Presentations *by the Humans* and *For the Humans*: Harnessing LLMs for Generating Persona-Aware Slides from Documents

Anonymous ACL submission

Abstract

Scientific papers and slides are two different representations of the same underlying information, but both require substantial work to prepare. While there had been prior efforts on automating document-to-slides generation (Fu et al., 2021; Sun et al., 2021), the concept of tailoring presentations to suit specific target audience or fit in a given time duration has been underexplored. This paper introduces *end-user specification-aware document-to-slides generation* that reflects end-user specifications into conversion process. First, we introduce a new dataset of papers and corresponding slide decks from recent *ACL conferences with four persona-aware configurations. Second, we present **Persona-Aware-D2S**, a novel approach by fine-tuning LLMs using target audience feedback to create persona-aware slides from scientific papers. Our evaluation using automated metrics and human surveys suggests that incorporating end-user specifications into conversion creates presentations that are not only informative but also tailored to cognitive abilities of target audience.

1 Introduction: Presentations are Everywhere…How can we make them customized to end user needs?

From business to education to research, presentations are everywhere (Zheng et al., 2022; Bhattacharyya, 2014; Tarkhova et al., 2020). A recent 2023 survey[1] reveals that 20.3 million people in the UK have used Powerpoint and over half (53%) of people in the UK have been required to create presentations either at work or in their personal lives, yet the creation of slide decks from documents poses significant cognitive load on users. This problem can be looked upon as a specific challenge within the broader context of summarizing

[1] https://www.acuitytraining.co.uk/news-tips/powerpoint-statistics/

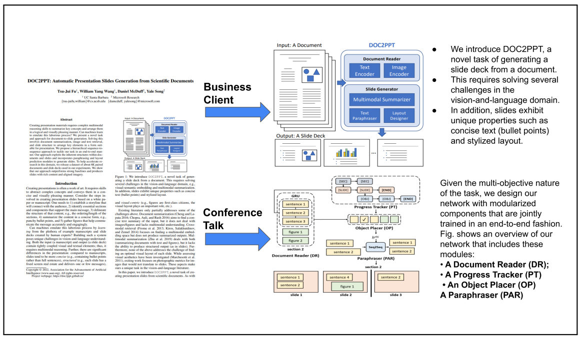


Figure 1: Output from our proposed **Persona-Aware-D2S** model showing the type of content preferred by end-users of two different persona while demonstrating the main pipeline of a conference paper.

long documents (Koh et al., 2022). Moreover, during conversion of a knowledge-rich scientific paper for a specific audience, it's crucial to consider pragmatic factors like audience expertise on the subject, duration of presentation, preferred communication style of audience, etc. Think of a scenario where you need to quickly create brief, audience-tailored presentations in just an hour for ACL conference attendees and a paper overview for business users, balancing complexity with time constraints. For instance (Figure 1), in a meeting with general public/businessmen, a lot of technical content might decrease engagement, as they might be only interested in knowing overall use-case instead of a detailed model architecture.

Existing work on automating document to slides creation (Fu et al., 2021; Sun et al., 2021) provides a strong foundation, but it lacks both mechanisms for users to customize the creation of slides and datasets that reflect that a single source document can be presented in multiple ways. In addition, these works are mostly aligned with fine-tuning based on a single gold standard (such as maximizing likelihood of ROUGE-measures) and are not aligned with expectations of humans having diverse expertise (Fu et al., 2021; Sun et al., 2021).

	N-S	N-L	E-S	E-L
#Slides	75	75	75	75
#Tokens	299.68	367.88	297.07	431.53
#Unique Tokens	37.29	40.11	38.91	45.23
#Sentences	13.85	24.89	18.2	32.74

Table 1: Statistics of **Persona-Aware-D2S-Dataset** where E, N, L, S stand for experts, non-experts, Long and Short persona-aware configurations respectively.

To address this gap, we make the following contributions: **[1]** To the best of our knowledge, we introduce a novel task of Human-In-the-Loop (HITL) persona-aware transformation of scientific documents to slides. **[2]** We introduce a new parallel corpus of document and persona-aware slides by repurposing *ACL papers from existing *SciDuet* dataset to create persona-aware presentations (section 2) to accomodate time constraints and end-user's technical background. **[3]** We are the first to propose a simple method that harnesses the power of LLMs to design end-user specification-aware presentations simply using natural language instructions (prompts) and **[4]** we propose **Persona-Aware D2S**, a novel pipeline for creating persona-aware presentations which comprises of *generating persona-specific slide outlines*, followed by a *persona-aware content extractor* to fetch relevant snippets from documents for each outline and *summarizing and aligning snippets on slides* (Section 3) and perform evaluation using both automatic metrics and human judgement (Section 5, 6).

2 Persona-Aware-D2S-Dataset Creation

Prior research has predominantly addressed preparing technical conference slides (Section 7), neglecting diverse presentation types, audiences, and durations. To fill this gap, we curate a novel benchmark evaluation dataset that encompasses a wider spectrum of presentation needs. Our dataset focuses only on a subset of 75 papers from *SciDuet* (Sun et al., 2021) dataset to create persona-aware configuration slides of each paper.

Data Annotation: We hope that our dataset will serve as a benchmark to train and evaluate persona-aware slide generation models, thus we conduct human annotation of our chosen subset of papers (75 papers) as mentioned in 2. Using Upwork, we hired two workers familiar with Machine learning and NLP (5 years of experience) and well-versed with creating presentations from documents (skill set: Presentation making) to create a parallel dataset containing paper and four persona-aware presentations: 1) **Expert-Long (E-L)** tailored for conference attendees and detailed presentation, 2) **Expert-Short (E-S)** tailored for conference attendees in a quick and spotlight fashion, 3) **Non-Expert-Long (N-L)** tailored for business attendees and detailed presentation, 4) **Non-Expert-Short (N-S)** tailored for business attendees in a quick and spotlight fashion). At the time of hiring, we showed them a paper, asked them to go through it, and answer 5 technical, conceptual and basic questions regarding that paper. We made a hiring decision if they could provide satisfactory answers and also made reasonably good presentations (See C.1).

After hiring, we ran a pilot phase to ensure that could create persona-aware presentations for each paper, when the task is to create four configuration of persona-aware presentations from two papers (as mentioned previously). Specific instructions were provided on choosing sentences/figures/tables from only the paper and no content should be included from external sources.

To ensure quality, the first two authors carefully checked the details of created presentations and started final round of annotation. After that, we randomly chose 200 documents (other than papers used during training) from the *SciDuet* dataset, and asked them to create four configuration of presentation slide decks for each of the chosen 200 documents. We exchange the presentations created between the two annotators amongst them and asked to rate the quality of presentations on a Likert scale of 1-5 and retained 75 PDFs and corresponding 4 slides per PDF where Likert scale rating $\geqslant$ 3.5.

Dataset Statistics and Analysis: Our dataset is split into train (20), dev (5) and test (50) set (number of papers in bracket). Each paper has four configuration of slides (total 75 papers and 300 slides). 56.3% slide outlines annotated are generic (e.g., method, results). Each slide comprises of content from more than one section of the paper, and on average each slide contain sentences selected from 2.5 sections. For short and long presentations, average number of slides are 4.56 and 7.6 and average number of tokens are 125.2 and 580.6 respectively (Table 1). 87.34% of slide outlines have fewer than 4 tokens, the top-3 frequent unigrams are Introduction, Motivation, Solution and top-3 bigrams include Problem Statement, Related Work, Solution Approach.

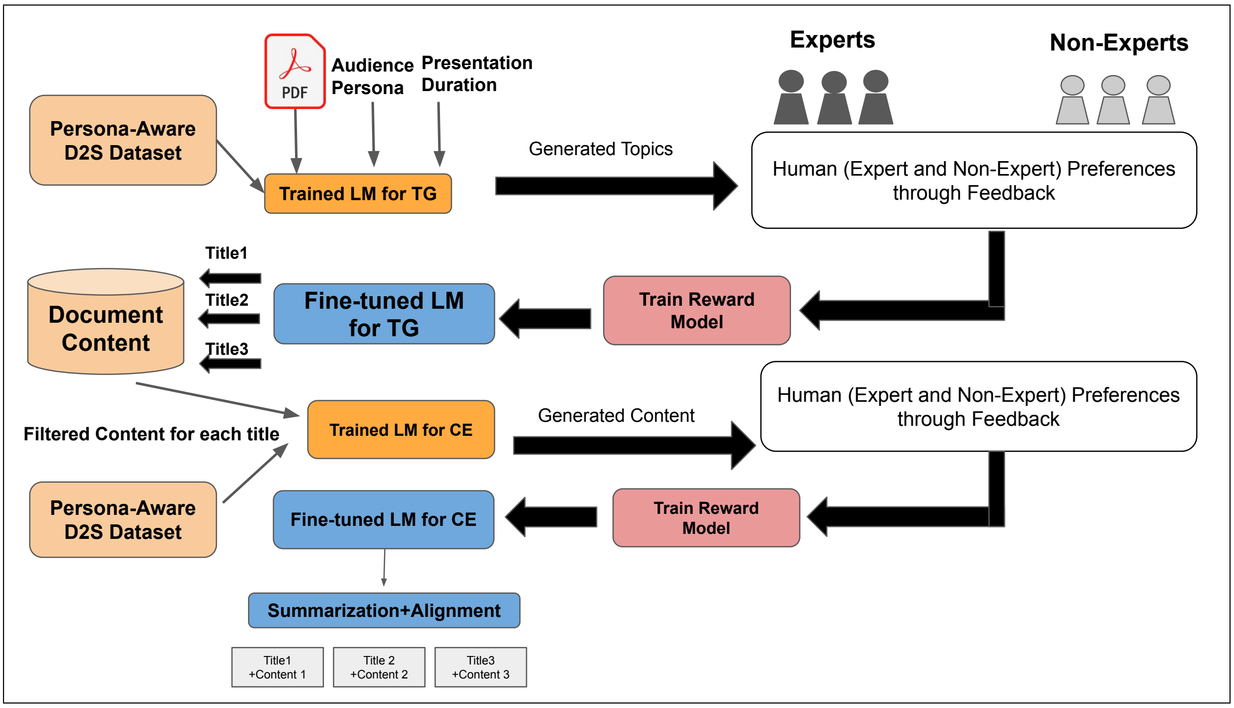


Figure 2: shows the entire information flow of Persona-Aware D2S - Model Pipeline. Initially, LLM for Topic Generator is trained with supervision from Persona-Aware D2S dataset, followed by finetuning using human-feedback to produce Fine-tuned LM for Topic Generator. For each generated slide outline, we filter content from document to extract relevant snippet for the title, the final content generator LLM is fine-tuned with Human Feedback. The content for all slide outlines are summarized and aligned to produce a logically coherent slide deck.

3 Persona-Aware D2S - Model Pipeline

Notations: A document D is organized into sections SE and a set of multimodal content figures/tables F. Each figure $F_q = \{I_q, Cap_q\}$ contains an image I_q and a caption Cap_q. Document content, the heading and abstract of paper are represented as C, H and A respectively.

Input and Output: Our model pipeline takes the document content C, audience background B ($B \in \{e, ne\}$ where e and ne stands for experts and non-experts respectively) and duration of presentation L ($L \in \{l,s\}$ where l and s stand for long and short presentations) as input and generates the final slide deck O, without including any external content. We denote input tuples $IN = \{C, B, L\}$ and output slide deck as O, where the probability of generating slide deck $p(O|C, B, L)$ has to be maximized. Our model pipeline is decomposed into following steps:

3.1 Persona-aware Slide Outline Generation

The first step is to have a mental model of how the slide outlines of the transformed document should look like, which comprises of choosing outline and the order in which the outline should be presented. Given A, H corresponding to a document, we generate slide outlines $t = \{t_1, t_2, ... \ t_j\}$ for each of the 4 possible combinations of persona-aware contraints B and L that strictly follow the order in which the slides in the slide deck O should be generated. Thus, we model the problem of persona-aware topic generation as conditional probability : $P(t|IN)$. Since B and L are binary variables, their combined set contains 4 possible combinations and for each combination, we generate topics for a fixed value of A, H.

3.1.1 Supervised Fine-tuning (SFT-F)

We fine-tune LLM using prompt created using persona-aware inputs (IN), and responses (slide outlines t) from the train split of **Persona-Aware-D2S-Dataset** in a supervised policy π_{SFT}. It adjusts weights in LLM by minimizing cross-entropy loss between generated topics (T') and ground-truth topics (T). We finetune such that for each configuration, we generate supervised policies $\pi_{SFT(B=ne,L=l)}$, $\pi_{SFT(B=ne,L=s)}$, $\pi_{SFT(B=e,L=l)}$ and $\pi_{SFT(B=e,L=s)}$.

3.1.2 Fine-tuning using Preference Data (P-F)

While LMs learn broad world knowledge, achieving precise control of their behavior is difficult due to unsupervised nature of their training. So it is imperative to gain steerability by collecting human labels of the relative quality of generations and further fine-tune the unsupervised LM to align with these preferences (reinforcement learning from human feedback (RLHF) (Christiano et al., 2017)).

Reward Modelling Inspired by the above motivation, we fine-tune our supervised policies to generate data that humans prefer on certain criteria, thus we need to model rewards for each criteria. On dev set, we generate set of topics using supervised policies $\pi_{SFT(B=ne,L=l)}$, $\pi_{SFT(B=ne,L=s)}$, $\pi_{SFT(B=e,L=l)}$ and $\pi_{SFT(B=e,L=s)}$ for each configuration. Using each policy, we vary temperature, top-K sampling and top-p nucleus sampling to generate 5 topic set for each persona-aware input (IN). Then we ask three experts to pairwise rank the topic set generated by $\pi_{SFT(B=e,L=l)}$ and $\pi_{SFT(B=e,L=s)}$ on two criteria **comprehensibility to target audience** and **length-based satisfaction**) and similarly three non-experts (see C.2) to pairwise rank the topics generated by $\pi_{SFT(B=ne,L=l)}$ and $\pi_{SFT(B=ne,L=s)}$ [2]. , we consider only those responses where there is a majority voting or consensus (E.g., for input prompt A, $r1$ is chosen over $r2$ by two experts on **comprehensibility to target audience** criteria, and $r2$ is chosen over $r1$ by another expert, we finally consider $r1$ over $r2$ on this criteria for prompt A), and discard those samples from the human-preference comparison data where there is no such consensus. Using this collected data, we train a reward model to generate reward (for each criteria) for a (prompt A, topic set t) pair by maximizing difference between the reward for the chosen response (s_w) and that of the rejected response (s_r), the goal is to minimize the expected loss for all training samples ($train$):

$$loss = -\mathbb{E}_{x \in train} \log_{\sigma} (s_w - s_r)) \qquad (1)$$

Now, we have 4 trained reward models: **RM-Comprehensibilty (RM-C-E)**, **RM-Length (RM-L-E)** for experts and **RM-C-NE** and **RM-L-NE** for non-experts.

Final Preference Fine-tune with estimated rewards and Inference Finally, we sample prompts (IN) from train set and generate 5 topic-sets by varying temperature using the π_{SFT} for each configuration. For each (sample, topic-set) pair, we use the RM-Comprehensibilty and RM-Length to generate rewards and further fine-tune LLM with the (prompt,reward) as input and topic-set as output, drawing on the principle of *Decision Transformer* (Chen et al., 2021) that abstracts Reinforcement Learning (RL) as a sequence modeling problem. During inference on test set, we provide the maximum reward for each criteria as input to each prompt, and obtain the sequence of topics that is optimal for that reward.

[2]these annotators are different from the ones asked to evaluate slides, just to mitigate any potential bias during evaluation

3.2 Persona-aware Content Extraction

Given the slide outlines t generated by persona aware slide outline generation module, this step selects a set of relevant sentences T_i and figure/table captions C_q for each title t_i from the document content C for the specified constraints B and L. We follow two steps to achieve this personalization goal. First, we make use of a retriever that fetches relevant content from source document (D) for each slide outline (t) 3.2. Since prompting an LLM to choose relevant sentences from entire paper with t as a query is an expensive operation, we use a non-LLM based sparse retriever (3.2) to ensure that the subset retrieved for each slide outline is small enough to make minimum number of LLM-calls and most of the gold- snippets for each title is included in the fetched content. So, we chunk C into a subset Su that serve as candidates for extracting persona-aware relevant content, and passed on to finally filter out information from Su. Therefore, we model the problem of persona-aware content extraction as conditional probability : $P(t|IN)$. Since B and L are binary variables, their combined set contains 4 possible combinations and for each combination, we generate content for a fixed value of A, H.

Topic-wise High Recall Section Filter First, we match each title in the slide t = $\{t_1, t_2 ... t_n\}$ to the most relevant section titles of the paper, which can serve as potential candidates for Su. Formally, given a candidate set of section headings SH, a query t_i we retrieve the top-k section headings using fuzzy match with a similarity score greater than th. Our choice of threshold (th) is determined after tuning on the development split. If none of the sections in the paper satisfy the above condition, we use sentence transformers (Reimers and Gurevych, 2019) to choose a section which has the highest similarity with the given slide outline. After choosing paper section titles for each t, we concatenate all the content (sentences and captions) belonging to the matched sections of the paper.

Persona-aware Content Extraction from Candidates Content Based on the output of retriever in step 3.2, we extract sentences tailored to the needs of end-user in this step. We follow the similar approach as persona-aware content extraction as performed in 3.1.1 where in **Step 1** we first fine-tune an LLM using slide outline t, persona-aware prompts with Su from candidate sentences per title, and responses (most relevant sentences $Su_{relevant}$) from the train split of **Persona-Aware-D2S-Dataset** in a supervised policy π_{SFT-CE}. It adjusts weights in LLM by minimizing cross-entropy loss between generated sentences and ground-truth sentences, then in **Step 2**, we follow the same principle (as mentioned in 3.1.2) of reward modelling and further finetuning LLM towards human preferences to choose the best set of sentences for each configuration per slide outline.

3.3 Summarization and Logical Alignment

The goal of this step is to convert extractive snippets from section 3.2 in a logically structured way such that the consumer of presentation can easily follow the content rendered from beginning to end. So, we summarize the content extracted for each slide outline t, then pass the summarized bullet points to an LLM asking for re-arranging the content inside a topic or across the topic to make it consumable by the audience (We use paper abstract and and concatenated summary of each slide content to generate slide decks, See Appendix).

4 Experimental Details

Our **Persona-Aware-D2S** pipeline is based on auto-regressive generative large language models (LLMs). We have experimented with GPT-2 (*text-davinci-002*), GPT-3 (*text-davinci-003*) and ChatGPT (*gpt-3.5-turbo*) as LLMs. In our pipeline, we have personalized both topic generation and content extraction steps and compared with non-personalized configurations.

Topic Generation Baselines We consider the following baselines for generating t from D (See E): 1) **Non-persona-aware Zero-shot Topic Generation (NZS-TG)**: Our prompt to the LLM comprises of only A and T of a document D, and we ask it to generate t. 2) **Persona-aware Zero-shot Topic Generation (ZS-TG)**: Apart from input to NZS-TG, we include B and L in the prompt and we ask it to generate t. 3) **Persona-aware Few-shot Topic Generation (FS-TG)**: Apart from input in ZS-TG, we provide $k1$ input-output samples from train-split of **Persona-Aware-D2S-Dataset**, along with $k1$ input-output samples and we ask it to generate t.

Model	Input	Evaluation Metrics		
		Precision	Recall	F1-score
NZS-CE	A+T	0.12 (0.08)	0.44 (0.11)	**0.18 (0.06)**
ZS-CE	A+T+B	0.30 (0.06)	0.47 (0.05)	0.38 (0.06)
	A+T+B+L	0.32 (0.03)	0.42 (0.01)	0.36 (0.04)
FS-CE	A+T+B	0.32 (0.06)	0.46 (0.05)	0.37 (0.06)
	A+T+B+L	0.34 (0.03)	0.47 (0.01)	0.40 (0.04)
SFT-F	A+T+B	0.41 (0.02)	0.70 (0.05)	0.51 (0.03)
	A+T+B+L	**0.45 (0.06)**	**0.72 (0.05)**	**0.54 (0.06)**
P-F	A+T+B	0.40 (0.02)	0.66 (0.03)	0.45 (0.01)
	A+T+B+L	**0.45 (0.04)**	**0.65 (0.05)**	**0.51 (0.05)**

Table 2: Benchmark Evaluation Results of content Extraction on test set. Rows for each model shows performance with different input features: Abstract (**A**), Title (**T**), Background of audience (**B**), and Length of presentation (**L**). The brackets indicate standard deviation after running on different prompt variations.

Content Extraction Baselines We consider the baselines for generating Su relevant to t from D (E): 1) **Non-persona-aware Zero-shot Content Extraction (NZS-CE)**: Our prompt to the LLM comprises of top-k content corresponding to t_i, and ask to select Su. 2) **Persona-aware Zero-shot Content Extraction (ZS-CE)**: comprises of top-k content element corresponding to t_i, B and L and ask to select Su. 3) **Personalized Few-shot Content Extraction (FS-CE)**: Apart from input in ZS-CE, we provide $k1$ input-output samples from train-split of **Persona-Aware-D2S-Dataset** and ask to select Su.

Hyperparameters and Model Details We fine-tuned GPT-3.5-turbo from OpenAI's standard API. The models are finetuned for 3 epochs, with learning rate 0.2, batch size 256. The zero-shot and few-shot experiments are carried out with temperature 0 to have a reproducible setup. We use distillbert-base[3] to calculate reward on comparison data collected during human feedback collection.

5 Evaluation: Automatic Measures

Our proposed candidate-filtering approach saves GPT-calls by 8 times Table 7 shows the

[3] https://huggingface.co/distilbert-base-cased

	Expert-Long			Expert-Short			Non-Expert-Long			Non-Expert-Short		
	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L	Rouge-1	Rouge-2	Rouge-L
Zero-shot	0.12	0.08	0.05	0.04	0.03	0.03	0.10	0.08	0.06	0.06	0.05	0.04
Few-shot	0.10	0.09	0.07	0.08	0.06	0.05	0.12	0.10	0.06	0.07	0.06	0.06
SFT-F	0.26	0.23	0.21	**0.17**	**0.15**	**0.15**	0.18	0.16	0.14	0.19	0.16	0.15
P-F	**0.28**	**0.24**	**0.22**	0.13	0.14	0.13	**0.19**	**0.17**	**0.17**	**0.19**	**0.18**	**0.16**

Table 3: Final Evaluation of Slides using the Persona-Aware-D2S pipeline (topic generation, content extraction, summmarization) for all four persona-aware configurations on Rouge-1, Rouge-2 and Rouge-L measures, showing that P-F models outperform others on all configuration except Expert-Short.

trade-off between using entire paper as candidates in 3.2 (higher number of GPT calls) vs the performance of recall in candidatate filtering. This step was mostly done to chunk the input prompt (for GPT3.5) to 4096 token limit, but we infer that making smaller number of GPT calls (1-5) might hurt performance of candidate retrieval.

Our proposed models outperform the baselines for module-wise and end-to-end evaluation. When we use chunked candidate set of relevant sentences and pass it to **CE** module, our maximum recall stands (token limit of the candidates is 2500) at 78.89%. Even after that, there is a significant improvement (12%) in average F1-scores after finetuning GPT3.5-turbo over baselines (Table 2). Moreover, Table 3 indicates that our P-F model outperforms all other baselines in terms of end-to-end performance evaluation of slide generation for all the configurations except Expert-Short where SFT-F is the winning candidate.

Generalizability of Approach with other LLMs Table 8 shows that almost any GPT-based LLMs can be leveraged with our approach. We conduct all experiments with GPT 3.5-turbo due to its decent decent performance with standard context window while being cheaper than GPT-3.

6 How *'good'* are the presentations according to the human raters?

Inspired by (Ribeiro et al., 2020), automatic evaluation metrics alone cannot accurately estimate the performance of a model. Thus, we assess whether the generated slides translate into lesser cognitive load of authors (Section 6.2) and better satisfaction in terms of personalization as judged by participants of diverse expertise (both quantitatively in 6.1 and qualitatively in 6.3), hired through Upwork (see C.2). The human evaluation task involves rating slide outputs by reading the corresponding papers from our dataset.

Figure 3: Average User Ratings by Experts on generated topics (Human-created and 3 model-created).

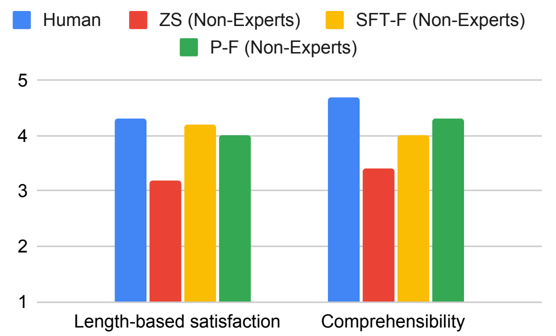


Figure 4: Average User Ratings by Non-Experts on generated topics (Human-created and 3 model-created).

6.1 Module-wise Evaluation and Findings

To assess effectiveness of every module in our model pipeline, we conduct an user study involving both technical experts and non-experts. We maintain consistent inputs at every intermediate step to ensure fair evaluation and employ non-personalized evaluation criteria like **Coverage**, **Relevance**, **Readability**, **Coherence** and persona-aware evaluation criteria like **Comprehensibility** and **Aptness of content volume with respect to length of Presentation** (Details in B).

6.1.1 Evaluation on Topic Generation

We randomly sample 10 papers from test set, generate 4 configurations of topic generation and show

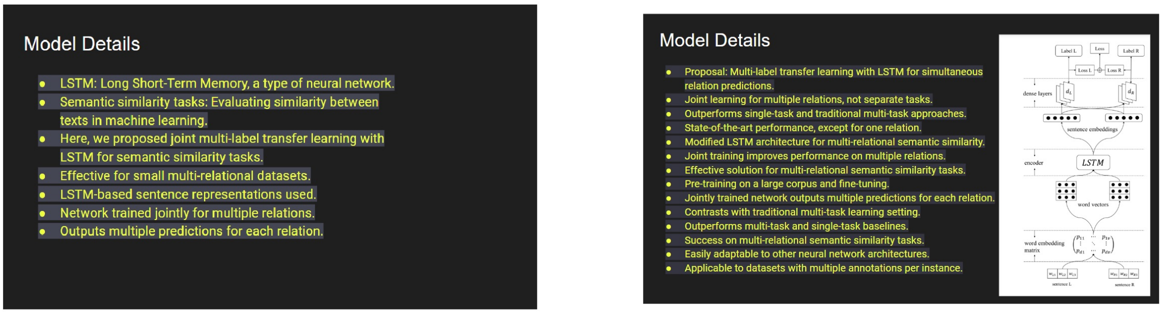


Figure 5: Source: (Zhang et al., 2019) (a) is produced by **P-F** model for non-experts on 'Model Details' with explanations of technical jargons and less details on network and training and (b) is generated by **P-F** model on 'Model Details' with content explaining the nitty gritty details of training and no explanations of technical jargons.

non-expert configuration to non-experts and vice-versa. For both groups, we also show topics customized for both long and short presentations: a) Human-written topics, b) ZS-TG output, c) SFT-F TG output and d) P-F TG output. These were rated by both groups on a 5-point Likert Scale along two persona-aware criteria. Ratings on same model's outputs are aggregated into average, resulting in 3 scores for each of 4 configurations.

Irrespective of presentation duration, technical experts gravitate towards comprehensible slide outlines while non-experts prefer concise titles. The most **comprehensible** and **length-based satisfactory** slide outlines were generated by humans (Figure 3). Experts have rated comprehensibility of slide outlines generated by our ZS and PR-model higher than the SFT-F model. Whereas, non-experts rated the comprehensibility of P-F higher than all other baselines, followed by SFT-F model (Figure 4). Even though the experts prefer more detailed, technical illustration-heavy topics that cater to their depth of knowledge, the non-experts prefer slide outlines that are less cluttered with technical jargons (table 6). On **Length-based satisfaction**, both the groups prefer SFT-F and PR-F outputs compared to that of ZS-F.

6.1.2 Evaluation on Content Extraction

As an evaluation set, we sample 20 random slides from the papers in the test set ensuring that the slide outlines are diverse *(E.g., Results, Methodology, Conclusion, Baseline Experiments, etc.)*. Next we generate 4 configurations of each slide (N-S, N-L, E-S and E-L). For each configuration, we choose the human-created slide from our dataset, our **Z-S**, **SFT-F** and **P-F** model generated slides and show the N-S and N-L configuration to non-experts and

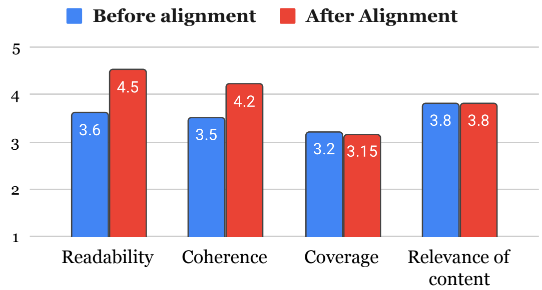


Figure 6: Average User Ratings (1-5) on 10 randomly sampled slide decks after Summarization+Alignment (Step-3) compared to extractive approach of slide generation (Step-2) indicating that *summarization and alignment is important for improved user experience.*

E-S and E-L to experts. Both groups rate the slides along the following dimensions (Coverage, Relevance, Length-based Satisfaction, Comprehensibility) on a 5-point Likert scale.

Experts rate our model-generated slides higher on all criteria compared to baselines, however on average non-experts' rate comprehensibility lower for all slides. (Figure 7) Experts prefer human-generated slides on all the criteria, except coverage of the paper (-0.8). ZS-TG provides the highest coverage but the least relevance, experts rate the SFT-F and P-F generated models equally high on coverage, length-based satisfaction and comprehensibility, indicating that experts prefer quality of our model (SFT-F and P-F) generated slides over baseline ZS-method. However, non-experts rate comprehensibility of all slides lower than their ratings on other criteria (Figure 8), on average their ratings displayed similar trends as followed by experts, thus we conduct a follow-up study (Section D).

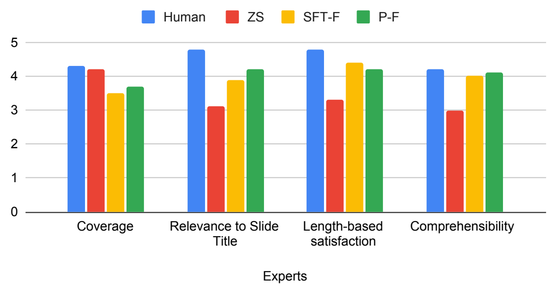


Figure 7: Average User Ratings by Experts on 4 slide configurations (Human-created and 3 model-created) where we found that experts rate our model-generated slides higher on all criteria compared to baselines, except coverage of paper.

6.1.3 Evaluation of Summarization and Alignment

During evaluation, we choose 10 papers and same set of experts and non-experts to evaluate how much does this step enhance *user's experience* on **Readability**, **Coherence**, **Coverage** and **Relevance of Content**. Figure 6 shows improvement on coherence (+0.5) and readability (+1), with minimal impact on coverage (-0.05) and relevance (0).

6.2 Reducing cognitive load of authors while making personalized presentations

We analyzed whether our model can reduce authors' cognitive load in creating persona-aware presentations. We generated N-S and N-L configurations using both baseline (**ZS**) and our model (**P-F**) for two random papers in test set and presented to 3 NLP experts asking how much time they would need to finalize presentations for non-experts (short and long) when starting with N-S and N-L configurations respectively from our proposed model, baseline model and compared to starting from scratch. Table 9 indicates a majority consensus between authors that making presentations from scratch takes over 1 hour, but utilizing **ZS** model's output can cut it down to 45-60 minutes, and **P-F** can bring it below 30 minutes.

6.3 Qualitative Analysis

Apart from quantitative human evaluation, we also randomly sample 10 slides and look at all the four configurations of those slides generated by our model P-F and the baseline. For instance, corresponding to the slide outline "Model Details", we obtain expert-long and non-expert-long configuration of slides (Figure 5) and similar set of configurations for slide outline "Results" in Figure 9. The striking difference between the technical and non-technical presentations is amount of technical complexity rendered in front of the audience on the same paper and on the same topic. In figures 14 and 15, non-relevant content based on slide outline is less compared to ones produced by baseline.

7 Related Work

Prior work on generating slides from documents have used both heuristic-based (Masum et al., 2005; Shibata and Kurohashi, 2005; Wang and Sumiya, 2013; Winters and Mathewson, 2019) (relying heavily on handcrafted features) and ML approaches (Bhandare et al., 2016; Syamili and Abraham, 2017; Sefid et al., 2019) to learn the importance of sentences and key phrases in each slide. However, they rely on extractive methods to fetch sentences from document as slide content. More recently, abstractive approaches based on diverse titles that summarize extracted content have been explored by (Sun et al., 2021; Fu et al., 2021). With respect to persona-aware response generation, some benchmark conversation datasets has been proposed to assess the conversation focusing on different personal attributes such as: (Xu et al., 2022b) presents a dialogue generation framework to update long-term persona memory without requiring datasets for model training. Recently, with the advent of LLMs, researchers have tried different ways as described in (Chen et al., 2023) to generate personalized dialogues (Lee et al., 2022; Xu et al., 2022a) and personalization in education (Li et al., 2023). However, a little attention has been paid to document to slides generation depending on target audiences' specifications.

8 Discussion and Conclusion

We introduce the concept of end-user specification-aware document to slides conversion that incorporates end-user specifications into the conversion process. Our novel three-step approach models human preferences in document to slide generation using human-in-the-loop. In future, we want to let humans exploit their creativity on top of the initial draft of persona-aware slides prepared by our models, through human-AI collaboration (Amershi et al., 2019), one could quickly create a slide deck improving the content and layout on-the-fly, generating or editing multimodal content through human textual feedback.

Limitations

Even though we receive good feedback from human experts on the created slides, we want to point out the two following limitations: 1) Our approach is limited to be faithful to document content, 2) Most of the technical jargons need to be explained to people with limited background either in terms of images or videos or definitions of jargons.

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

A. Ashray Bhandare, Chetan J. Awati, and Sonam Kharade. 2016. Automatic era: Presentation slides from academic paper. *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pages 809–814.

Ena Bhattacharyya. 2014. Walk the talk: Technical oral presentations of engineers in the 21st century. *Procedia - Social and Behavioral Sciences*, 123:344–352. TAYLOR'S 6TH TEACHING AND LEARNING CONFERENCE 2013: TRANSFORMATIVE HIGHER EDUCATION TEACHING AND LEARNING IN PRACTICE PROCEEDINGS OF THE TAYLOR'S 6TH TEACHING AND LEARNING CONFERENCE 2013 (TTLC2013), November 23, 2013, Taylor's University Lakeside Campus, Selangor Daruh Ehsan, Malaysia.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. 2023. When large language models meet personalization: Perspectives of challenges and opportunities.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, P. Abbeel, A. Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. In *Neural Information Processing Systems*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tsu-Jui Fu, William Yang Wang, Daniel J. McDuff, and Yale Song. 2021. Doc2ppt: Automatic presentation slides generation from scientific documents. In *AAAI Conference on Artificial Intelligence*.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating personalized dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach llms to personalize – an approach inspired by writing education.

Shaikh Mostafa Al Masum, Mitsuru Ishizuka, and Md. Tawhidul Islam. 2005. 'auto-presentation': a multi-agent system for building automatic multimodal presentation of a topic from world wide web information. *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 246–249.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Athar Sefid, Jian Wu, Prasenjit Mitra, and C. Lee Giles. 2019. Automatic slide generation for scientific papers. In *SciKnow@K-CAP*.

Tomohide Shibata and Sadao Kurohashi. 2005. Automatic slide generation based on discourse structure analysis. In *Second International Joint Conference on Natural Language Processing: Full Papers*.

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. 2021. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1418, Online. Association for Computational Linguistics.

S. Syamili and Anish Abraham. 2017. Presentation slides generation from scientific papers using support vector regression. *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 286–291.

Lyaylya Tarkhova, Sergey Tarkhov, Marat Nafikov, Ilshat Akhmetyanov, Dmitry Gusev, and Ramzid Akhmarov. 2020. Infographics and their application in the educational process. *International Journal of Emerging Technologies in Learning (iJET)*, 15(13):63–80.

Yuanyuan Wang and Kazutoshi Sumiya. 2013. A method for generating presentation slides based on expression styles using document structure. *Int. J. Knowl. Web Intell.*, 4(1):93–112.

Thomas Winters and K. Mathewson. 2019. Automatically generating engaging presentation slide decks. In *EvoMUSART*.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022a. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 201–211, New York, NY, USA. Association for Computing Machinery.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650, Dublin, Ireland. Association for Computational Linguistics.

Li Zhang, Steven Wilson, and Rada Mihalcea. 2019. Multi-label transfer learning for multi-relational semantic similarity. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 44–50, Minneapolis, Minnesota. Association for Computational Linguistics.

Chengbo Zheng, Dakuo Wang, April Yi Wang, and Xiaojuan Ma. 2022. Telling stories from computational notebooks: Ai-assisted presentation slides creation for presenting data science work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

	Correct	Incorrect	Can't Decide
Human-created	74.4%	15.6%	10%
SFT-P Generated	67.2%	17.3%	15.5%
P-F Generated	68.2%	12.5%	19.3%

Table 4: delves into the question of how accurately both experts and non-experts can discern whether a presentation is tailored for a technical audience or one with limited technical knowledge. The results underscore an intriguing aspect of human perception, revealing that *there is no unequivocal consensus*, and this observation holds true both when individuals are examining slides created by humans and those generated by our models.

	Correct	Incorrect	Can't Decide
Human-created	**94.4%**	3.2%	2.4%
SFT-P Generated	**91.2%**	7.3%	1.5%
P-F Generated	**89.7%**	8.2%	2.1%

Table 5: sheds light on the ability of both experts and non-experts to discern whether slides are tailored for short or long durations, revealing a *striking consensus among individuals* in making correct choice, whether they are examining slides crafted by human (94.4%) or those generated by our models (91.2%, 89.7%).

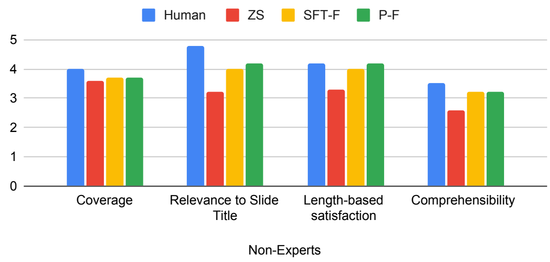


Figure 8: Average User Ratings by Non-Experts on 4 slide configurations (Human-created and 3 model-created) where we found that non-experts rate our model-generated slides higher on all criteria compared to baselines, but comprehensibility is low overall.

A Example Appendix

B Instructions to the Annotators for Evaluating the slide content

All the ratings for all outputs should be either 1, 2, 3, 4 or 5 (Likert Scale) Also, each of the presentation has table and figure captiions, You can consider that whenever table or figure is refered, they are present in slide deck. Now you can rate the quality of each slide based on the instructions below: **Coverage** (This criteria is based on how muc most of the content is present in a paper for a particular slide title): It speaks of whether all relevant details of a topic are present. Please assume

Configuration	Topics generated by ZS-TG	Topics generated by SFT-P TG	Topics generated by P-F TG
Non-Expert-Long	["Introduction to the WMT19 Metrics Shared Task", "Objective of the research paper", "Overview of the translation systems and metrics used", "Explanation of system-level evaluation", "Explanation of segment-level evaluation", "Importance of manual evaluation using direct assessment (DA)", "Summary of the results obtained", "Discussion on the impact of the research paper's approach", "Conclusion and future directions", "Q&A session"]	['Problem statement', 'Solution', 'System-level evaluation', 'Results', 'Segment-level evaluation', 'Analysis']	['Problem statement', 'Solution', 'Quality Estimation Metrics', 'Quality Analysis', 'Human Judgements', 'QE as a Metrics Analysis', 'Human Evaluations', 'Baseline Experiments', 'Data Set', 'Evaluation']

Table 6: Sample output predictions for topic generation algorithm.

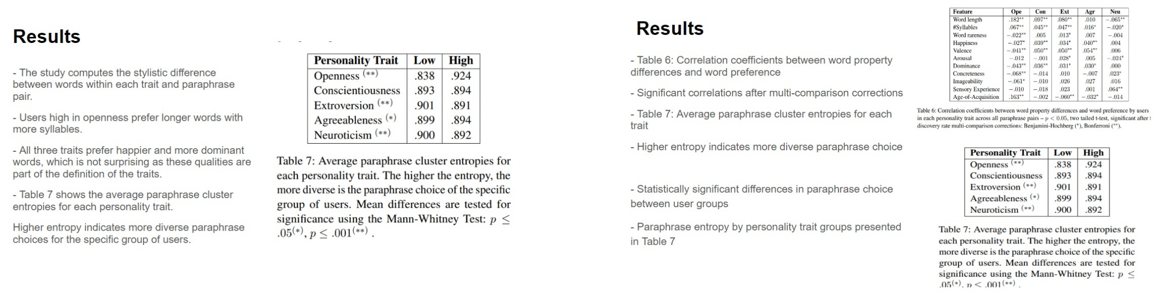


Figure 9: Here (a) is produced by **P-F** model for non-experts with explanations of phrases, and less technical jargons like 'statistical significance' and (b) is a technical results-heavy presentation for experts.

that this is a presentation, not every detail can be included
Relevance to Slide Title (How much are all the content in each slide relevant?): Whether all sentences, tables, figures in slides are relevant to the slide title
Fit for Length of Presentation or Length-based satisfaction: How much do you think that the slide title has sufficient amount of information (in a presentation) for long or short duration?) If the presentation is long, you can expect nitty gritty details on the paper, otherwise, we can settle on the most important and relevant content for a topic
Fit for the type of audience or Comprehensibility (How much do you think a technical expert or non-expert can follow the content well? You can see the type of presentation in Audience and Paper type.): Then you can rate whether output of each model are well understood by experts(who have prior knowledge) or non-experts (who have mild experience in research)?
Readability determines if the slide content is coherent, concise, and grammatically correct.

C Hiring Upwork Participants

C.1 Hiring Workers for Dataset Creation

Using Upwork, we hired two workers familiar with Machine learning and NLP with almost 5 years of experience and well-versed with creating presentations from documents, sorted by having a skill set of Presentation making. The hiring was made after shortlisting them through interviews, where they were initially asked to read the paper (Devlin et al., 2019) and answer questions like : 1) What is the novelty of this approach? 2) What is the motivation behind the main algorithm? 3) What are the strengths and weaknesses of this paper? 4) What is the state-of-art algorithm prior to this model? 5) What kind of evaluation has been made using this approach? Moreover, they were asked to make a presentation suitable for presenting it in an AI conference. Based on their answers and the quality of the presentation being made, the first two authors of the paper made a hiring decision.

C.2 Characterizing workers in Upwork into 'Experts' vs 'Non-Experts'

We wanted to have a clear distinction between who we call as technical 'experts' vs 'non-experts'. We hire twelve people using Upwork and characterize six of them into 'experts' and rest as 'non-experts'.

For understanding the depth and knowledge of the workers in NLP, Machine Learning research and their experience of attending prior AI conferences, we ask them to answer the following questions as shown in Figure 10 and Figure 11. The ones who have provided satisfactory answers to questions such as prior attendance to NLP conference, number of NLP papers they have read, answering convincing details about what they like and dislike in the paper, and also whether they had any rior publication. Three experts had prior publications, while other three had summarized the paper, strengths and weaknesses of the paper reasonably well. The non-experts community comprised mostly of data analysts, machine learning engineers who had no/limited prior experience in attending conferences.

We have used three experts and three non-experts for providing feedback (choosing one response over the other) on the model responses (both in topic generation and content extraction) during human-in-the-loop preference data collection as defined in Section 3.1.2.

The other three experts and three non-experts were asked to rate the quality of presentations at each step of the slide generation process as mentioned in Section 6. The instructions for both experts and non-experts are shown in Figure 12.

D Double checking Personalization of the Content Extraction module

Content customization for long vs short presentations were easy, but non-experts want more explanations of technical jargons. We hypothesize that asking users to distinguish generated samples between these two classes will serve as a proxy for assessing the *level of personalization* in the slides. We conduct a user study to assess the reader's capacity to identify whether the generated slides are tailored for long or short presentations/for technical experts or non-expert audiences. We sample 20 slides from papers in test set and generate variations for both long/short presentations, as well as for expert and non-expert audiences, using **human-created**, **SFT-P** and **P-F** models. Table 5 shows that 94.4% of the users could distinguish between the slides tailored for long vs short presentations. However, an interesting observation (Table 4) while distinguishing between technical vs non-technical presentation was that, the entropy between decision-making is quite high, revealing that *there is no unequivocal consensus*, and this observation holds true both when individuals are examining slides created by humans and those generated by our models. After uncovering these results, we talked to raters to explore the lack of consensus. Both human-created and model-generated slides contained technical content segments, making it difficult to choose one over the other. The key takeaway is the **pressing need for clearer technical explanations**.

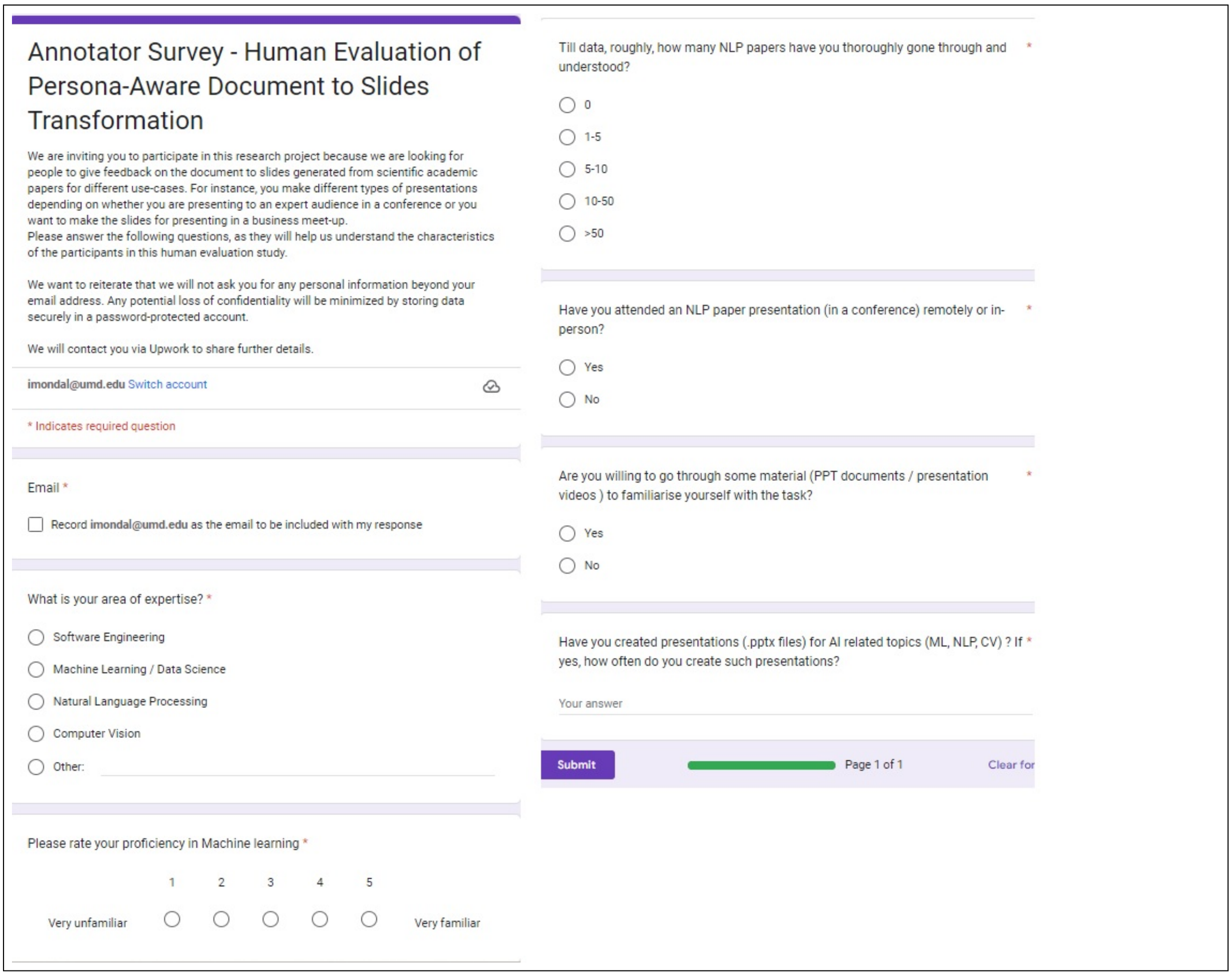
E Prompts

NZS-TG-Prompt="I want to present the paper with"+str(title)+" and abstract "+str(abstract)+" using a presentation. Can you create slide outlines for that? Format your response as JSON Object with keys as paperID and topics where paperID is the "+str(fileid)+" and the topics are a list of what you chose for making slides"

NZ-CE-prompt="You are creating a slide deck for presenting to people. In particular you want to create a slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

		Performance of Content Filter	
		Precision	Recall
Average GPT Calls	1	6.73	78.89
	5.3	5.93	81.34
	8.2	5.88	100

Table 7



Annotator Survey - Human Evaluation of Persona-Aware Document to Slides Transformation

We are inviting you to participate in this research project because we are looking for people to give feedback on the document to slides generated from scientific academic papers for different use-cases. For instance, you make different types of presentations depending on whether you are presenting to an expert audience in a conference or you want to make the slides for presenting in a business meet-up.
Please answer the following questions, as they will help us understand the characteristics of the participants in this human evaluation study.

We want to reiterate that we will not ask you for any personal information beyond your email address. Any potential loss of confidentiality will be minimized by storing data securely in a password-protected account.

We will contact you via Upwork to share further details.

imondal@umd.edu Switch account

* Indicates required question

Email *

☐ Record imondal@umd.edu as the email to be included with my response

What is your area of expertise? *

- Software Engineering
- Machine Learning / Data Science
- Natural Language Processing
- Computer Vision
- Other:

Please rate your proficiency in Machine learning *

	1	2	3	4	5	
Very unfamiliar	○	○	○	○	○	Very familiar

Till data, roughly, how many NLP papers have you thoroughly gone through and understood? *

- 0
- 1-5
- 5-10
- 10-50
- >50

Have you attended an NLP paper presentation (in a conference) remotely or in-person? *

- Yes
- No

Are you willing to go through some material (PPT documents / presentation videos) to familiarise yourself with the task? *

- Yes
- No

Have you created presentations (.pptx files) for AI related topics (ML, NLP, CV) ? If yes, how often do you create such presentations? *

Your answer

Submit Page 1 of 1 Clear for

Figure 10: Hiring of Expert and Non-Expert Annotators depending on their response to these questions.

	F1-score	Rouge-1	Rouge-L
GPT2 (text-davinci-002)	0.12	0.10	0.07
GPT3 (text-davinci-003)	0.32	0.13	0.12
GPT3.5-turbo	0.38	0.20	0.13

Table 8: Generalizability of our approach on three LLMs, where we report the zero-shot content extraction performance of all the models on the development set. All these models have the same set of slide outlines and the persona-aware constraints in their inputs in order to show a fair comparison. Stoked by the best performance of **GPT3.5-turbo**, we conduct all our experiments in the main paper using that model.

Zero-shot Personalized Content Extraction:

prompt for NS="You are creating a short slide deck for presenting to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

prompt for NL="You are creating a long slide deck for presenting to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

prompt for ES="You are creating a short slide deck for presenting to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

prompt for EL="You are creating a long slide deck for presenting to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

Few-shot Personalized Content extraction

prompt for NS="Follow the below example: Example: Output. You are creating a short slide deck for presenting to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

prompt for NL="Follow the below example: Example: Output. You are creating a long slide deck for presenting to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

prompt for ES="Follow the below example: Example: Output. You are creating a short slide deck for presenting to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

prompt for EL="Follow the below example: Example: Output. You are creating a long slide deck for presenting to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results. In particular you want to create slides on the topic of "+str(topic)+". Choose the sentences pertaining to the topic of "+str(topic)+" from the list of "+str(list of sentences) +" such that all the content should be informative, understandable, crisp, and all relevant and descriptive details. Only extract the sentences and format your answer as JSON with key as the topic "+str(topic)+"and value as the list of relevant sentences"

Zero-shot Topic Generator

NS="Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks." in this case can you make presentation slides which is short comprising of 4-5 topics.Format your response as JSON Object with keys as paperID and topics"

NL="Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks." in this case can you make presentation slides which is short comprising of 8-10 topics.Format your response as JSON Object with keys as paperID and topics"

ES="Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results in this case can you make presentation slides which is short comprising of 4-5 topics.Format your response as JSON Object with keys as paperID and topics"

EL="Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results in this case can you make presentation slides which is long comprising of 8-10 topics.Format your response as JSON Object with keys as paperID and topics"

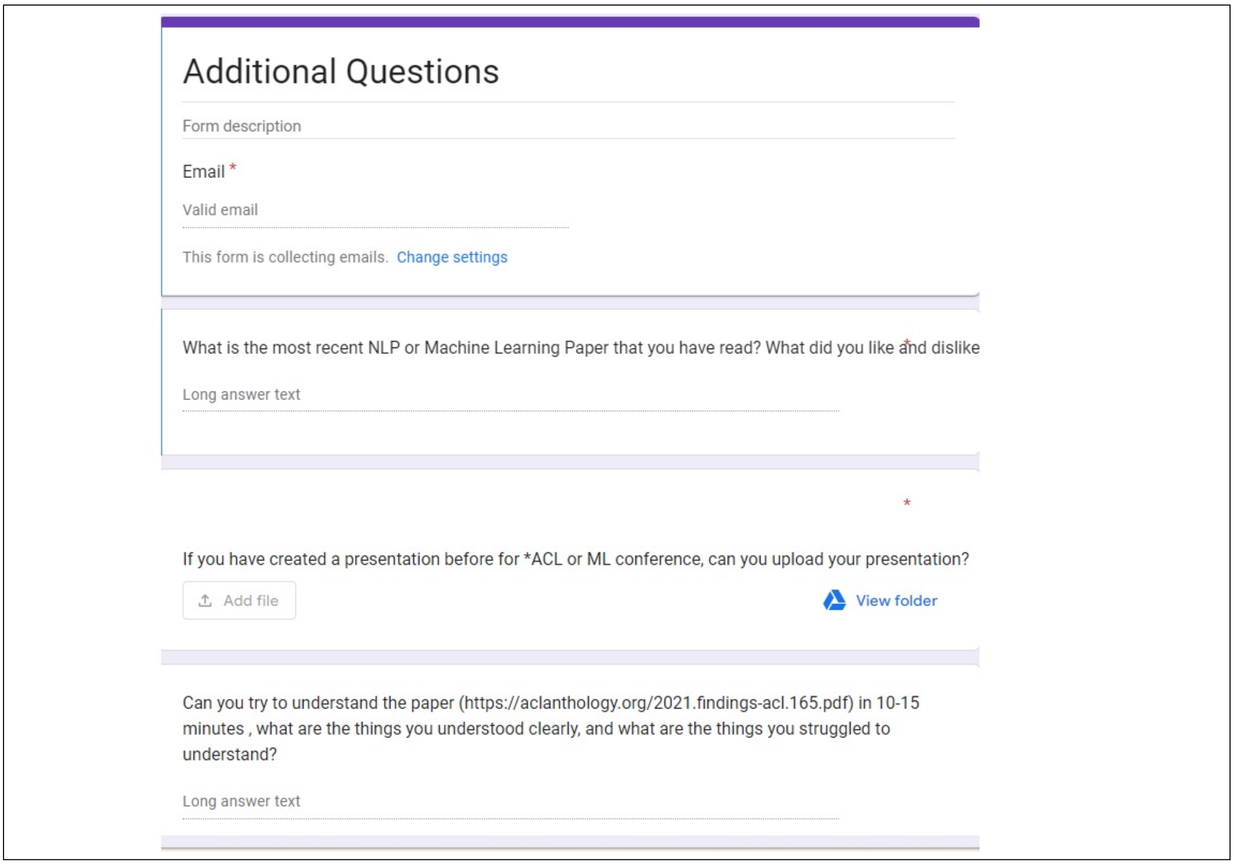
Few-shot Topic Generator:

NS="Follow the output of two examples: Example1: Output1, Example2: Output2. Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks." in this case can you make presentation slides which is short comprising of 4-5 topics.Format your response as JSON Object with keys as paperID and topics" \\

NL="Follow the output of two examples: Example1: Output1, Example2: Output2. Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the non-technical audience who cares mostly about the overall impact of the solution approach in the research paper. They don't understand any of the technical jargons used in the literature of machine learning and natural language processing tasks." in this case can you make presentation slides which is short comprising of 4-5 topics.Format your response as JSON Object with keys as paperID and topics" \\

ES="Follow the output of two examples: Example1: Output1, Example2: Output2. Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results in this case can you make presentation slides which is short comprising of 4-5 topics.Format your response as JSON Object with keys as paperID and topics" \\

EL="Follow the output of two examples: Example1: Output1, Example2: Output2. Find the answer for the prompt: 'Here is the title"+str(title) +"and abstract "+str(abstract)+" of the paper in the following usecase where I want to present the paper to the technical audience who wants to know the problem, solution, its impact, technical details, proofs and results in this case can you make presentation slides which is long comprising of 8-10 topics.Format your response as JSON Object with keys as paperID and topics" \\



Additional Questions

Form description

Email *

Valid email

This form is collecting emails. Change settings

What is the most recent NLP or Machine Learning Paper that you have read? What did you like and dislike *

Long answer text

*

If you have created a presentation before for *ACL or ML conference, can you upload your presentation?

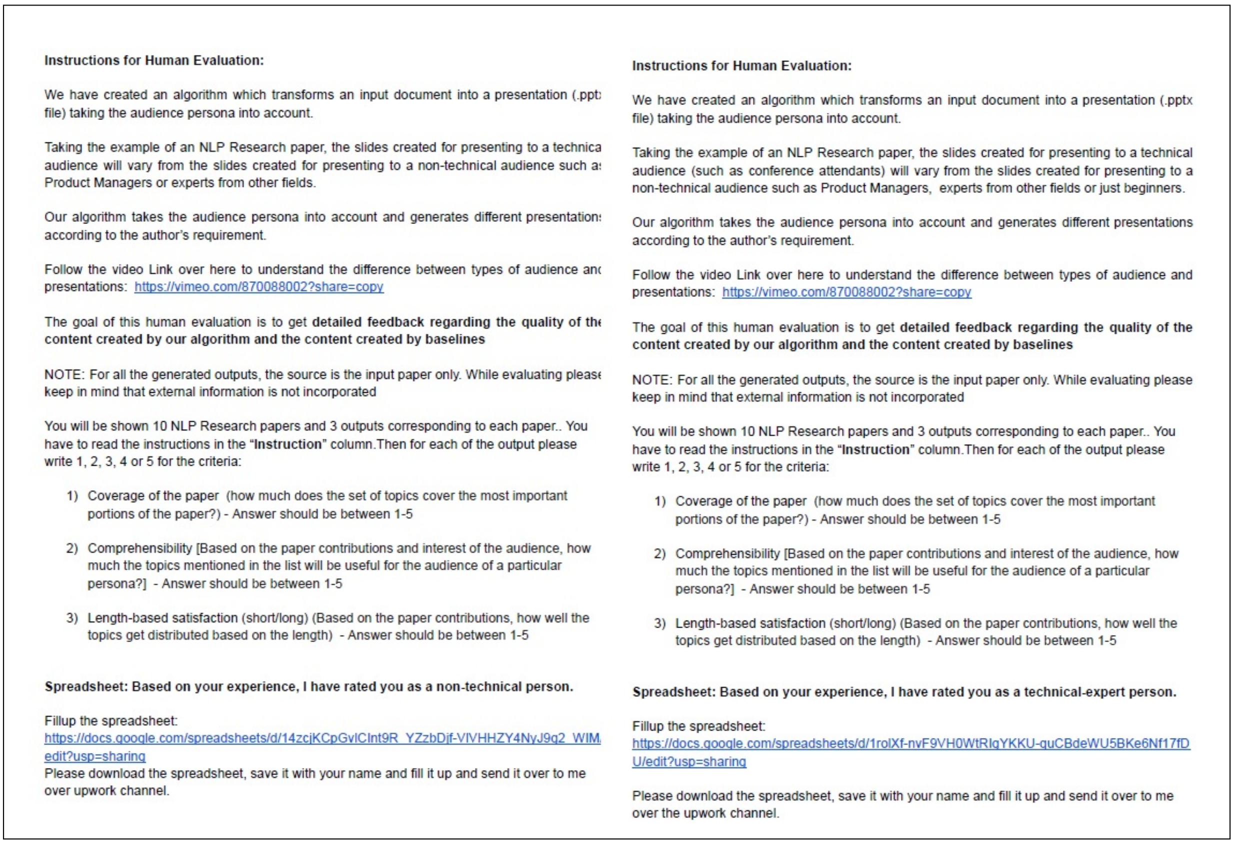
Add file

View folder

Can you try to understand the paper (https://aclanthology.org/2021.findings-acl.165.pdf) in 10-15 minutes , what are the things you understood clearly, and what are the things you struggled to understand?

Long answer text

Figure 11: Additional Questions while hiring the Expert and Non-Expert Annotators through Upwork.



Instructions for Human Evaluation:

We have created an algorithm which transforms an input document into a presentation (.ppt file) taking the audience persona into account.

Taking the example of an NLP Research paper, the slides created for presenting to a technica audience will vary from the slides created for presenting to a non-technical audience such a Product Managers or experts from other fields.

Our algorithm takes the audience persona into account and generates different presentation according to the author's requirement.

Follow the video Link over here to understand the difference between types of audience an presentations: https://vimeo.com/870088002?share=copy

The goal of this human evaluation is to get **detailed feedback regarding the quality of th content created by our algorithm and the content created by baselines**

NOTE: For all the generated outputs, the source is the input paper only. While evaluating pleas keep in mind that external information is not incorporated

You will be shown 10 NLP Research papers and 3 outputs corresponding to each paper.. You have to read the instructions in the "**Instruction**" column.Then for each of the output please write 1, 2, 3, 4 or 5 for the criteria:

1) Coverage of the paper (how much does the set of topics cover the most important portions of the paper?) - Answer should be between 1-5
2) Comprehensibility [Based on the paper contributions and interest of the audience, how much the topics mentioned in the list will be useful for the audience of a particular persona?] - Answer should be between 1-5
3) Length-based satisfaction (short/long) (Based on the paper contributions, how well the topics get distributed based on the length) - Answer should be between 1-5

Spreadsheet: Based on your experience, I have rated you as a non-technical person.

Fillup the spreadsheet:
https://docs.google.com/spreadsheets/d/14zcjKCpGvICInt9R_YZzbDjf-VIVHHZY4NyJ9q2_WIM edit?usp=sharing
Please download the spreadsheet, save it with your name and fill it up and send it over to me over upwork channel.

Instructions for Human Evaluation:

We have created an algorithm which transforms an input document into a presentation (.pptx file) taking the audience persona into account.

Taking the example of an NLP Research paper, the slides created for presenting to a technical audience (such as conference attendants) will vary from the slides created for presenting to a non-technical audience such as Product Managers, experts from other fields or just beginners.

Our algorithm takes the audience persona into account and generates different presentations according to the author's requirement.

Follow the video Link over here to understand the difference between types of audience and presentations: https://vimeo.com/870088002?share=copy

The goal of this human evaluation is to get **detailed feedback regarding the quality of the content created by our algorithm and the content created by baselines**

NOTE: For all the generated outputs, the source is the input paper only. While evaluating please keep in mind that external information is not incorporated

You will be shown 10 NLP Research papers and 3 outputs corresponding to each paper.. You have to read the instructions in the "**Instruction**" column.Then for each of the output please write 1, 2, 3, 4 or 5 for the criteria:

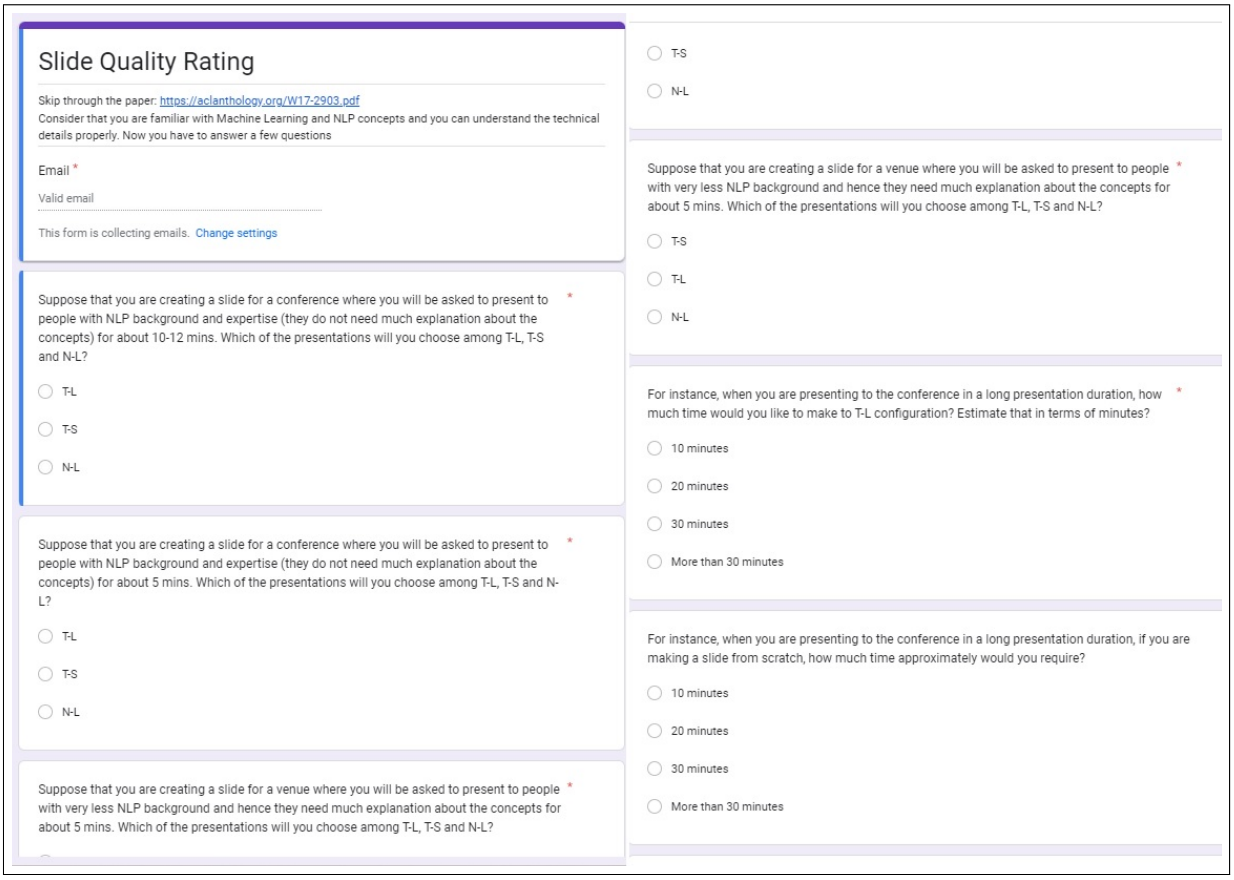
1) Coverage of the paper (how much does the set of topics cover the most important portions of the paper?) - Answer should be between 1-5
2) Comprehensibility [Based on the paper contributions and interest of the audience, how much the topics mentioned in the list will be useful for the audience of a particular persona?] - Answer should be between 1-5
3) Length-based satisfaction (short/long) (Based on the paper contributions, how well the topics get distributed based on the length) - Answer should be between 1-5

Spreadsheet: Based on your experience, I have rated you as a technical-expert person.

Fillup the spreadsheet:
https://docs.google.com/spreadsheets/d/1rolXf-nvF9VH0WtRIgYKKU-quCBdeWU5BKe6Nf17fDU/edit?usp=sharing

Please download the spreadsheet, save it with your name and fill it up and send it over to me over the upwork channel.

Figure 12: Instructions provided to the Expert and Non-Expert Audience to evaluate the slides.



Slide Quality Rating

Skip through the paper: https://aclanthology.org/W17-2903.pdf
Consider that you are familiar with Machine Learning and NLP concepts and you can understand the technical details properly. Now you have to answer a few questions

Email *

Valid email

This form is collecting emails. Change settings

Suppose that you are creating a slide for a conference where you will be asked to present to people with NLP background and expertise (they do not need much explanation about the concepts) for about 10-12 mins. Which of the presentations will you choose among T-L, T-S and N-L? *

- T-L
- T-S
- N-L

Suppose that you are creating a slide for a conference where you will be asked to present to people with NLP background and expertise (they do not need much explanation about the concepts) for about 5 mins. Which of the presentations will you choose among T-L, T-S and N-L? *

- T-L
- T-S
- N-L

Suppose that you are creating a slide for a venue where you will be asked to present to people with very less NLP background and hence they need much explanation about the concepts for about 5 mins. Which of the presentations will you choose among T-L, T-S and N-L? *

- T-S
- N-L

Suppose that you are creating a slide for a venue where you will be asked to present to people with very less NLP background and hence they need much explanation about the concepts for about 5 mins. Which of the presentations will you choose among T-L, T-S and N-L? *

- T-S
- T-L
- N-L

For instance, when you are presenting to the conference in a long presentation duration, how much time would you like to make to T-L configuration? Estimate that in terms of minutes? *

- 10 minutes
- 20 minutes
- 30 minutes
- More than 30 minutes

For instance, when you are presenting to the conference in a long presentation duration, if you are making a slide from scratch, how much time approximately would you require?

- 10 minutes
- 20 minutes
- 30 minutes
- More than 30 minutes

Figure 13: Assessing the reduction of cognitive Load (of expert authors) after creating persona-aware presentations from the documents

	Time required by Annotator 1	Time required by Annotator 2	Time required by Annotator 3
From Scratch	More than 1 hour	More than 1 hour	More than 1 hour
Z-S Generated	45-60 mins	More than 1 hour	45-60 mins
P-F Generated	Less than 30 mins	45-60 mins	Less than 30 mins

Table 9: presents the comparison of the ability of the expert authors (in terms of required time) to create their own presentations from scientific papers and tailored for non-expert audience having limited experience in NLP and Machine Learning with first-draft of slides generated from Zero-shot personalized approach (ZS-TG, ZS-CE, summarization and alignment), our proposed P-F approach and from scratch when they do not see any first draft.

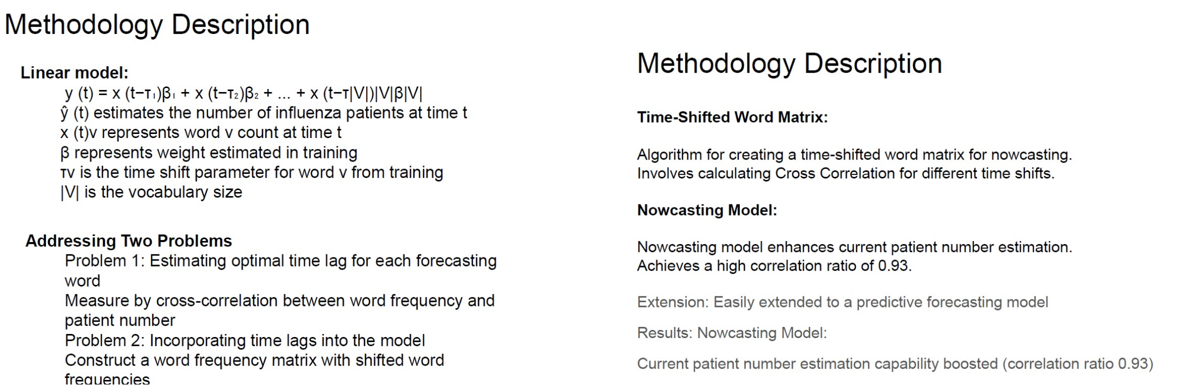
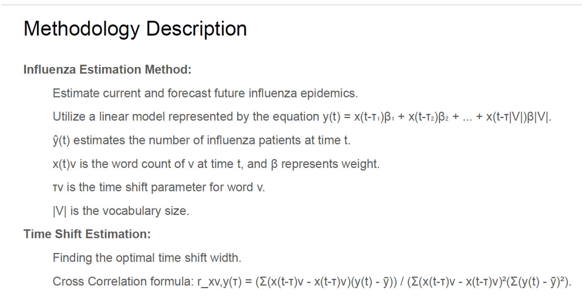


Figure 14: The slides generated from our baseline **ZS**-method based on the slide title "Methodology Description" which shows that in the first slide, we have some non-relevant content of "Addressing Two Problems", and in the second slide, we have non-relevant content on Results.



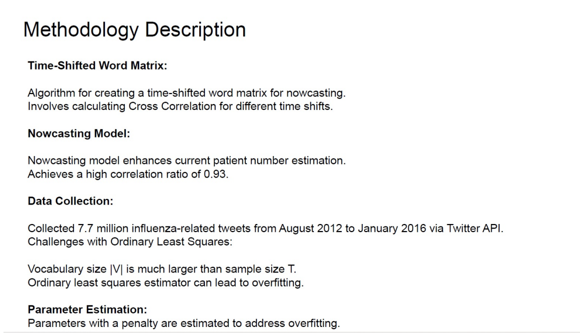


Figure 15: The slides generated from our proposed **Persona-Aware-D2S**-method based on the slide title "Methodology Description" which shows that in the first slide, we have some methods explained along with equations, and in the second slide, the model generates matrix, model and parameter estimation. Hence, non-relevant content is less compared to our baseline method. Moreover, it suffices the requirements of Expert Audience more than the content displayed by our baseline method.