation:* the method uses *no target audio or transcripts*, taking only the target's documents as input (here rendered into dialogs).

Source document (excerpt). "[. . .] The SilverShield Premier plan covers in-patient hospitalization up to $50,000 per year. A 30-day waiting period applies before claims can be submitted. To file a claim, customers [. . .]"

Generated dialog (excerpt).
AGENT: THANKS FOR CALLING SILVERSHIELD HOW CAN I HELP
CUST: HI UM YEAH I I WANTED TO ASK ABOUT THE PREMIER PLAN
AGENT: SURE
AGENT: SO THE SILVERSHIELD PREMIER PLAN IT COVERS HOSPITAL STAYS
AGENT: UP TO FIFTY THOUSAND DOLLARS A YEAR
CUST: OH OKAY
CUST: AND CAN I CLAIM RIGHT AWAY
AGENT: NOT QUITE THERE'S A THIRTY DAY WAITING PERIOD FIRST
CUST: HMM OKAY GOOD TO KNOW
. . .

Fig. 2. A generated dialog grounded in a source document. "[. . .]" marks omitted document text and the trailing ". . . " indicates the dialog continues. Colored spans mark facts shared by both views, realized verbatim or in normalized spoken form: document facts are preserved but rendered as short, spontaneous, uppercase segments with spelled-out numbers, matching the conversational ASR convention.

repetitions, hesitations, fillers, and other disfluencies characteristic of spontaneous speech rather than planned or read speech.

4) **ASR-style text normalization.** Numbers, dates, and email addresses are spelled out in full spoken form, initialisms such as "USA" and "FBI" are spelled out letter by letter ("U S A", "F B I") while other abbreviations are expanded (e.g. "Dr." as "Doctor"), and match the transcription format of the corpora (e.g. uppercased text, no punctuation).

For each target domain we generate N synthetic dialogs. In our experiments, to compare fairly against adaptation on ground-truth transcripts, we set N to the number of ground-truth dialogs the domain has in our experimental data (Section V). Each of the N generations independently samples a document uniformly at random from $\mathcal{C}_{\text{tgt}}$ and produces one grounded dialog from it; consequently a single document may seed several dialogs, and the number of generated dialogs is *decoupled* from the number of documents M (we do not assume one dialog per document). We match the *number of dialogs* rather than the number of segments or total duration: since each generation yields a full dialog of variable length, the segment count cannot be easily controlled, so dialog count serves as a simple, comparable budget between ground-truth-transcript and synthetic adaptation.

We use zero-shot generation (no in-domain example dialogs given in the prompt), which keeps the method free of any real target transcripts; we analyze the effect of few-shot prompting in Section VI-D. The model is asked to return a single JSON object containing an ordered list of {`speaker`, `text`} segments (speaker $\in$ {AGENT, CUSTOMER}), from which we use each `text` as an individual transcript; the document is supplied in Markdown so the model can exploit its headings and lists to locate key terms. We decode with a relatively high sampling temperature (2.0) to encourage diverse, spontaneous output across the N generations. The resulting synthetic transcripts (*i.e.* all the `text` segments) play the role of $\mathcal{D}_{\text{tgt}}$ in the adaptation step. Fig. 2 illustrates the transformation: the same factual content present in a document is re-expressed as short, disfluent, spoken-style segments with numbers spelled out, a "conversational view" of the document.

C. Adaptation backbone

Given the synthetic target transcripts, we adapt the LLM-based ASR with the multi-view text-denoising method of [13], which we treat as a fixed, previously validated component that is summarized below; see [13] for full details. The method interprets LLM-based ASR as a *denoising* task: the learnable projector $sp(\cdot)$, which maps the speech encoder's frame-level acoustic features for an utterance a into a sequence of vectors in the LLM input-embedding space, produces a sequence of soft tokens that the frozen LLM reads as a noisy, text-like rendering of the transcript t and then "cleans" into t. Text-only adaptation therefore amounts to training the LLM to denoise corrupted transcripts $(noise(t), t)$, with no audio required. Directly fine-tuning on target $(noise(t), t)$ pairs, however, collapses the speech–text alignment (catastrophic forgetting); the method avoids this by composing each mini-batch from four views with proportions $\sigma_a, \sigma_{ta}, \sigma_t, \tau$ that sum to one:

- σ_a: source audio–text pairs $(sp(a), t)$, $(a, t) \in \mathcal{D}_{\text{src}}$;
- σ_{ta}: projector-induced noisy source transcripts $(noise_a(t), t)$, where $noise_a(t)$ maps $sp(a)$ to its nearest LLM-vocabulary tokens (an empirical "optimal" noise);
- σ_t: synthetically noised source transcripts $(noise(t), t)$, $t \in \mathcal{D}_{\text{src}}$ (a cheap, audio-free approximation);
- τ: synthetically noised *target* domain transcripts $(noise(t), t)$, which drive adaptation.

The three source views preserve alignment and avoid forgetting, while τ controls adaptation strength. We use the same training prompt as the backbone, `Transcribe speech to text. Speech: {input}`, with `{input}` set to $sp(a)$, $noise_a(t)$, or $noise(t)$ during training and to $sp(a)$ at inference. We also reuse the backbone's noise function $noise(\cdot)$ (random character substitution then duplication; see [13] for the exact rates). In our setting the target transcripts are the synthetic document-derived dialogs of Section III-B. Whereas the original method [13] sets τ per domain by a size heuristic, we fix $\tau = 0.5$ for *all* experiments (so $\sigma_a = \sigma_{ta} = \sigma_t = (1-\tau)/3 \approx 0.17$), and keep the backbone otherwise unchanged (including noise settings and all optimization/training hyperparameters), so that the *text source* (ground-truth transcripts vs. raw document sentences vs. conversational renderings) is the only factor affecting adaptation. Algorithm 1 summarizes the end-to-end document-only procedure.

Algorithm 1 LLM-based ASR Document-only adaptation.

Require: source corpus $\mathcal{D}_{\text{src}}$; target documents $\mathcal{C}_{\text{tgt}} = \{d_1, \ldots, d_M\}$; pre-trained LLM-based ASR; generation LLM G (conversational rendering)
Render $\mathcal{C}_{\text{tgt}}$ into target adaptation text $\mathcal{D}_{\text{tgt}}$, either:
  Written: the documents' own sentences; or
  Conversational: N dialogs $G(d)$, $d \sim \text{Uniform}(\mathcal{C}_{\text{tgt}})$
for each training step **do**
  build batch from $\mathcal{D}_{\text{src}}$ (views $\sigma_a, \sigma_{ta}, \sigma_t$) and $\mathcal{D}_{\text{tgt}}$ (view τ=0.5)
  update the LLM-based ASR (denoising)
end for
return adapted LLM-based ASR

IV. Building a Document-Grounded Benchmark

Studying document-only adaptation requires, for each domain, documentation that (a) actually covers the domain's spoken content and (b) is paired with audio and reference transcripts for evaluation. No public corpus provides this combination. Crawling the web is also unsuitable here: one of our corpora uses fictitious product and service names that do not exist online, so external documents would not cover the relevant entities. To obtain a *controlled* instance of such documentation, we synthesize documents from each domain's *training* transcripts only, so no validation or test material enters the pipeline. To keep the documents from simply echoing the transcripts they replace, generation receives only *abstracted* signals extracted from those transcripts, i.e. topic labels, keywords, and entities (steps 2–3 below), and never

TABLE I
DATASET STATISTICS FOR DEFINEDAI AND SLIDESPEECH. THE BASE MODEL IS TRAINED ON THE *Source* DATA AND THEN TEXT-ONLY ADAPTED TO EACH *Target* DOMAIN.

	Role	Domain	Train			Val		Test	
			#Dlg	#Utt	Hrs	#Utt	Hrs	#Utt	Hrs
DefinedAI	Source	B/I/H	–	17,398	38.2	–	–	–	–
	Target	Banking	889	26,457	54.4	1,443	2.9	3,164	6.5
	Target	Insurance	1,126	30,720	65.1	1,789	3.9	3,667	7.7
	Target	Healthcare	164	4,702	11.5	236	0.5	488	1.1
	Target	Telco	616	18,835	50.8	1,123	3.1	2,238	6.0
	Target	Retail	354	10,910	27.6	618	27.6	1,300	3.2
SlideSpeech	Source	L/T/E	–	34,682	35.0	–	–	–	–
	Target	Agriculture	51	30,498	29.3	1,679	1.7	3,470	3.4
	Target	Animation	141	56,593	49.8	3,283	2.9	7,005	5.8
	Target	Musical Instr.	20	9,981	8.5	852	0.7	984	0.9

their raw text or phrasing; the documents thus share with the transcripts only the domain vocabulary, precisely what genuine customer documentation would also contain. We stress that this construction is used *only to build the benchmark*; the adaptation pipeline of Section III never sees transcripts of the target domain.

The construction proceeds in four steps (Fig. 1, top). (1) **Topic discovery:** We cluster the dialogs with a BERTopic-style pipeline [34] (class-based TF–IDF), tuning the configuration with Optuna [35] (100 trials) to maximize c_v coherence over preprocessing, vectorization, dimensionality reduction (LSA, NMF, or UMAP [36]), and clustering (KMeans or HDBSCAN [37]). (2) **Topic labeling:** An LLM labels each topic from its representative dialogs and top keywords. (3) **Entity/keyword extraction:** An LLM extracts, per dialog, the salient domain-specific text spans the ASR must potentially learn, deduplicated into a per-topic set: entities (e.g. product, plan, and company names), keywords (e.g. domain terms such as "claim" or "premium"), and factual phrases (e.g. coverage amounts and waiting periods). (4) **Document generation:** An LLM writes a structured Markdown document per topic from the label, keywords, and entities only (never the raw dialogs), integrating as many as possible and introducing no new facts; documents are capped at $9{,}000$ words and diversified by shuffling the entity list. After generation, we verify that at least 90% of the extracted entities and keywords appear in the document, regenerating otherwise, so that each document carries enough domain-relevant terminology. The same Gemma 4:31B model is used throughout. We release the generated documents and synthetic dialogs in the SlideSpeech domains to facilitate reproducibility and further research.

V. EXPERIMENTAL SETUP

ASR system and training configuration. We use the same architecture as the adaptation backbone [13]: the SLAM-ASR framework [1] with a WavLM-Large encoder [38] and a Llama-3.2-3B-Instruct decoder [39], connected by a projector

TABLE II
GENERATED-CORPUS STATISTICS PER TARGET DOMAIN.

Domain	#Document	Written		Conversational		
		#Sent	r_{w}	#Dlg	#Utt	r_{c}
Banking	23	941	0.04	889	43,365	1.64
Insurance	25	2,740	0.09	1,126	48,540	1.58
Healthcare	14	475	0.10	164	7,307	1.55
Telco	25	5,818	0.31	616	27,162	1.44
Retail	22	1,067	0.10	354	12,022	1.10
Agriculture	21	1,527	0.05	51	1,804	0.06
Animation	25	3,018	0.05	141	4,394	0.08
Musical Instruments	19	1,168	0.12	20	600	0.06

of a single hidden layer with a ReLU followed by a regression layer. We reuse the backbone's hyperparameters unchanged, the only difference being the fixed $\tau = 0.5$ (Section III-C). The *base model* for each corpus is trained on the source data with the encoder and LLM frozen and only the projector trained, for 5 epochs with learning rate 10^{-4}, $1{,}000$ warm-up steps, and batch size 10. The text-only denoising *adaptation* fine-tunes the LLM with LoRA on the self-attention query and value projections (rank 16, α=32), again for 5 epochs with learning rate 10^{-4} and $1{,}000$ warm-up steps, using the four-view batches of Section III-C. As a fully supervised topline we also fine-tune the same LoRA configuration directly on target audio–text pairs (*audio-adapted*); since this uses target audio (the resource our setting aims to replace) and our comparison is among text-only conditions, we report it only on the development domain (Fig. 3), not in the per-domain tables.

Generation LLM. Document generation, topic labeling, entity extraction, and document-to-dialog generation all use Gemma 4:31B via the SDialog toolkit [40].

Datasets. We use two corpora (Table I), reporting per domain (*Source* or *Target*) the total number of dialogs (#Dlg), utterances (#Utt; the speech segments the ASR is trained and scored on), and hours of audio (Hrs). *DefinedAI* is a proprietary corpus of manually transcribed customer–agent telephone calls licensed from the data-provider company Defined.ai[2]. The base model is trained on a *source* mixture of Banking (B), Insurance (I), and Healthcare (H); the *target* domains are these same three (from a disjoint subset, so source and target never overlap) plus the unseen domains Telco and Retail. *SlideSpeech* [16] is an open, multi-domain corpus of online conference talks and lectures, i.e., presentation-style rather than conversational speech; following [13] we use Life/Talent/English as the source mixture and Agriculture (Ag), Animation (An), and Musical Instruments (MI) as target domains. The two corpora thus differ acoustically and stylistically (telephone customer-service calls vs. presentation-style talks), so the cross-domain setting (Section VI) compounds lexical and recording-condition shifts.

Adaptation conditions (text source). Table II reports what the two generation pipelines produced as a corpus for domain

[2]https://defined.ai/

adaptation: the number of documents (#Document) and their sentences (#Sent), with #Dlg and #Utt as in Table I; r_{w} and r_{c} give #Sent and #Utt as fractions of the real training #Utt from Table I. The generated documents are compact (14–25 per domain): the #Sent pool is well below the real #Utt ($r_{\mathrm{w}} \leq 0.31$), while the #Utt pool is comparable to real on DefinedAI ($r_{\mathrm{c}} \approx 1.1$–$1.6$) but far smaller on SlideSpeech ($r_{\mathrm{c}} \approx 0.06$). We compare four experimental conditions, all using the *same* denoising backbone and $\tau = 0.5$: (i) *Base*, the unadapted source model; (ii) *Transcripts*, adaptation on ground-truth target transcriptions (#Utt, Table I), the upper bound for text-only adaptation; (iii) *Written*, the documents split into sentences and used directly as the target text (#Sent, Table II); (iv) *Conversational*, the same documents rendered as synthetic dialogs, using their utterances as the target text (#Utt, Table II). Conditions (iii) and (iv) are the *document-only* settings (no transcripts); (iii) is included to test the natural alternative of fine-tuning on the document text itself.

Metric. We report word error rate (WER, %) and, where useful, the relative improvement Δ over the corresponding base model.

Reproducibility. The generation, clustering, labeling, extraction, and document-creation pipelines are implemented on top of the SDialog toolkit [40] with fixed random seeds. We release all code (generation, benchmark construction, and experiments) together with the generated synthetic dialogs as well as the SlideSpeech documents for all its domains.[3]

VI. Results

All result tables use the four conditions from the previous section: *Base*, *Transcripts*, and the *Document-only* pair *Written* and *Conversational*; Δ is the relative WER improvement over *Base*; and **bold** marks the better document-only rendering, including near-ties within 0.1 WER. We organize results by corpus: adaptation within DefinedAI, whose targets include domains seen in the source (Banking, Insurance, Healthcare) as well as unseen ones (Telco, Retail, marked *ood*), and within SlideSpeech (all targets unseen), followed by the harder cross-domain transfer between the two corpora.

A. DefinedAI

Table III reports WER on the DefinedAI targets, where the base model is trained on the DefinedAI source. Two findings stand out. First, the conversational condition recovers most of the gain of real transcriptions: across the five domains the conversational–transcript WER gap is at most 0.28 absolute (Banking), and on Retail it even surpasses the transcript condition (17.18 vs. 17.59). Second, the written condition is consistently far weaker: near or below the base model on Banking and Insurance, and *worse* than the base model on Healthcare (7.63 vs. 6.85). The documents carry the relevant entities and keywords by construction (Sec. IV), but their written form lacks the spoken characteristics the ASR must model; the conversational view of the same content is far more effective here.

[3]Link is withheld for anonymous review.

TABLE III
WER (%) ON DEFINEDAI; Δ: RELATIVE IMPROVEMENT OVER BASE. TELCO AND RETAIL ARE UNSEEN DOMAINS (OOD).

Domain	Base	Transcripts		Document-only			
				Written		Conversational	
		WER	Δ	WER	Δ	WER	Δ
Banking	10.90	9.13	16.2	10.93	−0.3	**9.41**	13.7
Insurance	10.15	8.84	12.9	9.87	2.8	**8.89**	12.4
Healthcare	6.85	6.13	10.5	7.63	−11.4	**6.29**	8.2
Telco (ood)	20.12	17.61	12.5	19.19	4.6	**17.86**	11.2
Retail (ood)	20.48	17.59	14.1	18.79	8.3	**17.18**	16.1

TABLE IV
WER (%) ON SLIDESPEECH; ALL TARGETS ARE UNSEEN (OOD).

Domain	Base	Transcripts		Document-only			
				Written		Conversational	
		WER	Δ	WER	Δ	WER	Δ
Agriculture	16.25	16.01	1.5	**16.09**	1.0	16.27	−0.1
Animation	16.45	15.55	5.5	**16.22**	1.4	**16.13**	1.9
Musical Instr.	15.16	14.09	7.1	14.84	2.1	**14.15**	6.7

B. SlideSpeech

Table IV reports WER on the SlideSpeech targets (all unseen), with the base trained on the SlideSpeech source. Here even real transcripts gain little ($\Delta \leq 7.1\%$), so the headroom for any text-only method is small; within it, two effects stand out. First, unlike on DefinedAI, the written condition is competitive: SlideSpeech consists of presentation-style talks and lectures [16] rather than two-party conversations, so the documents' expository sentences already match the presentation register, whereas the customer-service dialog form produced by our DefinedAI-tuned prompt is a weaker fit. Second, the default conversational setting is data-limited: SlideSpeech has few dialogs, so generating N equal to the real dialog count yields only 51–141 dialogs and a few thousand utterances (Table II), far below the tens of thousands in the real data. Because generation is decoupled from the document count, we can simply generate more: raising Agriculture and Animation to 1,000 dialogs (40k and 31k utterances, matching the real data's order of magnitude) lowers WER from 16.27/16.13 to 16.17/15.68, approaching real transcriptions (16.01/15.55), with the larger gain on Animation. When headroom is small, then, the *amount* of synthetic data, not only its form, governs how much is recovered, and the conversational view offers a tunable knob to supply it.

C. Cross-domain (DefinedAI → SlideSpeech)

The hardest setting adapts a DefinedAI-source model to SlideSpeech targets, combining a domain shift with an acoustic/recording-condition shift, so the base WER is very high (47–59). Table V shows all three text sources roughly halve the WER and land close together: over the base, the conversational condition improves by 44.9/43.2/39.9% and the written condition by 44.1/38.0/45.3%, versus 49.5/39.6/45.2%

TABLE V
CROSS-DOMAIN WER (%): BASE TRAINED ON DEFINEDAI SOURCE, ADAPTED TO SLIDESPEECH TARGETS.

Domain	Base	Transcripts		Document-only			
				Written		Conversational	
		WER	Δ	WER	Δ	WER	Δ
Agriculture	59.40	29.98	49.5	33.22	44.1	**32.70**	44.9
Animation	48.18	29.12	39.6	29.86	38.0	**27.36**	43.2
Musical Instr.	47.02	25.76	45.2	**25.73**	45.3	28.25	39.9

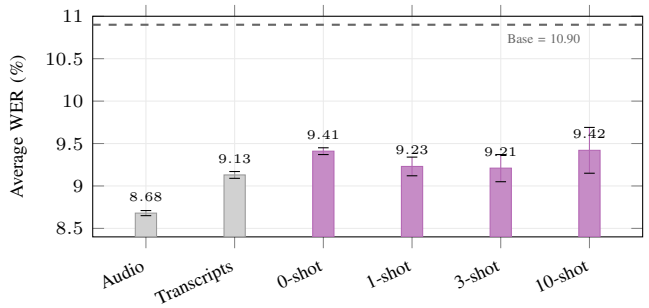

Fig. 3. Banking (development domain) average WER by adaptation text source; all conditions are averaged over three seeds (error bars). Light gray bars (*Audio*, audio-adapted; *Transcripts*, real transcriptions) are reference toplines our method avoids. Colored bars are our *Conversational* document-only condition with 0/1/3/10 real example dialogs in the prompt (0-shot, the transcript-free setting used throughout). The dashed line marks the unadapted base model.

for real transcriptions; the conversational condition surpasses real on Animation (27.36 vs. 29.12), while written and real perform virtually the same on Musical Instruments (25.73 vs. 25.76). Strikingly, this holds with the *default* small dialog count: because the source-trained base is far from the targets, even a little in-domain text recovers most of a large gain, in contrast to the data-limited in-corpus setting above. This is our strongest evidence: starting from a model trained on an unrelated corpus and using *only* target documents, one recovers most of the improvement that would otherwise require real in-domain transcripts, exactly the onboarding scenario our work targets.

D. Effect of few-shot prompting

We also test whether seeding the document-to-dialog prompt with a few real example dialogs (n-shot) helps, on Banking (Fig. 3). All conditions use three seeds; few-shot adds a second source of randomness, since each seed samples different example dialogs. Few-shot changes WER only marginally (1/3-shot 9.23/ 9.21 vs. zero-shot 9.41; 10-shot 9.42), with all conditions within 0.3 WER of the *Transcripts* topline (9.13) and the best (3-shot) within 0.1; 10-shot adds no gain and is the most variable, plausibly because its longer prompt (ten example dialogs) is harder to use reliably. Since *any* few-shot variant requires real in-domain transcripts (violating our document-only premise) and the gap is small, we adopt zero-shot generation throughout.

E. Prompt design

We stress that our goal is *feasibility*, not prompt optimality. We designed the generation prompt of Section III-B once, on Banking (our development domain), guided by lightweight checks that the synthetic dialogs matched real ones (turn-length, n-gram perplexity, and authorship style [41]) together with downstream Banking WER, then *froze* it across all domains and corpora. Since it was never tuned per domain and is likely suboptimal, the reported results are a conservative lower bound; refining the prompt is an orthogonal direction left to future work.

F. Discussion

Across all three settings (Tables III–V), document-only adaptation recovers much of the improvement real transcripts give, but which rendering wins, and how much text it needs, depends on the target through three factors. **Content** is fixed by construction: every condition carries the same entities and keywords. **Form** is decisive for conversational speech: on DefinedAI the written sentences barely help and on Healthcare even hurt, whereas rendering the same content as dialogs recovers nearly all the transcript gain. On the presentation-style SlideSpeech targets form matters far less, since the documents' expository sentences already fit that register and the written condition stays competitive (Table IV). **Quantity** takes over when the form fits but the pool is small: it is roughly real-sized on DefinedAI ($r_c \approx 1$) but tiny on SlideSpeech (≈ 0.06, Table II), where scaling generation to 1,000 dialogs improves the two domains with room to spare, most clearly Animation ($16.13 \rightarrow 15.68$, near its 15.55 transcript WER, Table IV). When the base instead starts far from the target (cross-domain), the headroom is so large that even the small default recovers most of it, with all text sources yielding similar gain over base (Table V). In short, the conversational view helps in two ways at once, turning compact documentation into spoken-style text and supplying as much of it as needed. These results hold with a single, fixed, non-tuned prompt (Section VI-E), suggesting that document-only adaptation is robust rather than an artifact of prompt engineering.

VII. CONCLUSION

We introduced document-only adaptation of LLM-based ASR: adapting to a new domain from raw documents alone, with no target audio and no target transcripts. We use a domain's documents as adaptation text in two ways: directly, as their own written sentences, or rendered into synthetic spoken dialogs. Fed to a denoising-based adaptation method, the conversational rendering generally approaches, and sometimes matches or exceeds, ground-truth-transcript adaptation across the three settings on two corpora; the written rendering is competitive when its register fits the target or the model starts far from it, but can fall short, or even hurt, otherwise. This makes adaptation feasible from documentation a customer already owns, without recording or transcribing sensitive conversations. We release the generated SlideSpeech documents and dialogs to support future research.

ACKNOWLEDGMENT

This work was supported by Idiap Research Institute and Uniphore collaboration project. Part of this work was also supported by EU Horizon 2020 project ELOQUENCE (grant number 101070558). During the preparation of this manuscript, the authors used Claude Opus 4.8 (Anthropic) to assist with editing, rephrasing, and polishing the writing throughout the paper. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of this article.

REFERENCES

[1] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang, and X. Chen, "Speech recognition meets large language model: Benchmarking, models, and exploration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i23.34666

[2] S. Kumar, I. Thorbecke, S. Burdisso, E. Villatoro-Tello, M. KE, K. Hacioglu, P. Rangappa, P. Motlicek, A. Ganapathiraju, and A. Stolcke, "Performance evaluation of slam-asr: The good, the bad, the ugly, and the way forward," in *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2025, pp. 1–5.

[3] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for asr," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 637–12 641.

[4] M. Yang, S.-J. Chen, J. Xie, and H. L. John Hansen, "Bridging the modality gap: Softly discretizing audio representation for llm-based automatic speech recognition," in *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025, pp. 1–7.

[5] S. Burdisso, E. Villatoro-Tello, S. Kumar, S. Madikeri, A. Carofilis, P. Rangappa, M. K. E, K. Hacioglu, P. Motlicek, and A. Stolcke, "Reducing prompt sensitivity in LLM-Based Speech Recognition through learnable projection," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16 372–16 376.

[6] E. Villatoro-Tello, S. Burdisso, S. Kumar, H. Watawana, S. Madikeri, M. K E, J. Prakash, T. Bañeras-Roux, K. Hacioglu, P. Motlicek, and A. Stolcke, "Context projector: Complementary keyword and dialogue context embeddings for LLM-based ASR," in *Proc. Interspeech*, 2026.

[7] P. D'Alterio, C. Hensel, and B. A. S. Hasan, "Can unpaired textual data replace synthetic speech in asr model adaptation?" in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.

[8] H. Sato, T. Komori, T. Mishima, Y. Kawai, T. Mochizuki, S. Sato, and T. Ogawa, "Text-only domain adaptation based on intermediate CTC," in *Interspeech*, 2022, pp. 2208–2212.

[9] V. Bataev, R. Korostik, E. Shabalin, V. Lavrukhin, and B. Ginsburg, "Text-only domain adaptation for end-to-end asr using integrated text-to-mel-spectrogram generator," in *Proc. Interspeech 2023*, 2023.

[10] B. Kurian, A. Upadhyay, and A. Sengupta, "Domain-specific adaptation for asr through text-only fine-tuning," in *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, 2025, pp. 78–85.

[11] Y. Fang, J. Peng, X. Li, Y. Xi, C. Zhang, G. Zhong, and K. Yu, "Low-resource domain adaptation for speech llms via text-only fine-tuning," *arXiv preprint arXiv:2506.05671*, 2025.

[12] Y. Ma, Z. Liu, and O. Kalinli, "Effective text adaptation for llm-based asr through soft prompt fine-tuning," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 64–69.

[13] A. Carofilis, S. Burdisso, E. Villatoro-Tello, S. Kumar, K. Hacioglu, S. Madikeri, P. Rangappa, M. K. E, P. Motlicek, S. Venkatesan, and A. Stolcke, "Text-only adaptation in LLM-based ASR through text denoising," 2026. [Online]. Available: https://arxiv.org/abs/2601.20900

[14] T. Bañeras-Roux, S. Burdisso, E. Villatoro-Tello, D. Sánchez-Cortés, S. Liu, S. Baroudi, S. Kumar, H. Watawana, M. K. E, K. Hacioglu, P. Motlicek, and A. Stolcke, "Closing the speech-text gap with limited audio for effective domain adaptation in LLM-Based ASR," 2026. [Online]. Available: https://arxiv.org/abs/2604.06487

[15] S. Burdisso, E. Villatoro-Tello, T. Bañeras-Roux, S. Kumar, S. Madikeri, P. Rangappa, M. K E, P. Motlicek, and A. Stolcke, "Avoiding catastrophic forgetting in text-only adaptation of LLM-based ASR via multi-view text denoising," in *Proc. Interspeech*, 2026.

[16] H. Wang, F. Yu, X. Shi, Y. Wang, S. Zhang, and M. Li, "SlideSpeech: A large scale slide-enriched audio-visual corpus," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 076–11 080.

[17] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," 2017. [Online]. Available: https://arxiv.org/abs/1708.06426

[18] Z. Meng, Y. Wu, N. Kanda, L. Lu, X. Chen, G. Ye, E. Sun, J. Li, and Y. Gong, "Minimum word error rate training with language model fusion for end-to-end speech recognition," *arXiv preprint arXiv:2106.02302*, 2021.

[19] T. Hori, M. Kocour, A. Haider, E. McDermott, and X. Zhuang, "Delayed fusion: Integrating large language models into first-pass decoding in end-to-end speech recognition," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[20] J. Gung, E. Moeng, W. Rose, A. Gupta, Y. Zhang, and S. Mansour, "Natcs: Eliciting natural customer support dialogues," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[21] A. Echeverria, S. S. T. de Oliveira, and F. M. Federson, "Call2instruct: Automated pipeline for generating q&a datasets from call center recordings for llm fine-tuning," *arXiv preprint arXiv:2601.14263*, 2025.

[22] S. Mohammadi, M. Paldhe, A. Chhabra, Y. Son, and V. Seshagiri, "Lingvarbench: Benchmarking llms on entity recognitions and linguistic verbalization patterns in phone-call transcripts," in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, 2026, pp. 545–561.

[23] S. K. Suresh, W. Mengjun, T. Pranav, and E. S. Chng, "Diasynth: Synthetic dialogue generation framework for low resource dialogue applications," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 673–690.

[24] J. Gao, L. Xiang, H. Wu, H. Zhao, Y. Tong, and Z. He, "An adaptive prompt generation framework for task-oriented dialogue system," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1078–1089.

[25] Y.-S. Lee, C. Gunasekara, D. Contractor, R. F. Astudillo, and R. Florian, "Multi-document grounded multi-turn synthetic dialog generation," *arXiv preprint arXiv:2409.11500*, 2024.

[26] A. Ahmad, S. Hillmann, and S. Möller, "Simulating user diversity in task-oriented dialogue systems using large language models," *arXiv preprint arXiv:2502.12813*, 2025.

[27] S. Dou, C. Sun, X. Wang, X. Liu, K. Shao, C. Chen, B. Zhang, and H. Hu, "Finua: Generating diverse user interactions for financial dialogue systems through user simulation," in *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2026, pp. 16 292–16 296.

[28] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[29] A. J. Oche, A. G. Folashade, T. Ghosal, and A. Biswas, "A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions," *arXiv preprint arXiv:2507.18910*, 2025.

[30] J. Li, R. Gadde, B. Ginsburg, and V. Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," *arXiv preprint arXiv:1811.00707*, 2018.

[31] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE, 2019, pp. 996–1002.

[32] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7069–7073.

[33] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You do not need more data: Improving end-to-end

speech recognition by text-to-speech data augmentation," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 439–444.

[34] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[35] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. ACM SIGKDD*, 2019, pp. 2623–2631.

[36] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[37] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[38] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[39] A. Dubey *et al.*, "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[40] S. Burdisso, S. Baroudi, Y. Labrak, D. Grünert, P. Cyrta, Y. Chen, S. Madikeri, E. Villatoro-tello, R. Marxer, and P. Motlicek, "SDialog: A python toolkit for end-to-end agent building, user simulation, dialog generation, and evaluation," in *Proceedings of EACL*. Rabat, Marocco: ACL, Mar. 2026, pp. 320–340. [Online]. Available: https://aclanthology.org/2026.eacl-demo.23/

[41] R. A. Rivera-Soto, O. E. Miano, J. Ordonez, B. Y. Chen, A. Khan, M. Bishop, and N. Andrews, "Learning universal authorship representations," in *Proc. EMNLP*, 2021, pp. 913–919.