Are Large Language Models In-Context Personalized Summarizers? Get an iCOPERNICUS Test Done!

Divya Patel [†*] **Pathik Patel** [†*] **Ankush Chander** [†*]
Sourish Dasgupta [†*] **Tanmoy Chakraborty** [‡]
[†] KDM Lab, Dhirubhai Ambani Institute of Information & Communication Technology, India
‡ Indian Institute of Technology, Delhi, India
{202001420, 202003002, ankush_chander}@daiict.ac.in,
✉ sourish_dasgupta@daiict.ac.in, ✉ tanchak@iitd.ac.in

Abstract

Large Language Models (LLMs) have succeeded considerably in In-Context-Learning (ICL) based summarization. However, saliency is subject to the users' specific preference histories. Hence, we need reliable *In-Context **Personalization** Learning* (IC**PL**) capabilities within such LLMs. For any arbitrary LLM to exhibit ICPL, it needs to have the **ability to discern contrast in user profiles**. A recent study proposed a measure for *degree-of-personalization* called EGISES for the first time. EGISES measures a model's responsiveness to user profile differences. However, it cannot test if a model utilizes all three types of cues provided in ICPL prompts: (i) example summaries, (ii) user's reading histories, and (iii) contrast in user profiles. To address this, we propose the iCOPERNICUS framework, a novel **I**n-**Co**ntext **Per**sonalization Lear**ni**ng S**cru**tiny of **S**ummarization capability in LLMs that uses EGISES as a comparative measure. As a case-study, we evaluate 17 state-of-the-art LLMs based on their reported ICL performances and observe that 15 models' ICPL degrades (min: 1.6%↓; max: 3.6%↓) when probed with richer prompts, thereby showing lack of *true* ICPL.

1 Introduction

With the constant influx of information, we need efficient models capable of summarizing essential content from lengthy documents for faster comprehension and prioritization (Ter Hoeve et al., 2022). Yet, defining what constitutes "salient" information remains subjective, particularly in documents covering multiple aspects. To tackle this, contemporary summarizers should be personalized to users' preference histories and interests.

Specialized PLMs as summarizers. Ao et al. (2021) proposed the most direct method to train models to learn personalization using user reading histories. These models (called PENS models) use variants of pointer-generator networks (See et al., 2017) that are injected with representations of user reading history for user preference alignment. Other indirect approaches include aspect-based models (Frermann and Klementiev, 2019; Hayashi et al., 2021) that produce summaries coherent with the document themes but lack adaptability to changes in the reader's profile. In contrast, interactive human-feedback-based models allow for iterative refinement based on user feedback, thereby better personalization (Ghodratnama et al., 2021).

LLMs as personalized summarizers. Recent studies on the state-of-the-art (SOTA) LLMs show unprecedented In-Context Learning (ICL) based summarization performance (Wang et al., 2023; Laskar et al., 2023; Tang et al., 2023). This opens the possibility of In-Context Personalization Learning (ICPL) in these LLMs. At the same time, it also underscores the necessity for robust and dependable methods of evaluating the degree of ICPL within such models. Although benchmarking of LLMs for summarization has been done for accuracy, fluency, and consistency (Zhang et al., 2024), so far, no study has been done yet on the probing and evaluation of ICPL in LLMs for the summarization task. In this paper, we propose iCOPERNICUS, an **I**n-**Co**ntext **Per**sonalization Lear**ni**ng S**cru**tiny of **S**ummarization capability in LLMs.

iCOPERNICUS framework. iCOPERNICUS is a prompt-based probing framework that investigates whether LLMs exhibit true ICPL using a 3-pronged approach: (i) whether few-shot prompting of **examples (i.e., reader-generated gold references) enhances ICPL**, (ii) whether adding reader's **reading history improves ICPL**, and (iii) whether **contrastive profile information** showing subjective differences in reader-preferences for the same document **induce better ICPL**. Since iCOPERNICUS is a *comparative* framework, it needs a personalization measure for analyzing the influence of the injected profile information in the prompts. We use

*Equal contributions.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16820–16842
November 12-16, 2024

Model Variants Probed	
Base-models	Llama 2 (7B, 13B) (Touvron et al., 2023)
	Mistral v0.1 (7B) (Jiang et al., 2023)
Instruct-tuned	Mistral 7B Instruct v0.1 (Jiang et al., 2023)
	Mistral 7B Instruct v0.2 (Jiang et al., 2023)
	Tulu V2 (7B, 13B) (Ivison et al., 2023)
	Orca 2 (7B, 13B) (Mitra et al., 2023)
	Stable Beluga (7B, 13B) (Mahan et al., 2023)
RLHF-tuned	Llama 2 Chat (7B, 13B) (Touvron et al., 2023)
DPO-tuned	Tulu V2 DPO (7B, 13B) (Ivison et al., 2023)
	Zephyr 7B α (Tunstall et al., 2023)
	Zephyr 7B β (Tunstall et al., 2023)

Table 1: LLMs probed for ICPL w.r.t summarization.

EGISES-JSD, the only known measure for personalized summarization (Vansh et al., 2023). EGISES measures the ability of models to discern the differences in user profiles and generate summaries that are proportionately different.

Observations As a case-study of the application of `iCOPERNICUS`, we probe ten SOTA LLMs that exhibit reasonably good ICL on standard benchmark tasks, six of which have 7B and 13B model size variants (see Table 1). We use the PENS dataset (Ao et al., 2021) as in (Vansh et al., 2023) to compare the ICPL results with the baseline specialized personalized summarization models evaluated therein. We observe that all the studied models, except for Orca-2 7B and Zephyr 7B β, exhibit performance degradation in all the three probes of `iCOPERNICUS` - i.e., injection of examples, history, and contrastive profile information.

Key Contributions. We present for the first time: (i) a detailed introduction of ICPL for personalized summarization, (ii) `iCOPERNICUS` as a formal framework of evaluation of ICPL in LLMs, and (iii) a detailed case-study of the application of `iCOPERNICUS` tests for determining ICPL in the selected SOTA LLMs.

2 Preliminaries

2.1 Personalization w.r.t Summarization

As proposed in Vansh et al. (2023), the *degree-of-personalization* is a quantitative measure of how much a summarization model fine-tuned for personalization is adaptive to a user's (i.e., reader's) subjective expectation. This also implies that it measures how accurately a model can capture the *user's "evolving"* ***profile*** *reflected through* ***reading history*** (i.e., a temporal span of the reading and skipping actions of a user on a sequence of documents that is interleaved by the actions of generating and reading summaries). This is because the *subjective expectation is a function of the reading history*. **A low degree of personalization, by definition, implies poor user experience (UX)**. If a model does not efficiently capture the user's profile, it may lead to irrelevant summaries. In this situation, poor UX would mean that the user would have to spend more time getting to the information he/she is interested in or suffer from information overload and fatigue. However, this irrelevant information can be useful for a different user with a different profile. To illustrate this, we borrow the example given by Vansh et al. (2023) where if reader Alice, who has been following "*civilian distress*" in the Hamas-Israeli conflict, reads a news summary whose content is primarily about "*warfront events*", her UX will drop down due to information overload and high time-to-consume, even though her interest is also covered to a fair extent. In contrary, a reader Bob, who has been mostly following war news, would have quite high UX.

2.2 EGISES: Personalization Measure

Vansh et al. (2023) showed theoretically and empirically that a model could have high accuracy scores in both the examples in the previous section, although the individual degree of personalization differs. This can *mislead selection of a model for a fairly high accuracy score even though it suffers from poor UX*. To address this, they proposed EGISES as, to the best of our knowledge, the only known measure for personalized summarization evaluation. Informally, EGISES measures the extent to which a model can discern the differences between user profiles and generate summaries that have proportionate differences. Difference is modeled as a chosen divergence on a metric space. See Appendix A.2 for formulation.

2.3 In-Context Personalization Learning

In-Context learning (ICL) is an emergent phenomenon exhibited LLMs (first highlighted in Brown et al. (2020) for GPT-3), where models acquire proficiency in unknown tasks on which they are not pre-trained from limited examples, called *prompts*, with no update in their parameters (i.e., the models are frozen).[1] In-Context Personalization Learning (ICPL) for summarization is a special case of ICL where, for a document query d_q, an LLM is expected to generate the user-specific optimal summary $s^*_{(d_q,h_j)}$, for the j-th user expecting

[1]For formalizations of ICL see Appendix A.3.

the summary of d_q. Here, h_j is the user's **reading history** (temporal sequence of the user's click and skip history of documents). $s^*_{(d_q,h_j)}$ is the same as the j-the user's expected summary u_{qj} (i.e., gold-reference), and hence can be denoted $s^*_{u_{qj}}$.

Definition 1. ***Prompt4ICPL****: A prompt $\mathcal{P}_{ICPL}$ to a language model M consists of an user's reading history (h), an optional sequence of n concatenated ($\oplus$) demonstration examples (i.e., input-label pairs) as: $h_j \oplus \bigoplus_{i=1}^{n}(d_i \oplus u_{ij})$ where, d_i is the example document to be summarized for j-th user (u_{ij} being the **gold-reference summary example**), and a query document d_q, such that $d_i \neq d_q$.*

A **zero-shot prompt (0-shot)** is the special case when demonstration examples are not provided (i.e., $u_{ij} = \emptyset$) while d_q and the user reading history h_j are given in the prompt. The few-shot version can be of two types: (i) **with history (k-shot w/hist.)**, and (ii) **without history (k-shot w/o hist.)**. In the second type, the user profile is not represented by reading history but rather by the expected summaries (or gold-reference summaries). This can seen as the user's "*writing history*".

3 The iCOPERNICUS Framework

iCOPERNICUS is a novel prompt-based three-pronged probing framework for evaluating ICPL in LLMs. It tests *whether the model can harness **three types of profile information** included within the test prompts: (i) **examples**, (ii) **history**, and (iii) **contrastive information** (in terms of history and examples)*. Before we provide a detailed outline of the framework, we first discuss the contrastive probing setup in the following section.

3.1 Contrastive Probing

One of the key probes of the iCOPERNICUS framework is testing LLMs for ICPL with *contrastive examples*, i.e., input-label pairs containing at least two user (i.e., reader) profiles (can be reading or writing history) with the query document d_q. An LLM capable of ICPL should be able to discern the difference between the reader profiles and generate summaries accordingly (i.e., $s_{u_{ij}}$ and $s_{u_{ik}}$) in line with the notion of insensitivity-to-subjectivity as defined by Vansh et al. (2023) (see Appendix A.1). Contrastive Prompt4ICPL ($\mathcal{P}_C$) is defined as:

Definition 2. ***Contrastive Prompt4ICPL*** *($\mathcal{P}_C$): A $\mathcal{P}_C$ given to a language model M is a sequence of n concatenated ($\oplus$) **contrastive** demonstration examples $\mathcal{D}_{\mathcal{P}_C}$ as: $(\bigoplus_{j=1}^{m} h_j) \oplus (\bigoplus_{i=1}^{n}(d_i \oplus \bigoplus_{j=1}^{m} u_{ij}))$, each having m **subjective** expected summaries (u_{ij}), and a query document d_q, s.t. $d_i \neq d_q$.*

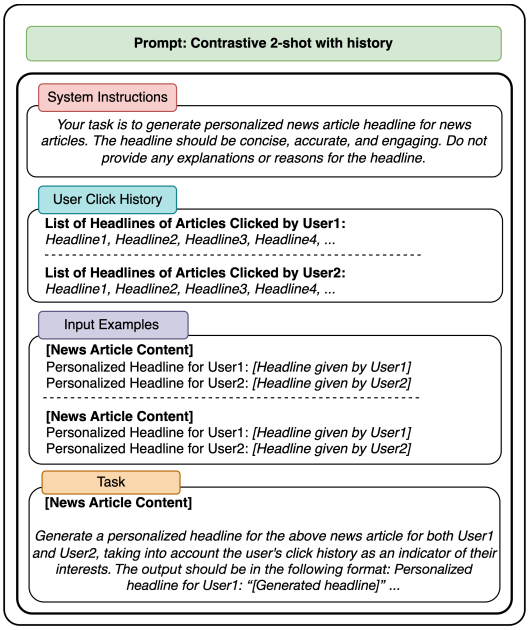


Figure 1: $\mathcal{P}_C$ type: Contrastive (C)-2-shot w/ history.

A **contrastive zero-shot prompt (C-0-shot)**, contains h_j and h_k representing the contrastive reading-histories of two users j and k with no demonstration examples. The few-shot version, similar to k-shot (plain) prompt, is of two types: (i) **C-k-shot w/hist.** (Figure 1) and (ii) **C-k-shot w/o hist**. In the second case, gold-reference summaries (writing history) of **both users** are given in the prompt as examples but not their reading histories.[2] We now define ICPL (weak and strong cases) as:

Definition 3. ***Weak ICPL.*** *Given a contrastive prompt $\mathcal{P}_C$, an LLM $M_{\boldsymbol{\theta},u}$ exhibits weak ICPL, if $\forall(u_{qj}, u_{qk})$ w.r.t query document d_q, $(\sigma(u_{qj}, u_{qk}) \leq \tau^U_{max}) \iff (\sigma(s_{u_{qj}}, s_{u_{qj}}) \leq \tau^S_{max})$; $\tau^U_{max}, \tau^S_{max}$ are bounds within which users' expected (gold-reference) summary pair (u_{qj}, u_{qk}) and LLM-generated summary pair $(s_{u_{qj}}, s_{u_{qj}})$ are indistinguishable; σ is an arbitrary distance metric on the metric space $\mathcal{M}$, where d, u, s are defined.*

Definition 4. ***Strong ICPL.*** *Given $\mathcal{P}_C$, $M_{\boldsymbol{\theta},u}$ exhibits strong ICPL, if $\forall(u_{qj}, u_{qk})$ w.r.t d_q, $M_{\boldsymbol{\theta},u}$ satisfies: (i) weak ICPL, and (ii) $(\sigma(u_{qj}, u_{qk}) >$*

[2]NT: *For sake of clarity, contrastive prompts are not chain-of-thought (CoT) prompts as it does not contain thought breakdown or require thought generation.*

$\tau^U_{max}) \iff (\sigma(s_{u_{qj}}, s_{u_{qj}}) > \tau^S_{max})$.

The following sections describe the three-pronged iCOPERNICUS framework.

3.2 Probe 1: Do example summaries help?

The first probe within the iCOPERNICUS framework studies the impact of k-shot prompts (in contrast to 0-shot prompts) on LLM models.[3] The (plain) k-shot w/o hist. prompt-based probing investigates whether the model improves ICPL performance w.r.t EGISES-scores by mapping the key latent concepts in each example summary with those in the corresponding document for any given user. This is a much richer cue than the 0-shot case, where the model does not get much assistance but can only observe inter-document conceptual associations (i.e., user's click and skip patterns) provided as a part of the reading history h. A richer version of this probe is that with k-shot w/ hist. prompts, where additional history information is also provided to investigate if the model can associate that with the examples provided. These two probes should also be performed in the contrastive prompt setting (C-k-shot prompts (w/ and w/o hist.)) to investigate the presence of the same kind of associations with the additional capacity of associating user-specific profile concepts. More specifically, ideally, a model should be able to **harness example summaries** in the following order of ICPL performance:

1. (plain) 0-shot $\prec$ **(plain) k-shot w/o hist.**, violation leads to **Paradox 1 (PX-1)**.
2. (plain) 0-shot $\prec$ **(plain) k-shot w/ hist.**, violation leads to **Paradox 1 w/ hist. (PX-1-h)**.
3. C-0-shot $\prec$ **C-k-shot w/o hist.**, violation leads to **PX-1 (contrastive)**.
4. C-0-shot $\prec$ **C-k-shot w/ hist.**, violation leads to **PX-1-h (contrastive)**.

3.3 Probe 2: Does reading-history help?

The second probe investigates whether a model can utilize the temporal sequence of a specific user's reading (i.e., document clicking and skipping) history and associate the innate latent concepts with that of the corresponding example summaries provided in the prompt (both plain and contrastive). Hence, a model should be able to **harness reading history** in the following ICPL performance order:

1. (plain) k-shot w/o hist. $\prec$ **(plain) k-shot w/ hist.**, violation leads to **PX-2**.
2. C-k-shot w/o hist. $\prec$ **C-k-shot w/ hist.**, violation leads to **PX-2 (contrastive)**.

3.4 Probe 3: Do contrastive prompts help?

The third probe investigates whether models can capitalize on additional contrasting (i.e., similarity/differences) information about user profiles in contrastive prompts. Hence, a model should be able to **harness contrastive information** in the following order of ICPL performance:

1. (plain) 0-shot $\prec$ **C-0-shot**, violation leads to **PX-3**.
2. (plain) k-shot w/o hist. $\prec$ **C-k-shot w/o hist.**, violation leads to **PX-4**.
3. (plain) k-shot w/ hist. $\prec$ **C-k-shot w/ hist.**, violation leads to **PX-5**.

3.5 Limitations of EGISES w.r.t ICPL

As can be seen from the previous section, iCOPERNICUS is a *comparative* framework that needs a personalization measure for analysis of the influence of profile information. Since it was reported in Vansh et al. (2023) that the EGISES-JSD variant performed well regarding human-judgment correlation, we choose this specific variant as the comparative measure within the iCOPERNICUS framework to evaluate system-level strong degree-of-ICPL (i.e., σ in definitions 3 and 4 is JSD (Jensen-Shannon Divergence)).[4] However, it is to be noted that, although iCOPERNICUS uses EGISES scores as a comparative measure within the context of the tests in the three probes, it is not tightly coupled with EGISES (or, for that matter, any personalization measure). In other words, iCOPERNICUS **is not a metric but a test-suite**. Even a hypothetical "perfect" personalization measure would still be a system-level metric that would evaluate an LLM in terms of its ability to be responsive to the difference in users' subjective expected summaries **in a specific test setting only** (out of the identified 9 probe settings) **but would not be able to detect whether the LLM improves on this hypothetical perfect metric when aided with more user information**, which iCOPERNICUS does. In section 5.5, we empirically show that absolute EGISES score-based ICPL leaderboards can be misleading.

[3] It is to be noted that, as pointed out in Section 3.1, 0-shot prompts contain the reading histories (h).

[4] NT: EGISES as a standalone measure **has not been designed to test ICPL** (see Appendix A.2 for formulation).

Prompt Style	Reading Hist.	Examples	Article Body	# Prompts
0-shot	1200 Tokens	–	2500 Tokens	6856
C-0-shot	1000 x 2 Tokens	–	1700 Tokens	5246
2-shot w/o hist.	–	950 x 2 Tokens	1800 Tokens	6798
C-2-shot w/o hist.	–	950 x 2 Tokens	1800 Tokens	5246
2-shot w/ hist.	1200 Tokens	600 x 2 Tokens	1300 Tokens	6798
C-2-shot w/ hist.	850 x 2 Tokens	450 x 2 Tokens	1100 Tokens	5246

Table 2: iCOPERNICUS prompt composition (w.r.t # of tokens) for all prompt styles; **NT**: overall prompt size is fixed.

4 Evaluation: Setup

4.1 Model Benchmarking Dataset

Evaluating the selected models w.r.t iCOPERNICUS requires the test dataset to contain (i) example (and expected) gold summaries, (ii) user's reading history, and (iii) contrastive examples (i.e., *subjective* gold summaries). We selected the test data sourced from the PENS dataset (Ao et al., 2021)[5] for our purpose since, to the best of our knowledge, it is the only dataset containing all the above three. 103 college students, having English as their native language, were invited as voluntary participants. A two-phase process was adopted to construct the test set. Initially, the participants selected at least 50 articles of personal interest from a pool of 1000 news articles, which were then sorted based on exposure time. **This formed their reading histories**. Subsequently, participants in the second phase created preferred headlines (gold references) for 200 news articles without prior knowledge of the original headlines. **This formed the set of examples and expected (personalized) summaries**. The two-stage process ensures an average of four gold-reference summaries per article, **thereby enabling contrastive prompts to be sampled out**. For details, see Appendix C.

4.2 Probing Dataset Creation

We engineer six distinct prompt templates in accordance with the iCOPERNICUS framework (see Table 2). The prompts were sampled from the PENS dataset (section 4.1) with sample size of 3840 news articles such that the total number of tokens for all the settings were 3700 - i.e., **the overall prompt size remained constant**. This was done so that the probes were comparable in a controlled environment. Depending on the specific test, each prompt has been broken up into history, examples, and article body.[6] The dataset is released for research purposes at KDM-Lab_iCOPERNICUS_prompt-dataset_v1.0.

4.3 Probed SOTA LLMs

We probe ten SOTA LLMs with their 7B and 13B variants (see Tables 1 and 3), totaling seventeen variants. Models are chosen based on their recency, training data diversity, and performance on key benchmark tasks requiring comparative reasoning.[7] We could not evaluate 13B+ models due to compute resource constraints. However, the core contribution of the paper is the iCOPERNICUS framework itself which is *applicable for selection decision of any sized model*. We do not intend to make any claim regarding the generalizability of the findings but rather point out the fact that if LLMs are to be used as personalized summarizers, one should let them go through the iCOPERNICUS test-suite since unless we do that, we will never know whether the paradoxes exist. Hence, the evaluations is primarily a *case study of the application of* iCOPERNICUS. Appendix B.1 has model descriptions.

4.3.1 Baseline Summarization Models

To understand whether the probed LLMs are superior to specialized PLMs trained on personalized summarization tasks, we examine the same ten SOTA summarization models as in (Vansh et al., 2023) for comparative benchmarking. Five of them are specifically trained personalized models that follow the PENS framework (Ao et al., 2021): (i) PENS-NRMS Injection-Type 1 (PENS-NRMS T1), (ii) PENS-NRMS Injection-Type 2 (PENS-NRMS T2), (iii) PENS-NAML T1, (iv) PENS-EBNR T1, and (v) PENS-EBNR T2. These models encode the document article using the Transformer encoder (Vaswani et al., 2017), deep-neural-model-based user history encoders (Okura et al., 2017; Wu et al., 2019a,b), and a Pointer-generator-network-based

[5] We comply with the Microsoft Research License Terms.

[6] See Figure 4 for the structure of the prompts and Figures 6-10 in the Appendix for examples.

[7] Tasks: commonsense reasoning (e.g., Winogrande (Sakaguchi et al., 2019), Hellaswag (Zellers et al., 2019)), math (e.g., GSM8k (Cobbe et al., 2021)), code (e.g., MBPP (Austin et al., 2021)), and multi-task benchmarks (e.g., MMLU (Hendrycks et al., 2020), AGI Eval (Zhong et al., 2023))

	LLM Model Variants	0-shot	2-shot w/o hist.	2-shot w/ hist.	C-0-shot	C-2-shot w/o hist.	C-2-shot w hist.
Base models	Llama 2 7B	0.408	0.367	0.367	0.408	0.46	0.458
	Llama 2 13B	0.412	0.371	0.357	0.418	0.5	0.474
	Mistral 7B v0.1	0.398	0.353	0.354	0.464	0.406	0.469
Instruction-tuned	Mistral 7B Instruct v0.1	0.4	0.366	0.378	0.395	0.405	0.399
	Mistral 7B Instruct v0.2	0.391	0.348	**0.354**	0.359	**0.339**	0.369
	Tulu V2 7B	0.364	0.376	0.356	0.36	0.38	0.37
	Tulu V2 13B	0.376	0.389	0.385	0.387	0.405	0.392
	Orca 2 7B	0.44	0.436	0.433	0.359	0.351	0.347
	Orca 2 13B	0.445	0.442	0.447	0.347	0.366	0.356
	Stable Beluga 7B	0.371	0.395	0.407	0.377	0.398	0.396
	Stable Beluga 13B	0.388	0.394	0.404	0.4	0.405	0.412
RLHF-tuned	Llama 2 7B Chat	0.383	0.391	0.362	**0.333**	0.349	**0.338**
	Llama 2 13B Chat	0.36	0.393	0.383	0.341	0.365	0.357
DPO-tuned	Tulu V2 DPO 7B	**0.325**	**0.345**	0.356	0.338	0.355	0.348
	Tulu V2 DPO 13B	0.359	**0.345**	0.385	0.37	0.383	0.368
	Zephyr 7B α	0.360	0.351	0.357	0.343	0.352	0.353
	Zephyr 7B β	0.384	0.359	0.369	0.35	0.356	0.345

Table 3: **iCOPERNICUS Probe Results**: Master table for all comparative analysis including the detection of the nine potential paradoxes that can arise due to the three probes outlined in Section 3; EGISES-JSD is used for the comparative evaluation (lower is better); Table 4, a summary of the observed paradoxes, is derived from this table. Evaluation Script: https://github.com/KDM-LAB/iCOPERNICUS-EMNLP24

(See et al., 2017) decoder for generating the personalized summaries. The other five non-personalized models are generic SOTA summarizers – BRIO (Liu et al., 2022), SimCLS (Liu and Liu, 2021), BigBird-Pegasus (Zaheer et al., 2020), ProphetNet (Qi et al., 2020), and T5-base (Orzhenovskii, 2021). These models were evaluated by providing documents enriched with headlines (reference summaries), serving as cues (Vansh et al., 2023). Since the baseline models *are incapable of ICPL*, **iCOPERNICUS** ***tests are inapplicable*** for them, and EGISES-based evaluation is sufficient. The model descriptions are in Appendix B.2.

4.3.2 Hyperparameter Selection

We conduct temperature ablation within the interval $[0.5, 0.75]$ to balance accuracy and diversity for personalized summarization and **observe almost similar results** to the selected 0.6.[8] Better ICPL performance might be seen for specific models under more comprehensive ablation, but finding an optimal configuration that generalizes for all LLMs is hard. Nevertheless, the current evaluation is a *useful indication of* ***potential paradoxes*** *and the* ***possibility of misguided model selection*** *if one relies solely on an EGISES-based leaderboard* (see section 5.5 for empirical results).[9]

[8] See Appendix F for inference environment (LLM settings and compute resources).

[9] In a way, iCOPERNICUS tests show the need for detailed hyperparameter optimization before model selection.

5 Observations and Insights

5.1 Effect of Examples (User-Summaries)

In the probe 1, we find that **10 model variants (out of 17) exhibit an increase in ICPL** (i.e., EGISES-JSD scores) by an average of 2.6%↑ for 2-shot prompts (w/o reading history (hist.)) w.r.t zero-shot (see PX-1 col. of Table 4 for the performing models (denoted as: ✗)). However, seven models degrade with (plain) 2-shot (w/o history) prompts (average drop of 1.7%↓), leading to the first of the five observed ***paradoxes of less is more*** as outlined in section 3 (see Table 4 for result summary).

(PX-1) Implicit is more than explicit: We believe that these seven models learn more from the latent concept association in the (temporal) reader's history at a broader level than the specific concepts within the example summaries (i.e., explicit reader-profile as "*writing-history*"), the replacement of which makes them deviate from their earlier ICPL performance. *However, all these models show significant ICPL boost w.r.t their respective base models for the 2-shot w/o hist. case* (denoted by †). A real example of PX-1 can be seen in Figure 6.

5.2 Effect of User's Reading History

As a part of the second probe w.r.t iCOPERNICUS, we observe that **10 model variants exhibit an increase in ICPL** (avg. boost: 2.5%↑) for 2-shot w/ hist. w.r.t zero-shot. It is also observed that the set of LLMs that show ICPL in the previous case does not take advantage of the additional history data, leading to the second of the paradoxes (i.e., PX-2).

	Model Variants	PX-1	PX-2	PX-3	PX-4	PX-5
Base-Model	Llama 2 7B	✗	✗	✓	✓	✓
	Llama 2 13B	✗	✗	✓	✓	✓
	Mistral 7B v0.1	✗	✗	✓	✓	✓
Instruct-tuned	Mistral 7B Instr. v0.1	✗	✗	✗	✓	✓
	Mistral 7B Instr. v0.2	✗	✗	✗	✗	✓
	Tulu V2 7B	✓†	✗	✗	✓	✓
	Tulu V2 13B	✓†	✓	✓	✓	✓
	Orca 2 7B	✗	✗	✗	✗	✗
	Orca 2 13B	✗	✓	✗	✗	✓
	Stable Beluga 7B	✓†	✓	✓	✓	✓
	Stable Beluga 13B	✓†	✓	✓	✓	✓
RLHF-tuned	Llama 2 7B Chat	✓†	✗	✗	✗	✓
	Llama 2 13B Chat	✓†	✓	✗	✗	✓
DPO-tuned	Tulu V2 DPO 7B	✓†	✓	✓	✓	✓
	Tulu V2 DPO 13B	✗	✓	✓	✓	✗
	Zephyr 7B α	✗	✗	✗	✓	✓
	Zephyr 7B β	✗	✗	✗	✗	✗

Table 4: **Paradox (PX) of less is more (✓: PX exists):** PX-1/2: 2-shot w/o & w/ hist.; PX-3/4/5: C-0-shot/C-2-shot w/o & w/ hist.; † denotes improvement over base models; for examples see Figures 6-10 in Appendix.

(PX-2) Reading history distracts: Seven model variants show worse ICPL with the additional history data (avg. drop: 2%↓; see Table 4). We believe that these models tend to learn more from the format of the prompt and latent-concept association but at a rather broader thematic level and get "*distracted*" (Shi et al., 2023) by the concept distribution within histories. A real example of PX-2 can be seen in Figure 7.

5.3 Effect of Contrastive Prompts

The third probe tests if contrastive user profiles (C-0-shot, C-2-shot w/o hist., C-2-shot w/ hist.) induce better ICPL in the models. This probe is central for any model to pass the `iCOPERNICUS` test.

Case of C-0-shot: We observe that **9 model variants seem to harness the additional contrastive information about the readers' reading history** in comparison with the 0-shot case (avg. boost: 3.8%↑, see PX-3 col. of Table 4 for performing models), which is better than the overall boost observed with (plain) 2-shot (w/ and w/o history). This indicates that these models might actually be utilizing the contrastive information in the reading histories rather than the examples without the contrast. However, this ICPL behavior is not observed in the remaining eight model variants, leading to a special case of the **paradox of contrastive user profiles** - PX-3. We will discuss this subsequently.

Case of C-2-shot w/o history: We observe that only **six model variants pass this probe test and seem to harness the contrastive examples quite notably** (p-value < 0.01) when compared to the case of 2-shot w/o history (avg. boost: 4.1%↑) (see PX-4 col. in Table 4 for these 6 models). In fact, Llama 2 13B Chat, one of the 6 variants, degrades in the cases of (plain) 2-shot (w & w/o hist.). Again, most variants (11) are non-compliant, leading to the other special case – PX-4.

Case of C-2-shot w/ history: We find that **only three model variants are compliant with this probe and seem to marginally (i.e., not notable) utilize the additional contrastive history** along with the contrastive examples when compared to the case of C-0-shot (avg. boost: 0.6%↑; Orca 2 7B has max. boost of 1.2%↑), while 14 model variants clearly seem to get distracted with an average drop of 1.6%↓, leading to the paradox – PX-5.

(PX-3/4/5) Contrast can be confusing: Surprisingly, additional contrastive reader profile information doesn't enhance ICPL in many model variants. Eight out of 17 models show distraction and a 1.6% average drop when contrast is injected into the history component, struggling to map intra-concept associations within the history (**PX-3**). A real example of PX-3 can be seen in Figure 8. Moreover, 11 models exhibit a 3.6% average drop when provided with contrastive examples in a 2-shot setup without history (**PX-4**). This decline may be due to overfitting the format of the 2-shot setup, causing distraction. A real example of PX-4 can be seen in Figure 9. Additionally, 14 model variants fail to effectively utilize the richest prompt, C-2-shot with history, with seven unable to harness any form of contrastive profile information (**PX-3/4/5**). Three variants use contrastive history in C-0-shot but not contrastive examples in C-2-shot w/o hist., thus learning from broader historical topics rather than label instances (**PX-4/5**). Surprisingly, four variants can utilize contrastive histories in C-0-shot and contrastive examples in C-2-shot w/o hist., but not both, potentially due to spurious (and cross) linking between history and example concepts. A real example of PX-5 can be seen in Figure 10.

5.4 Effect of Article Length

In probe 1, we substitute longer articles with shorter ones, keeping the prompt length the same (see Table 2). Arguably, personalized summarization of shorter articles should be an easier task. However, several models exhibit the PX-1 and PX-1 (contrast) paradoxes (i.e., adding examples led to poor ICPL) under probe-1. Hence, we ***cannot generalize*** *that the length of the articles has influence*. Also, several models pass tests involving longer

Baseline Models	EGISES-JSD	LLM Models	EGISES-JSD	Paradoxes (PX) Observed
BigBird-Pegasus	**0.429**	**Tulu V2 DPO 7B** (0-shot / C-0-shot / 2-shot)	**0.325** / 0.338 / 0.345	PX-1/2/3/4/5
SimCLS	0.557	**Llama 2 7B Chat** (C-0-shot / C-2-shot / 2-shot)	0.333 / 0.338 / 0.345	PX-1/5
BRIO	0.661	**Llama 2 13B Chat** (C-0-shot)	0.341	PX-1/2/5
PENS-NAML T1	0.899	**Tulu V2 DPO 13B** (2-shot)	0.345	PX-2/3/4
PENS-NRMS T1	0.916	**Orca 2 13B** (C-0-shot)	0.347	PX-2/5

Table 5: **EGISES Leaderboard Misleads**: Top-5 LLMs as per EGISES-JSD, apparently beating top-5 baselines (top-2 personalized and top-3 non-personalized w/ summary cue), do not pass critical iCOPERNICUS tests exhibiting several paradoxes.

Models	PX-3	PX-4	PX-5
Llama 2 13B Chat	1.97% ↑	0.28% ↑	1.54% ↑
Mistral 7B Inst. v0.2	1.96% ↑	0.46% ↑	1.78% ↑
Tulu V2 DPO 7B	1.94% ↑	0.56% ↑	2.2% ↑

Table 6: **Adversarial Validation:** PX-3/4/5 exists. **%↑** value indicates how much the paradox worsens.

articles or history sequences, thereby reinforcing this finding.

5.5 EGISES Leaderboard Misleads

We observe from Table 4 that the top-5 best-performing LLMs in terms of EGISES-JSD as per Table 5[10] - i.e., Tulu v2 DPO 7B, Llama 2 7B Chat, Llama 2 13 B Chat, Tulu v2 DPO 7B, and Orca 2 13 B show several paradoxes with Tulu v2 DPO 7B being the worst. In fact, the best performing models (Orca 2 7B and Zephyr 7B β), passing all the iCOPERNICUS tests (see Table 4) do not rank high in Table 5. This empirically establishes that *EGISES as a standalone measure is* ***inadequate for ICPL evaluation*** *in LLMs*.

6 Validation of Paradoxes

The paradoxes raise a serious question: *do these models exhibit true ICPL?* (including the top-5 LLMs in Table 5) Hence, we need to first confirm that the paradoxes exist, which is done via adversarial probes. We select three model variants for three scenarios: (i) **worst case**: Tulu V2 DPO 7B showing all the paradoxes, (ii) **average case**: Llama 2 13B Chat showing non-contrastive (PX-1/2) and contrastive (PX-5) paradoxes, and (iii) **best case**: Mistral 7B Instruct v0.2 having PX-5 only.

Adversarial Probe-based Validation: In the adversarial probe setup, one of the user profiles (i.e., ground-truth history or examples) in the contrastive prompts (C-0-shot/C-2-shot w/o hist./C-2-shot w/ hist.) was replaced with a random history/examples. Since it is randomly sampled, the noise would be completely irrelevant to the document article to be summarized. Upon injection of such a noise, a model's personalization performance, and thereby the exhibited paradoxes, should degrade further. However, *if a model shows a better or similar EGISES score, then it would mean that the contrastive tests of iCOPERNICUS are no better than tests based on random choices*. In other words, iCOPERNICUS probes may not necessarily provide conclusive insights. However, we observe in Table 6 that the paradoxes do worsen (PX-3 avg. spike: 1.9%↑; PX-4 avg. spike: 0.43%↑; PX-5 avg. spike: 1.8%↑). **This validates the robustness of the iCOPERNICUS tests**. Details in Appendix E.1.

Human-Judgment Validation: We further validated PX-5 (the most serious paradox) using survey-based unbiased judgments (i.e., similarity ratings (1 (low) - 6 (very high)) on summary-pairs, both reference and model-generated from 339 respondents.[11] The core objective is to validate the extent to which human evaluators would *agree with the design principles of EGISES-JSD at a cognitive level* and thereby agree with the iCOPERNICUS test results that are based on EGISES-JSD. We, therefore, model a "*human version*" of EGISES-JSD (i.e., EGISES-HJ) and used the survey responses to estimate EGISES-HJ). We find that PX-5 persists except for Mistral 7B Instruct v0.2 (best case; ICPL boost: 1.42%↑; Table 9). We also find that PX-1 (contrastive) persists in the other models.[12]

7 Related Work

ICL for summarization: ICL capabilities in LLMs w.r.t summarization were first observed within reinforcement learning frameworks, utilizing human-feedback-trained reward models to predict human ratings for specific summary actions

[10] All the models are ***seemingly*** **better** than the strongest baseline model (BigBird-Pegasus), with 57%↑ for Tulu V2 DPO 7B when compared to PENS-NAML-T1.

[11] Grad student volunteers in 20-30 age group from Computer Sc., Maths, and Humanities; ∼70% male, ∼30% female. For details of survey methodology see Appendix E.2.

[12] Statistical Significance: p-value < 0.01.

(Stiennon et al., 2020; Nguyen et al., 2022). Since then, LLMs have shown unprecedented ICL-based summarization performance (Wang et al., 2023; Laskar et al., 2023; Tang et al., 2023). This opens the possibility of ICPL in these LLMs. At the same time, it also underscores the necessity for robust and dependable methods of evaluating the degree of ICPL within such models. Benchmarking LLMs for summarization was done for accuracy, fluency, and consistency (Zhang et al., 2023), but not ICPL.

Personalization evaluation: Personalization evaluation is studied in recommendation systems (Zangerle and Bauer, 2022), with metrics based on the Jaccard Index, MAE/RMSE/Hit-Ratio (Li et al., 2024), and nDCG (normalized Discounted Cumulative Gain) (Matthijs and Radlinski, 2011). However, these metrics are not useful for summarization. The only personalization evaluation metric for summarizers is EGISES (Vansh et al., 2023). However, as established, relying solely on EGISES to evaluate ICPL can be misleading.

8 Conclusion

We propose iCOPERNICUS, a novel evaluation framework for analyzing the capability of true In-Context Personalization Learning (ICPL) in LLMs for the personalized summarization task. The central goal is to detect whether models can pass all the probes (or exhibit the "*paradox of less is more*"). We showed that relying solely on EGISES-scores can be misleading (as in top-4 LLMs beating baselines). Only 2 out of the 17 SOTA LLM variants probed passed the iCOPERNICUS test, hence the need for further research on ICPL-driven LLMs.

Limitations

In this work, we restrict probing of In-Context Personalization Learning (ICPL) w.r.t summarization to 7B and 13B model variants. It would be interesting to observe how smaller SLMs ($<$ 7B), such as the Phi model suite (Gunasekar et al., 2024), fare the iCOPERNICUS test. Also, studies around the effect of more recent fine-tuning techniques such as PEFT (LoRA and QLoRA) need to be analyzed. On similar lines, it would be interesting to observe whether ICPL, which otherwise is not emerging by doubling the models' size, finally starts emerging at an even larger scale (which often is the case for several emerging properties). At the same time, robust and reliable ICPL measures should be designed for aggregated leaderboard generation of models w.r.t ICPL within the iCOPERNICUS framework. Finally, more robust and systematic adversarial probing of ICPL is required to analyze true ICPL in models.

Ethics Statement

We would like to declare that we used the PENS dataset prepared and released by Microsoft Research. Our human-judgment survey was conducted according to the norms set by the Institutional Review Board (IRB) and respects participant anonymity as per guidelines.

Acknowledgements

This research has been supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

References

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92. Association for Computational Linguistics.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *ArXiv*, abs/2108.07732.

Ayaka. 2023. Jax implementation of the llama 2 model.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Abdul Ghafoor Etemad, Ali Imam Abidi, and Megha Chhabra. 2021. Fine-tuned t5 for abstractive summarization. *International Journal of Performability Engineering*, 17(10).

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Samira Ghodratnama, Mehrdad Zakershahrak, and Fariborz Sobhanmanesh. 2021. Adaptive summaries: A personalized concept-based summarization approach by learning from users' feedback. In *Service-Oriented Computing – ICSOC 2020 Workshops*, pages 281–293, Cham. Springer International Publishing.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Conti Kauffmann, Gustavo Henrique de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Behl, Xin Wang, Sebastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2024. Textbooks are all you need.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv e-prints*, pages arXiv–2310.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Shu Li, Yuan Zhao, Longjiang Guo, Meirui Ren, Jin Li, Lichen Zhang, and Keqin Li. 2024. Quantification and prediction of engagement: Applied to personalized course recommendation to reduce dropout in moocs. *Information Processing & Management*, 61(1):103536.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. 2023. Stable Beluga models. `huggingface.co/stabilityai/StableBeluga2`. [Online; accessed 1-December-2023].

Nicolaas Matthijs and Filip Radlinski. 2011. Personalizing web search using long term browsing history. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 25–34.

M.L. Menéndez, J.A. Pardo, L. Pardo, and M.C. Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes Ribeiro, Sahaj Agrawal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca-2: Teaching small language models how to reason. arXiv.

Duy-Hung Nguyen, Nguyen Viet Dung Nghiem, Bao-Sinh Nguyen, Dung Tien Tien Le, Shahab Sabahi,

Minh-Tien Nguyen, and Hung Le. 2022. Make the most of prior data: A solution for interactive text summarization with preference feedback. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1919–1930, Seattle, United States. Association for Computational Linguistics.

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1933–1942, New York, NY, USA. Association for Computing Machinery.

Mikhail Orzhenovskii. 2021. T5-long-extract at fns-2021 shared task. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 67–69.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410. Association for Computational Linguistics.

GS Ramesh, Vamsi Manyam, Vijoosh Mandula, Pavan Myana, Sathvika Macha, and Suprith Reddy. 2022. Abstractive text summarization using t5 architecture. In *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2021*, pages 535–543. Springer.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. An adversarial winograd schema challenge at scale.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

T Tawmo, Mrinmoi Bohra, Pankaj Dadure, Partha Pakray, et al. 2022. Comparative analysis of t5 model for abstractive text summarization on different datasets. In *Proceedings of the International Conference on Innovative Computing Communication (ICICC) 2022*. SSRN.

Maartje Ter Hoeve, Julia Kiseleva, and Maarten Rijke. 2022. What makes a good and useful summary? Incorporating users in automatic summarization research. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 46–75, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv e-prints*, pages arXiv–2307.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv e-prints*, pages arXiv–2310.

Rahul Vansh, Darsh Rank, Sourish Dasgupta, and Tanmoy Chakraborty. 2023. Accuracy is not enough: Evaluating personalization in summarizers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2582–2595, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv e-prints*, pages arXiv–2303.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

International Joint Conferences on Artificial Intelligence Organization.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019b. Neural news recommendation with multi-head self-attention. 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 6389–6394, Hong Kong, China. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

Eva Zangerle and Christine Bauer. 2022. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8).

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied Sanosi Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *ArXiv*, abs/2304.06364.

A Preliminaries

A.1 Degree-of-personalization

In this section, we recall the notion of *insensitivity-to-subjectivity* of a personalized summarization model ($M_{\boldsymbol{\theta},u} : (d, u) \mapsto s_u$; where s_u is the personalized summary for reader u on document d) as defined in Vansh et al. (2023).

Definition 5. ***Weak Insensitivity-to-Subjectivity.*** *A summarization model $M_{\boldsymbol{\theta},u}$ is (weakly) Insensitive-to-Subjectivity w.r.t reader u, if $\forall(u_i, u_j)$, $(\sigma(u_i, u_j) \leq \tau^U_{max}) \iff (\sigma(s_{u_i}, s_{u_j}) > \tau^S_{max})$, where σ[13] is an arbitrary distance metric defined on the metric space $\mathcal{M}$ where d, u, s are defined, τ^U_{max} is the maximum limit for u_i, u_j to be mutually indistinguishable, and τ^S_{max} is the maximum limit for s_{u_i}, s_{u_j} to be mutually indistinguishable.*

Definition 6. ***Strong Insensitivity-to-Subjectivity.*** *A summarization model $M_{\boldsymbol{\theta},u}$ is (strongly) Insensitive-to-Subjectivity w.r.t reader u if $\forall(u_i, u_j)$, $M_{\boldsymbol{\theta},u}$ satisfies: (i) the condition of weak insensitivity, and (ii) $(\sigma(u_i, u_j) > \tau^U_{max}) \iff (\sigma(s_{u_i}, s_{u_j}) \leq \tau^S_{max})$.*

A.2 EGISES and Degree-of-Personalization

We generalize the definition of summary-level deviation (or, ***Deg**ree-of-**R**esponsiven**ess*** (DEGRESS))[14] proposed by Vansh et al. (2023) as follows:

Definition 7. ***Summary-level* DEGRESS.** *Given a document d_i and a user-profile u_{ij} (user j's expected summary), the summary-level responsiveness of a personalized model $M_{\boldsymbol{\theta},u}$, (i.e., $\texttt{DEGRESS}(s_{u_{ij}}|(d_i, u_{ij}))$), is defined as the proportional divergence between model-generated summary $s_{u_{ij}}$ of d_i for j-th user from other user-specific summary versions w.r.t a corresponding divergence of u_{ij} from the other user-profiles.*

[13] $\sigma(u_i, u_i) = 0$; $\sigma(u_i, u_j) \in [0, 1]$; σ satisfies positivity, reflexive, maximality, symmetry, and the triangle inequality.

[14] This is based on the notion of weak and strong *insensitivity-to-subjectivity*, as defined by (Vansh et al., 2023) (see Appendix A.1).

$\texttt{DEGRESS}(s_{u_{ij}}|(d_i, u_{ij}))$ is formulated as:

$$\texttt{DEGRESS}(s_{u_{ij}}|(d_i, u_{ij})) = \frac{1}{|\mathbf{U}_{d_i}|} \sum_{k=1}^{|\mathbf{U}_{d_i}|} \frac{min(X_{ijk}, Y_{ijk}) + \epsilon}{max(X_{ijk}, Y_{ijk}) + \epsilon}$$

$$X_{ijk} = \frac{\exp(w(u_{ij}|u_{ik}))}{\sum_{l=1}^{|\mathbf{U}_{d_i}|} \exp(w(u_{ij}|u_{il}))} \cdot \sigma(u_{ij}, u_{ik})$$

$$Y_{ijk} = \frac{\exp(w(s_{u_{ij}}|s_{u_{ik}}))}{\sum_{l=1}^{|\mathbf{U}_{d_i}|} \exp(w(s_{u_{ij}}|s_{u_{il}}))} \cdot \sigma(s_{u_{ij}}, s_{u_{ik}})$$

$$w(u_{ij}|u_{ik}) = \frac{\sigma(u_{ij}, u_{ik})}{\sigma(u_{ij}, d_i)}; \; w(s_{u_{ij}}|s_{u_{ik}}) = \frac{\sigma(s_{u_{ij}}, s_{u_{ik}})}{\sigma(s_{u_{ij}}, d_i)} \quad (1)$$

Here, $|\mathbf{D}|$ is the total number of documents in the evaluation dataset, $|\mathbf{U}|$ is the total number of users who created gold-reference summaries that reflect their expected summaries (and thereby their subjective preferences or profiles), and $|\mathbf{U}_{d_i}|$ $(= |\mathbf{S}_{d_i}|)$ is the number of users who created gold-references for document d_i. A lower value of $\texttt{DEGRESS}(s_{u_{ij}}|(d_i, u_{ij}))$ indicates that while reader-profiles are different, the generated summary $s_{u_{ij}}$ is very similar to other reader-specific summaries (or vice versa), and hence, is not responsive at the summary-level. The system-level DEGRESS and EGISES have been formulated as follows:

$$\texttt{DEGRESS}(M_{\boldsymbol{\theta},u}) = \frac{\sum_{i=1}^{|\mathbf{D}|} \frac{\sum_{j=1}^{|\mathbf{U}_{d_i}|} \texttt{DEGRESS}(s_{u_{ij}}|(d_i, u_{ij}))}{|\mathbf{U}_{d_i}|}}{|\mathbf{D}|} \quad (2)$$

$$\texttt{EGISES}(M_{\boldsymbol{\theta},u}) = 1 - \texttt{DEGRESS}(M_{\boldsymbol{\theta},u}) \quad (3)$$

EGISES measures the degree of insensitivity-to-subjectivity for relative benchmarking of how much models lack personalization (i.e., a lower score is better within the range: $[0, 1]$) instead of assigning an absolute goodness score. In this paper, we choose Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997), where d, u, and s_u are defined as word distributions on a probability space. JSD has a strong human-judgment correlation and has been used in evaluating the ten specialized baseline models (Vansh et al., 2023).

A.3 In-Context Learning

ICL is a method employed by LLMs, notably emphasized in Brown et al. (2020) (GPT-3's ICL behavior was first highlighted), where models acquire proficiency in apparently unknown tasks (i.e., tasks on which the models are not pre-trained) from limited examples, called *prompts*, with no update in their parameters (i.e., the models are frozen). Formally, it is defined below.

Definition 8. ***Prompt****: A prompt $\mathcal{P}$ given to a language model M is a sequence of n concatenated ($\oplus$) demonstration examples (i.e., input-label pairs: $\left(x'_i \oplus y'_{x_i}\right)$) as $\langle\left(x'_1 \oplus y'_{x_1}\right) \oplus \left(x'_2 \oplus y'_{x_2}\right) \oplus \ldots \oplus \left(x'_n \oplus y'_{x_n}\right)\rangle$ and an input query x_q appended, such that $x_i \neq x_q$.*

The input often includes a description of the task or a system command $\mathcal{T}$ before the demonstration sequence $\mathcal{D}_\mathcal{P}$. We now provide a formal definition of ICL as follows:

Definition 9. ***In-Context Learning (ICL):*** *A model M is said to exhibit ICL if given a prompt $\mathcal{P} \sim \mathbb{D}$ (where $\mathbb{D}$ is an unseen demonstration dataset) and an unseen task $\mathcal{T}$, $M : (\mathcal{T} \oplus \mathcal{D}_\mathcal{P} \oplus x_q) \mapsto y^*_{x_q}$; $y^*_{x_q} \in Y^*_{x_q}$; where $Y^*_{x_q}$ is the expected set of output labels for the given query x_q.*

M predicts (i.e., maps) using the prompt's conditioning only, requiring it to discern essential aspects such as input-label mapping, input text distribution, label space, and formatting (lexico-syntactic structural relationship between the prompt components).

A.4 ICL is latent concept mapping

Xie et al. (2021) suggested that LLMs acquire latent document-level concepts during pretraining to generate coherent subsequent tokens. ICL occurs when LLMs identify shared latent concepts among prompt examples. Min et al. (2022) revealed that input-label mapping, input-text distribution, label space, and format in prompts matter more. Wei et al. (2023) provided supportive evidence that label association learning becomes more pronounced with larger model sizes. Although further analysis needs to be done, we consider the latent-concept association hypothesis the most compelling explanation of ICL so far and use this as a primary tool for understanding our own findings on ICPL in this paper. In the following section, we first define ICPL (w.r.t personalized summarization) as a special case of ICL and then propose iCOPERNICUS as a framework for evaluating ICPL.

B Model Descriptions

B.1 LLM Model Descriptions

We provide concise descriptions of the LLMs analyzed in this study for understanding ICPL.

1. **Llama 2** - Llama 2 (Touvron et al., 2023) is a family of transformer-based autoregressive

causal language models, ranging in scale from 7 billion to 70 billion parameters. Llama 2 models are trained on 2 trillion tokens and have double the context length of Llama 1.

2. **Llama 2 Chat** - Llama 2 Chat (Touvron et al., 2023) is a fine-tuned version of Llama 2, optimized for dialogue applications using reinforcement learning from human feedback (RLHF). Llama 2 Chat models demonstrate improved helpfulness and safety compared to other open models and achieve comparable performance to ChatGPT according to human evaluations.

3. **Mistral 7B** - Mistral 7B (Jiang et al., 2023) is a language model that outperforms Llama 2 13B and Llama 1 34B in various tasks, such as natural language inference, mathematics, and code generation. It leverages grouped-query attention (GQA), and sliding window attention (SWA). GQA significantly accelerates the inference speed and also reduces the memory requirement during decoding.

4. **Mistral 7B Instruct** - Mistral 7B Instruct (Jiang et al., 2023) is a fine-tuned version of Mistral 7B that leverages instruction datasets to enhance its generalization and adaptation capabilities. The model has two versions: v0.1 and v0.2. It exhibits superior performance compared to all 7B models on MT-Bench and is comparable to 13B – Chat models. Mistral 7B Instruct v0.2 is an updated version with improvements in instruction following and generalization capabilities.

5. **Tulu V2 Suite** - The Tulu V2 Suite (Ivison et al., 2023) is a collection of fine-tuned large language models (LLMs) based on Llama 2. The models in this suite are fine-tuned on a mix of publicly available, synthetic, and human datasets. The suite includes Tulu V2 models as well as the DPO fine-tuned Tulu V2 DPO models.

6. **Orca 2** - Orca 2 (Mitra et al., 2023) is a suite of models that are fine-tuned on Llama 2 using synthetic dataset. Orca models are designed to enhance the reasoning abilities of smaller language models by imitating the step-by-step reasoning traces of more capable LLMs. Orca 2 models surpass models of similar size and attain performance levels similar to or better than models five times larger on complex reasoning tasks.

7. **Stable Beluga** - Stable Beluga (Mahan et al., 2023) is a collection of models that have been fine-tuned on the Llama 2 using an internal Orca style dataset. The primary objective of these models is to generate responses that are not only responsive to user prompts and queries, but also emphasize reasoning and helpfulness.

8. **Zephyr 7B** - Zephyr 7B (Tunstall et al., 2023) is a series of language models trained to act as helpful assistants developed by the HuggingFace H4 team. This includes Zephyr 7B α and Zephyr 7B β. It is a fine-tuned on Mistral 7B v0.1 and it was trained on on a mix of publicly available, synthetic datasets using DPO.

B.2 Baseline Model Descriptions

We briefly introduce the SOTA baseline summarization models that were analyzed to understand their degree-of-personalization below:

1. **PENS-NRMS Injection-Type 1**: The PENS framework (Ao et al., 2021) utilizes NRMS (Neural News Recommendation with Multi-Head Self-Attention) (Wu et al., 2019b) for personalized summary generation. NRMS employs a news encoder using multi-head self-attention to understand news titles and learn user representations based on browsing history. Additive attention enhances learning by selecting important words and articles. In Injection-Type 1, NRMS user embedding is injected into PENS by initializing the decoder's hidden state of the headline generator.

2. **PENS-NRMS Injection-Type 2**: To generate a personalized summary, NRMS user embedding is injected into attention values (Injection-Type 2) of PENS that helps to personalize attentive values of words in the news body.

3. **PENS-NAML Injection-Type 1**: NAML (Neural News Recommendation with Attentive Multi-View Learning) (Wu et al., 2019a) incorporates a news encoder employing a multi-view attention model for comprehensive news representations. The user encoder

learns user representations based on interactions with browsed news, allowing the selection of informative news. In Injection-Type 1, this user embedding is injected into the PENS model for personalization.

4. **PENS-EBNR Injection-Type 1**: EBNR (Embedding-based News Recommendation for Millions of Users) (Okura et al., 2017) proposes a method for user representations by using an RNN model that takes browsing histories as input sequences. This user embedding is injected using Type 1 into the PENS model for personalization.

5. **PENS-EBNR Injection-Type 2**: This personalized model injects EBNR user embedding into PENS using type-2.

6. **BRIO**: BRIO (Liu et al., 2022) assumes a non-deterministic training paradigm that assigns probability mass to different candidate summaries according to their quality, thereby helping it to better distinguish between high-quality and low-quality summaries.

7. **SimCLS**: SimCLS (A Simple Framework for Contrastive Learning of Abstractive Summarization) (Liu and Liu, 2021) uses a two-stage training procedure. In the first stage, a Seq2Seq model (BART (Lewis et al., 2020)) is trained to generate candidate summaries with MLE loss. Then, a RoBERTa-initiated evaluation model is trained to rank these using contrastive learning.

8. **BigBird-Pegasus**: BigBird (Zaheer et al., 2020) is an extension of Transformer based models designed specifically for processing longer sequences. It utilizes sparse, global, and random attention mechanisms to approximate full attention which enables it to handle longer contexts more efficiently.

9. **ProphetNet**: ProphetNet (Qi et al., 2020) is a seq2seq pre-trained model that employs n-gram prediction using the n-stream self-attention mechanism. It enhances n-step ahead prediction by predicting the next n tokens at once, based on previous tokens, thus avoiding overfitting on local correlations.

10. **T5**: T5 (Text-To-Text Transfer Transformer) is based on the Transformer Encoder-Decoder architecture that operates on the principle of the unified text-to-text task for any NLP problem, including summarization. See (Tawmo et al., 2022; Ramesh et al., 2022; Etemad et al., 2021) for recent T5 summarization analyses.

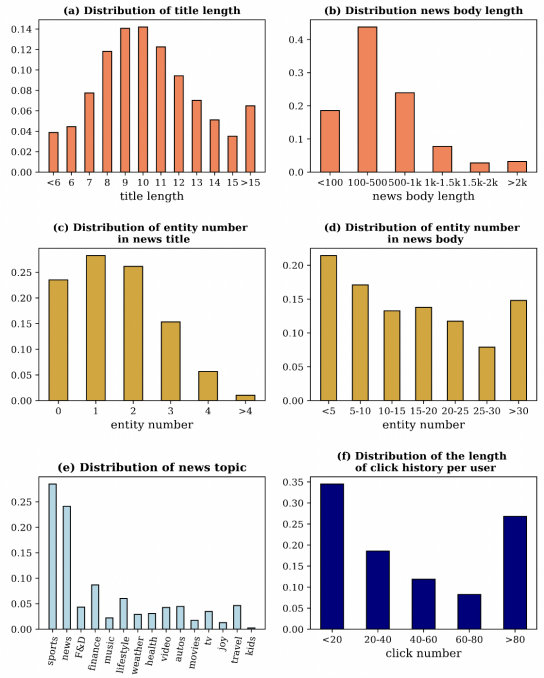


Figure 2: The statistics of news corpus and training set of the PENS dataset.

C PENS Dataset

The PENS dataset is a comprehensive collection of 113,762 news articles, each of which is categorized into one of 15 distinct topics. Each article in the dataset includes a unique news ID, a title, a body, and a category that has been manually tagged by editors. The average length of a news title is 10.5 words, while the average length of a news body is 549.0 words (Refer to Figure 2 for statistics of news articles). Entities from each news title are extracted and subsequently linked to corresponding entities in WikiData.

For the purpose of training, 500,000 user-news impressions were sampled from June 13, 2019, to July 3, 2019. An impression log records the news articles displayed to a user and the user's click behaviors on these articles during a specific visit to the news website. Each labeled sample in the training set follows the format [uID, tmp, clkNews, uclkNews, clkedHis], where 'uID' represents the anonymous ID of a user, 'tmp' denotes the timestamp of the impression record, 'clkNews' and 'uclkNews' are the clicked and un-clicked news in

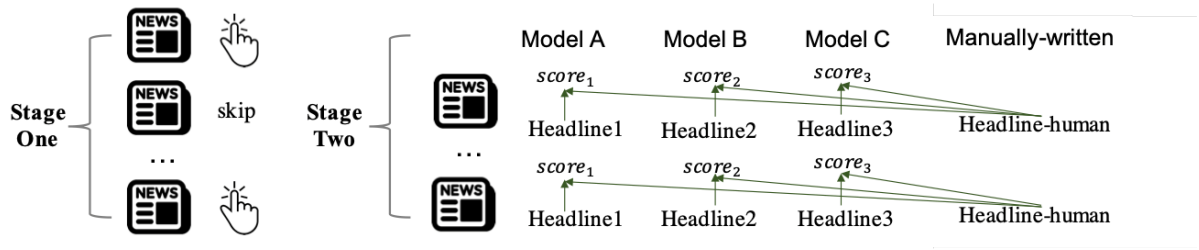


Figure 3: Stages of creation of testing dataset consisting of personalized headlines

Column	Example Context	Description
userid	NT1	The unique ID of 103 users
clicknewsID	N108480, N38238, N35068, ...	The user's historical clicked news collected at the first stage
posnewID	N24110, N62769, N36186, ...	The exhibited news for each user at the second stage
rewrite_titles	'Legal battle looms over Trump EPA's rule change of Obama's Clean Power Plan rule ...'	The manually-written news headlines for the exhibited news articles and can be split by '#TAB#'

Table 7: Dataset Format of data collected from different users consisting of the articles clicked by the user and details of articles for which personalized headlines created by the user

the impression, respectively, and 'clkedHis' represents the news articles previously clicked by the user. All samples in 'clkNews', 'uclkNews', and 'clkedHis' are sorted by the user's click time.

C.1 Test Set Construction Process

In order to establish an offline testbed, 103 English native speakers, all of whom are college students, were invited to manually create a test set in two stages as represented in Figure 3.

In the first stage, each participant browses 1,000 news headlines and marks at least 50 pieces that they find interesting. These selected news articles were randomly chosen from the news corpus and were arranged according to their first exposure time.

In the second stage, participants are asked to write down their preferred headlines for another 200 unseen news articles from the dataset without being shown the original news titles. They are also asked to highlight important segments in the original news articles. These unseen news articles are evenly sampled, and they are redundantly assigned to ensure that each news article is reviewed by an average of four people. The quality of these manually-written headlines is checked by professional editors from the perspective of the factual aspect of the media frame. Headlines that are of low quality, such as those containing incorrect factual information, those inconsistent with the news body, or those that are too short or too long, are excluded. The remaining headlines are considered to be the personalized reading focuses of the annotators on the articles, and are taken as gold-standard headlines in the PENS dataset.

D Prompt Design Principles

We use six different types of prompting styles, each tailored to provide varying levels of context and personalization to the LLMs. We present the structure of each prompt style in Figure 4, along with the composition of tokens in the prompts as described in Table 2, to provide a comprehensive understanding of the composition of prompting techniques employed in our study. If the length of any portion of the prompt is greater than the limit, it's truncated to fit it in the context length. For prompt examples refer to Figures 6-10 in Appendix [15]

1. **Zero-shot**: In this approach, we provide the user's click history followed by the target article for which we aim to generate a personalized headline.

2. **Contrastive zero-shot**: This approach provides the click history of two users (User1 and

[15]News article statistics from the sample of 3840 news articles used in this study – The mean length was 659.91 tokens, the median was 493 tokens, the 90th percentile was 1180 tokens, and the 95th percentile was 1659 tokens.

User2), followed by the target article clicked by both users.

3. **2-shot w/o history**: In this approach, we provide two example articles and their corresponding headlines given by the user, followed by the target article for which we aim to generate a personalized headline.

4. **Contrastive 2-shot w/o history**: This method is similar to the 2-shot approach but involves providing two example articles and their corresponding headlines given by two users (User1 and User2).

5. **2-shot w/ history**: In this approach, we provide the user's click history, two examples of articles clicked by the user and their corresponding summaries, along with the article body that needs to be summarized.

6. **Contrastive 2-shot w/ history**: This method involves providing the click history for both users, two examples of common articles clicked by both users and their corresponding summaries, along with the article body that needs to be summarized.

In the study, a substantial number of prompts were employed for probing, as indicated in Table 2. The generated headlines were extracted from the produced text using simple regular expression matching. The output format was explicitly demonstrated in the examples for 2-shot prompts, and an example format was provided for 0-shot prompts.

E Analysis of Models w.r.t iCOPERNICUS

E.1 Adversarial Testing: Results

Adversarial testing was conducted for contrastive prompts across three distinct types of Large Language Models (LLMs): Mistral 7B Inst. v0.2 which exhibits a single paradox; Llama 2 13B Chat that displays a moderate number of paradoxes; and Tulu V2 DPO 7B which presents all paradoxes as shown in Table 4. The details of User-2, such as their click history and the headlines they wrote in the contrastive prompt, are not accurate. Instead, the reading history of another user and the headline of a random article written by a random user were used. Table 6 verifies the existence of the paradoxes, as these models show higher perplexity for three different types of contrastive prompts: PX-3, which is tested by the contrastive zero-shot prompt; PX-4, which is tested by the contrastive two-shot prompt without history; and PX-5, which is tested by the contrastive two-shot prompt with history. These results indicate that these models are sensitive to the quality and relevance of the information provided in the prompts, and that they perform worse when the prompts contain incorrect details about the user's reading history or the headlines written by them in the examples. Intriguingly, these three models exhibit a significant performance decline when the prompts include user's reading history. However, they only show a minor performance drop for contrastive two-shot prompts without history. This suggests that these models are capable of discerning that an incorrect/irrelevant headline is given by User-2 in the examples.

Models	C-0-shot	C-2-shot w/o hist	C-2-shot w/ hist
Llama 2 13B Chat	q_1	q_2	q_3
Mistral 7B Inst. v0.2	q_4	q_5	q_6
Tulu V2 DPO 7B	q_7	q_8	q_9

Table 8: **(Survey) Questionnaire Structure:** A respondent fills up the survey for document d_i in the sequence: $\langle u_{ij}, u_{ik} \rangle \rightarrow q_{i1} \rightarrow q_{i2} \rightarrow \ldots \rightarrow q_{i9}$, where $q_{i\bullet}$: $\langle s_{u_{ij}}, s_{u_{ik}} \rangle^{(M_{\boldsymbol{\theta},u}, \mathcal{P}_C)_\bullet}$; s_u is the summary generated by each of the 3 models for a specific prompt type (i.e., the model-contrastive prompt-type pair $(M_{\boldsymbol{\theta},u}, \mathcal{P}_C)_\bullet$).

E.2 Human Judgment Validation: Results

Survey Structure: Human judgment-based validation was conducted on a set of contrastive prompts, systematically evaluating three distinct LLMs: (i) Mistral 7B Instruct v0.2 (exhibiting only PX-5), (ii) Llama 2 13B Chat (exhibiting a moderate number of paradoxes - PX-1/2/5), and (iii) Tulu V2 DPO 7B (exhibiting all the paradoxes), as outlined in Table 4. We have randomly sampled multiple documents and three corresponding users (i.e., readers) who have generated summaries for those documents eg.$(d_i, (u_{i1}, u_{i2}, u_{i3}))$ from the PENS dataset. For each $(d_i, (u_{i1}, u_{i2}, u_{i3}))$ there are 3 combinations possible $((d_i, (u_{i1}, u_{i2}))$, $(d_i, (u_{i2}, u_{i3}))$, $(d_i, (u_{i3}, u_{i1})))$. Each survey respondant is shown a set of 10 questions $(\langle u_{ij}, u_{ik} \rangle, q_{i1}, q_{i2} \ldots q_{i9})$ for a given combination corresponding to a document d_i eg.$(d_i, (u_{ij}, u_{ik}))$ as shown in the Table 8; where $q_{i\bullet} \in \{q_{i1}, q_{i2} \ldots q_{i9}\}$ contains a pair of the corresponding prompts and model generated summaries $(\langle s_{u_{ij}}, s_{u_{ik}} \rangle^{(M_{\boldsymbol{\theta},u}, \mathcal{P}_C)_\bullet})$ of user-pairs and $\langle u_{ij}, u_{ik} \rangle$ contains a pair of user-generated reference summaries of user-pairs.

Models	PX-5	PX-6
Llama 2 13B Chat	✓	✓
Mistral 7B Inst. v0.2	✗	✓
Tulu V2 DPO 7B	✓	✓

Table 9: **Human-Judgment Validation:** PX-5 exists for worst and medium case; Contrastive PX-1 (PX-6) is also confirmed (Human-agreement denoted as ✓).

Survey details: We designed and conducted an online survey to gather insights from participants on a voluntary basis (see Figure 5 in Appendix). A total of 339 responses were obtained, encompassing evaluations of 113 documents from the PENS evaluation dataset. The respondent pool comprised 262 males and 77 females. Participants represented diverse educational backgrounds, including undergraduate and graduate students specializing in computer science, mathematical science, electronic engineering, and humanities. To maintain objectivity, participants were not informed that the questions presented were summary-pairs. Instead, each summary pair was displayed as regular text, prompting participants to rate their similarity on a scale ranging from 1 (low) to 6 (very high). This methodology ensured unbiased judgments regarding the proximity of subjective user reference summaries and their corresponding model-generated summaries.

Computing EGISES-HJ-JSD: The similarity scores provided by users were utilized as the basis for computing similarity scores for summary-summary pairs. For summary-document distance, we used JSD (since it is not viable for a user to read the whole document and assign scores to a summary). Using these distances (summary-summary distances based on user ratings, summary-document distances using JSD), we then calculated EGISES-HJ-JSD for all three models and prompt types. We observed that EGISES-HJ-JSD scores based on human judgment showed agreement with **2/3** models for PX-5 and **3/3** models for PX-6 (the contrastive version of PX-1) as shown in Table 9.

F Inference Setup and Configuration

For the inference process in our study, we utilized the JAX library. Our implementation heavily relied on the repository "llama-2-jax" (Ayaka, 2023). In order to generate outputs from our models, we configured the inference process to ensure optimal performance and output quality.

F.1 LLM Settings

We used the following settings to control the behavior of the LLMs during inference:

- **Temperature:** We set the temperature to a value of 0.6, striking a balance between predictability and diversity.
- **Sampling Method:** We adopted the top-k sampling method with k = 16 for generating outputs.
- **Maximum Length:** Maximum length for inference was set to the 4096 tokens.
- **Inference Precision:** We used bfloat16 precision for inference, consistent with the precision in which the model weights were originally published.

F.2 Compute Platform

For high-performance inference, we utilized TPU v3-8/v4-8 VMs through Google Cloud Platform.

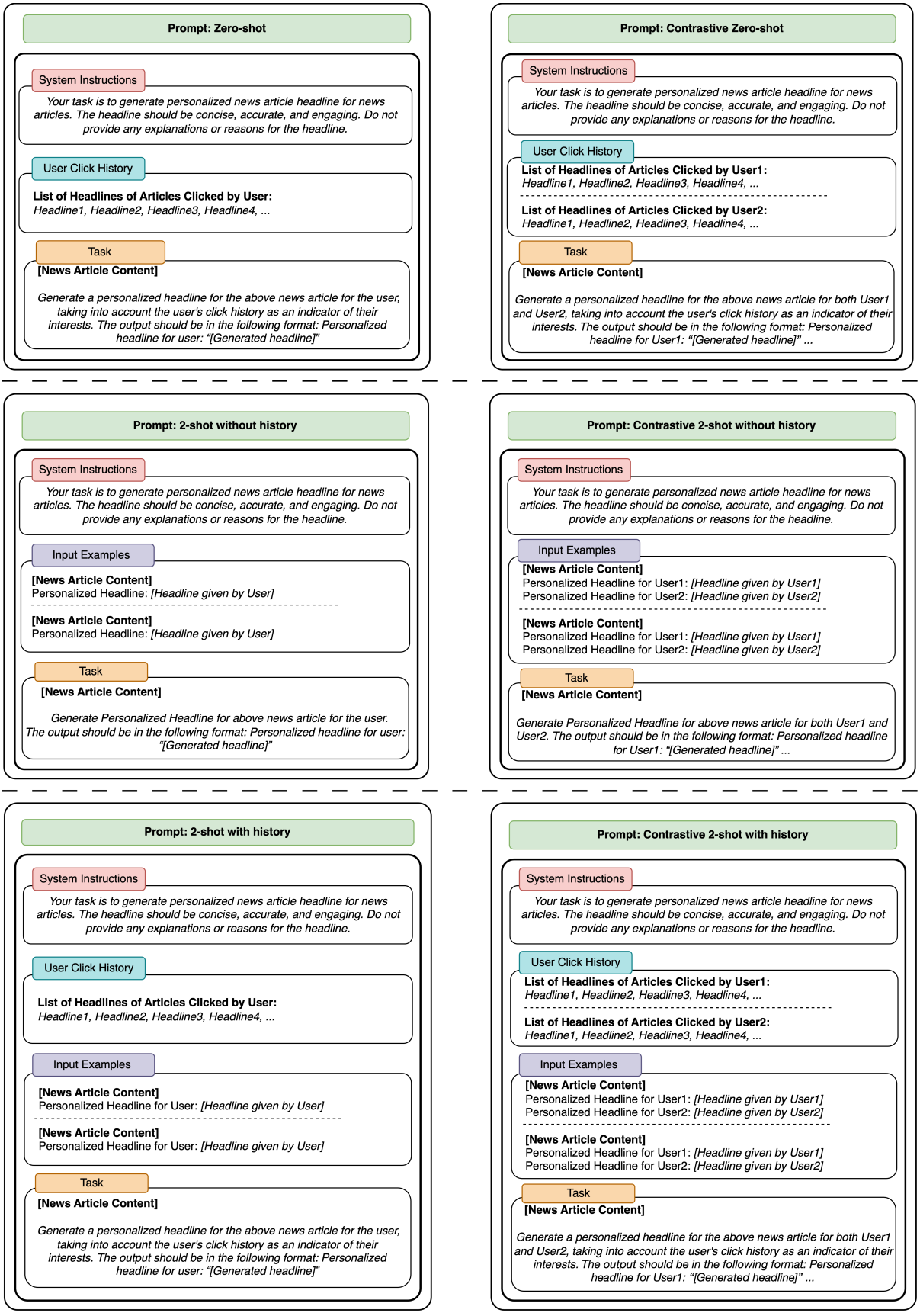
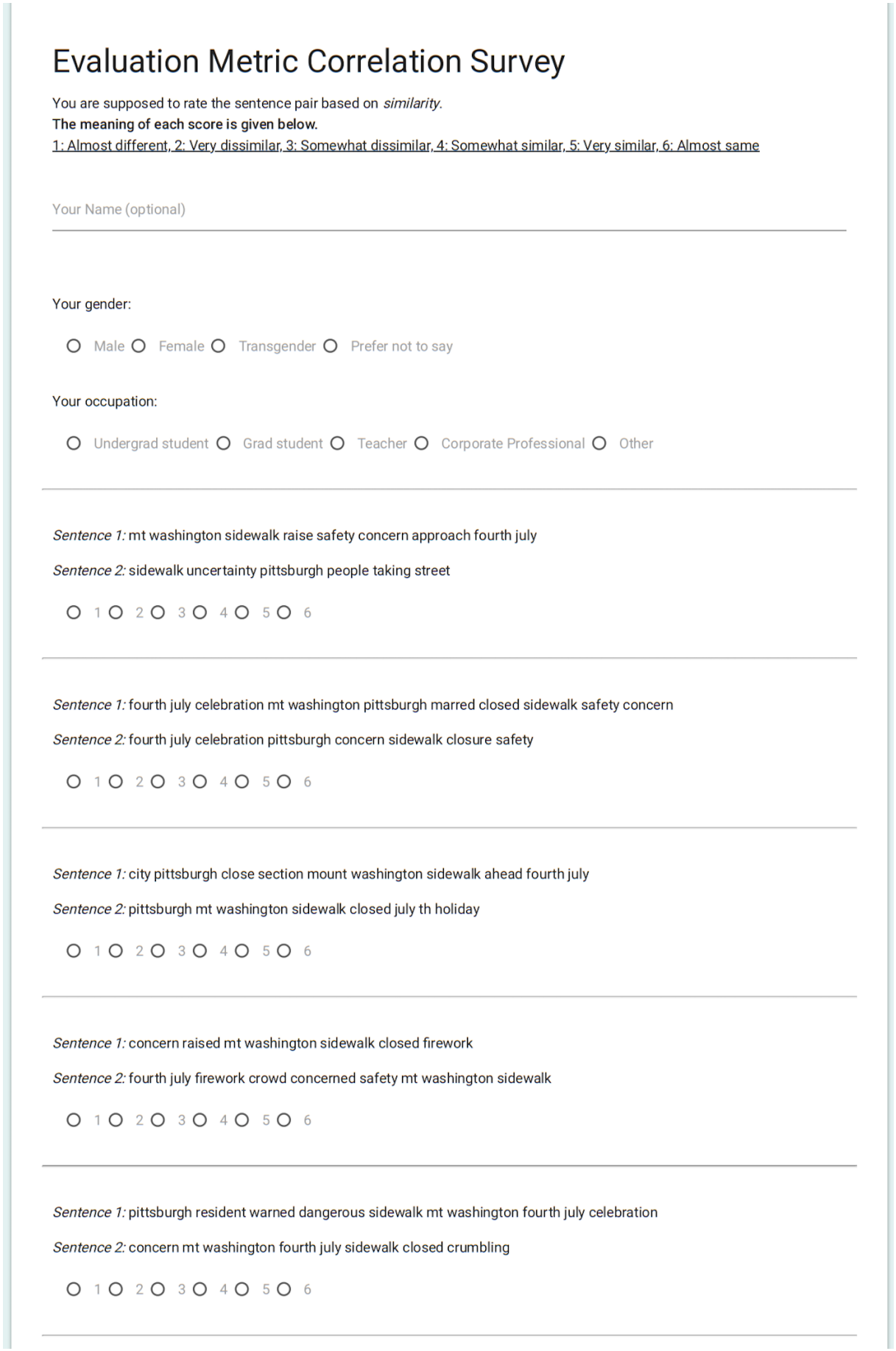


Figure 4: **Prompt Templates within the iCOPERNICUS framework**: Prompts on the left probe whether models utilize richer reader profiles; prompts on the right probe whether models utilize contrastive information for real personalization.



Evaluation Metric Correlation Survey

You are supposed to rate the sentence pair based on *similarity*.
The meaning of each score is given below.
1: Almost different, 2: Very dissimilar, 3: Somewhat dissimilar, 4: Somewhat similar, 5: Very similar, 6: Almost same

Your Name (optional)

Your gender:

Male Female Transgender Prefer not to say

Your occupation:

Undergrad student Grad student Teacher Corporate Professional Other

Sentence 1: mt washington sidewalk raise safety concern approach fourth july

Sentence 2: sidewalk uncertainty pittsburgh people taking street

1 2 3 4 5 6

Sentence 1: fourth july celebration mt washington pittsburgh marred closed sidewalk safety concern

Sentence 2: fourth july celebration pittsburgh concern sidewalk closure safety

1 2 3 4 5 6

Sentence 1: city pittsburgh close section mount washington sidewalk ahead fourth july

Sentence 2: pittsburgh mt washington sidewalk closed july th holiday

1 2 3 4 5 6

Sentence 1: concern raised mt washington sidewalk closed firework

Sentence 2: fourth july firework crowd concerned safety mt washington sidewalk

1 2 3 4 5 6

Sentence 1: pittsburgh resident warned dangerous sidewalk mt washington fourth july celebration

Sentence 2: concern mt washington fourth july sidewalk closed crumbling

1 2 3 4 5 6

Figure 5: Website portal designed for conducting survey for collecting human judgements on the similarity between user references and model generated summaries for contrastive prompts.

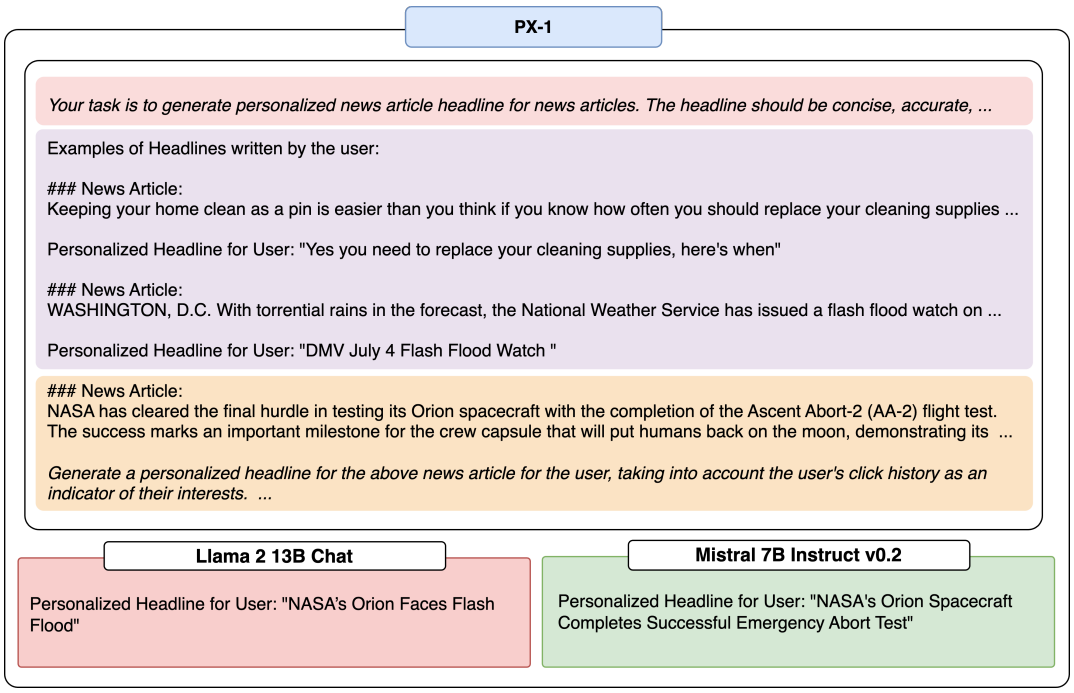


Figure 6: **Illustration of PX-1** (*effect of personalized examples*): The left column shows the output of Llama 2 13B Chat, which hallucinates generating irrelevant information (marked in red) due to the distraction caused by the examples; the right column shows the output of Mistral 7B Instruct v0.2, which generates the expected response (marked in green).

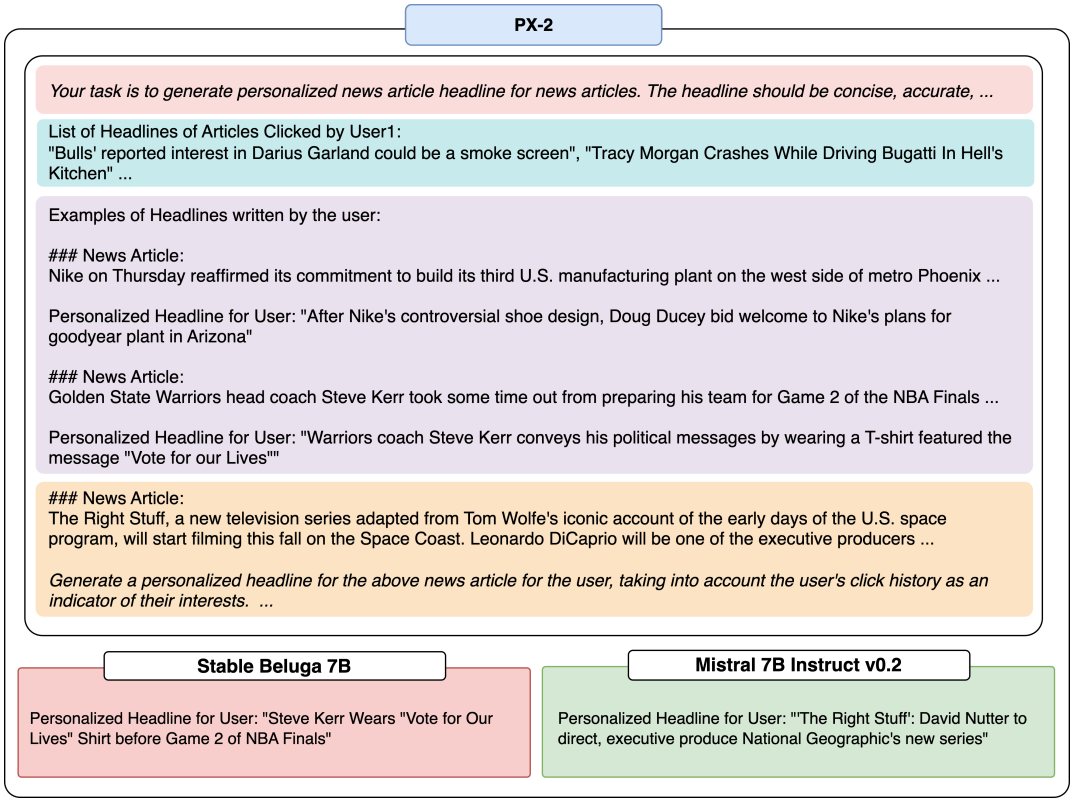


Figure 7: **Illustration of PX-2** (*effect of personalized headline click history*): The left column shows the output of Stable Beluga 7B, which hallucinates incorrect information (marked in red) due to the inability to reinforce historical interest on TV-series with the current interest (i.e., query concepts); the right column shows the output of Mistral 7B Instruct v0.2, which generates the expected response (marked in green).

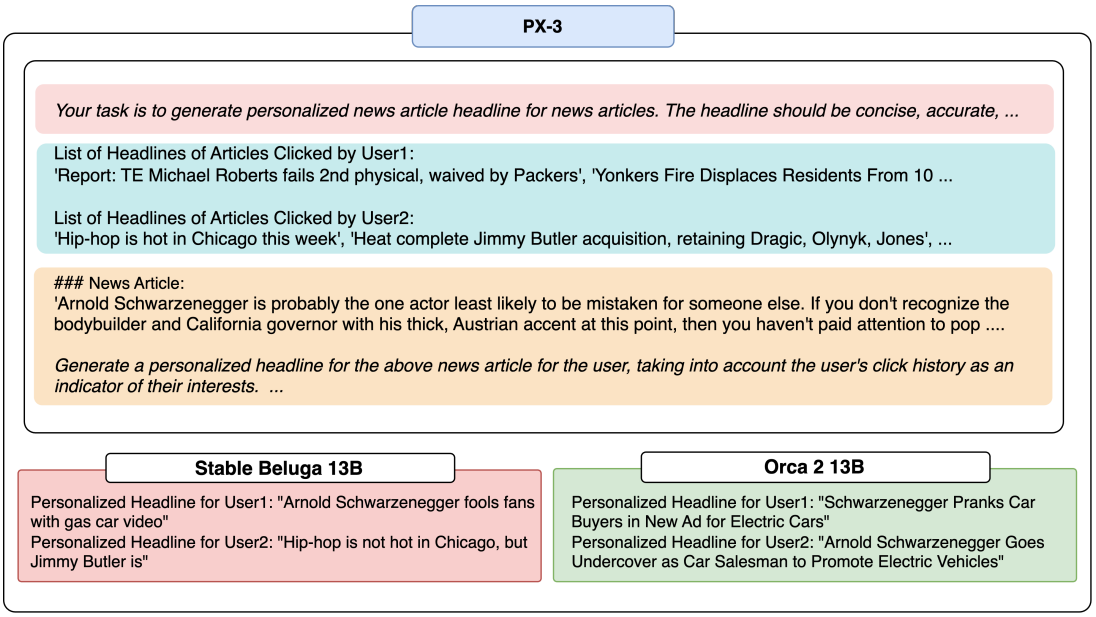


Figure 8: **Illustration of PX-3** (*effect of contrastive personalized click history*): The left column shows the output of Stable Beluga 13B, which hallucinates inaccurate information (marked in red) due to the distraction caused by the list of articles clicked by two users. The right column shows the output of Orca 2 13B, which generates the expected response (marked in green).

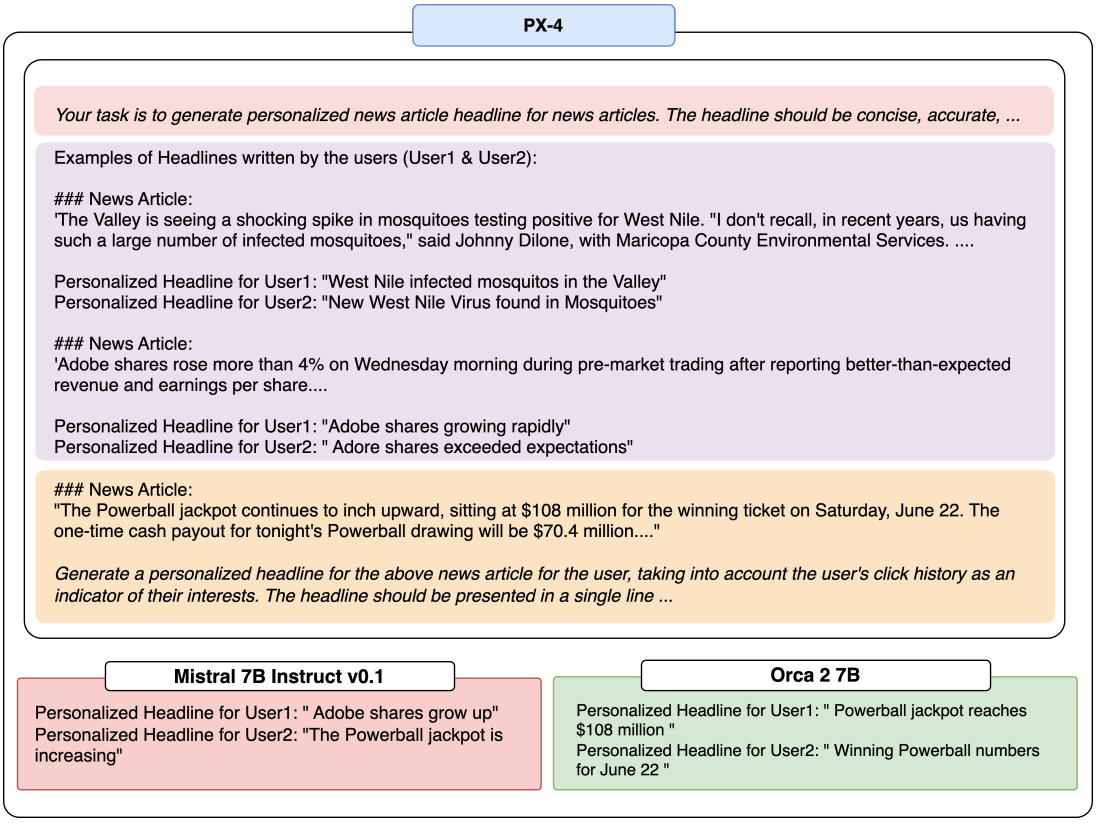


Figure 9: **Illustration of PX-4** (*effect of contrastive personalized examples*): The left column shows the output of Mistral 7B Instruct v0.1, which hallucinates inconsistent information (marked in red) due to the distraction resulting from the unintended reinforcement of common concepts (*adobe*, *share*, *grow*) in both the contrastive personalized headline examples provided by different users that overflows to the response of the query document; the right column shows the output of Orca 2 7B, which generates the expected response (marked in green) and does not suffer from any such overflow.

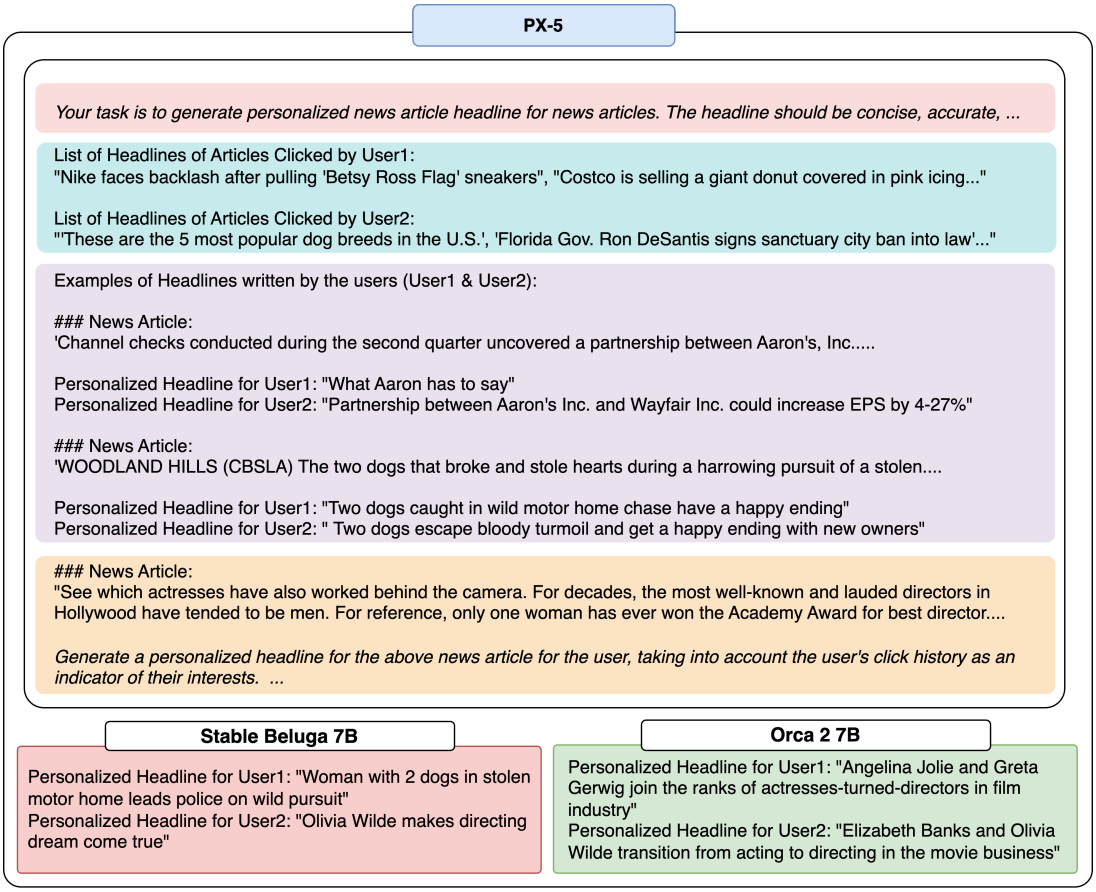


Figure 10: **Illustration of PX-5** (*effect of contrastive personalized examples with click history*): The left column shows the output of Stable Beluga 7B, which hallucinates irrelevant information (marked in red) due to the distraction caused by the cross-association of concepts in the click history of user-1 with that of user-2; the right column shows the output of Orca 2 7B, which generates the expected response (marked in green).