A Diachronic Analysis of Human and Model Predictions on Audience Gender in How-to Guides

Nicola Fanton$^{\alpha}$ **Sidharth Ranjan**$^{\beta}$ **Titus von der Malsburg**$^{\beta}$ **Michael Roth**$^{\alpha,\gamma}$
$^{\alpha}$Institute for Natural Language Processing, University of Stuttgart
$^{\beta}$Institute of Linguistics, University of Stuttgart
$^{\gamma}$Natural Language Understanding Lab, University of Technology Nuremberg
Nicola.Fanton@ims.uni-stuttgart.de

Abstract

We examine audience-specific how-to guides on wikiHow, in English, diachronically by comparing predictions from fine-tuned language models and human judgments. Using both early and revised versions, we quantitatively and qualitatively study how gender-specific features are identified over time. While language model performance remains relatively stable in terms of macro F_1-scores, we observe an increased reliance on stereotypical tokens. Notably, both models and human raters tend to overpredict women as an audience, raising questions about bias in the evaluation of educational systems and resources.

Bias Statement

In the present work, the how-to guides are categorized based on the intended audience and according to their performative construct of gender (Butler, 1989), into: "(for) Women" and "(for) Men". This binary choice has been based on limited data availability for other gender groups. We do not intend to marginalize or exclude any genders or identities, nor to perpetrate any form of representational bias (Blodgett et al., 2020). For the following, the examined biases are the different standards with regards to the gender groups of the intended audience of instruction material. These biases are evidence of binary gender roles, with the masculine gender usually dominating over the feminine gender, and the other genders are excluded by the structure of this system. Therefore the ultimate risk of biased data consists in perpetrating harms in terms of exclusion and inequality. This research not only engages in the awareness of what we excluded here, but also in the development of what data and technical systems could not reiterate, with the broader goal towards fairer socio-technical futures.

	Look Rich Without Being Rich (for Guys)
Early version	"Buying one pair of shoes to go with your wardrobe is impossible[.] Most men think that having 2-3 pair of shoes is enough[.] That's nice but WRONG!"
Revised version	"Get plenty of pairs of shoes and wear them in rotation. If you only have 2 or 3 pairs of shoes, chances are that they'll all be worn and will look old after a while."

Table 1: Example from wikiHow in English. The **title (with audience indicator)** of a guide, an *early* passage and a *revised* passage extracted from the guide.

1 Introduction

Marginalization and discrimination are central topics of recent advances in computational research on educational resources, like school textbooks, which contribute to shaping the sociocultural knowledge of learners (Curdt-Christiansen, 2017). As an example, Crawfurd et al. (2024) study sexism in textbooks, reporting gender bias on various dimensions across the over 30 countries examined. Besides textbooks, also children's stories as sources of educational data drew the attention of the recent advances. Adukia et al. (2022) focus on gender roles: Their research shows women to be, despite progress in terms of representation, still subjected to different treatments. Later work on children's books accounts for intersectional perspectives as well, comprising in the analyses not only texts but also images (Adukia et al., 2023).

Another form of learning material are the collaboratively edited how-to guides from the online platform wikiHow[1]. With this paper, we present a diachronic analysis of texts from wikiHow guides that explicitly indicate a target audience based on gender in the guide's title (like in the example of

[1] www.wikihow.com

Table 1), investigating how predictions by language models and humans differ over time. The predictions in question regard the audience groups *to whom* the instructional texts are tailored.

In sociology, the concept of audience is not new. Erving Goffman's theory about the presentation of self designs the elements of interaction (offline) and conceptualizes the *audience* as part of the (social) performance (Goffman, 1959; Kernaghan and Elwood, 2013). The audience is featured as the entity according to whom the performers act. However, different audiences might lead different performers to act differently. In the case of instructions, the performers are the writers, who might design and revise their texts with having in mind a specific target audience group. As instructional texts are resources meant to guide people in conducting activities, it becomes crucial to assess the variations according to the different intended audiences. The risk of leveraging stereotypes concerning the addressed audience groups might eventually lead to, for example, unfair treatments (cf. Blodgett et al., 2020). Sociolinguistic research also focuses on audience adaptation. This can be traced back at least to the 80s, with early studies on radio speakers (Bell, 1984), but previous research accounting for phenomena related to different audiences in instructional texts is just recent. For example, Fanton et al. (2023) inspect audience-specific guides from wikiHow qualitatively and quantitatively, revealing superficial differences in writing as well as gender-specific standards.

This work extends previous work by including a diachronic perspective as well as by comparing model predictions with human judgments, following two main research questions:

RQ1. Have the patterns learned by the training data changed over time in the task of distinguishing the gendered audience-specific instructional texts?

RQ2. How do model and human predictions in the task of distinguishing the (gendered) audiences of instructional texts differ over time?

Answers contribute to both the Computational Social Science and the Natural Language Processing sub-communities with a focus on gender-related topics. Furthermore, researchers in psychology or marketing domains, especially on the (perceived-)personalization of advertisement (De Keyzer et al., 2015, 2022) could benefit from our research as well. By detailing the gender biases that exist in audience-specific texts and investigating how people and language models use such biases, our work informs efforts to debias instructional text generation and system evaluation.

2 Related Work

In this section, we review gender bias in NLP (§2.1) and connect the present work to previous literature concerning audiences (§2.2).

2.1 Gender Bias in NLP

Gender bias can be defined as preferring and/or having prejudices against one gender (Moss-Racusin et al., 2012; Sun et al., 2019).[2] Studies of gender bias in NLP are nowadays well established, despite their inconsistencies (Blodgett et al., 2020). The mere existence of the workshop series "GeBNLP" (Gender Bias in Natural Language Processing), for the 5th time in 2024 (Faleńska et al., 2024), is on its own a clear sign of the attention of research communities towards the topic of gender bias within NLP. Beyond studies of subtle biases in data (Swim et al., 2004; Falenska and Çetinoğlu, 2021; inter alia), we find: the line of work on biases and debiasing word embeddings (Bolukbasi et al., 2016; Basta et al., 2019; inter alia) and the line of work on LMs (Martinková et al., 2023; Oba et al., 2024; inter alia), or on algorithms suitable for debiasing both (Omrani et al., 2023). In the domain of instructional texts from wikiHow, Suhr and Roth (2024) provide an analysis of gender-neutral language, over time, and on how the editing process includes/excludes efforts towards gender-neutrality. Specifically, they reveal the tendency to add gender-specific, rather than inclusive, language. However, research on gender bias accounting for the different target audience groups is limited.

2.2 Audiences

Compared to the various lines of research on gender bias in NLP, the interest in computational studies on texts for different audiences shows to be smaller – so far. For example, the formalization of the task of profiling the *recipients* is proposed by Borquez et al. (2024). They anchor their work to author profiling, especially to early contributions, including for instance Koolen and van Cranenburgh (2017). While recipient profiling does not fully correspond to distinguishing audiences, it contributes to switching the focus of the well-established author profil-

[2] *One* (preferred) *gender* consists usually of men (see "androcentrism" in Gilman, 1911; cf. Bailey et al., 2019). Prejudices *over the other-* means usually over women (see "sexism" in Swim et al., 1995; and "misogyny" in Zeinert et al., 2021).

ing task (Koppel et al., 2002; Schler et al., 2006; Panicheva et al., 2010; Sap et al., 2014; Mishra et al., 2018; Hsieh et al., 2018; Chen et al., 2024; inter alia). Namely, from addressing the *who?* question about communication – to the other rather new aforementioned tasks – answering to the *to whom?* question. Furthermore, Borquez et al. refer to the Language Accommodation phenomenon, based on the Communication Accommodation Theory (Giles, 1973; Giles et al., 1991) and finding not only several applications, but also addressing diversified audience groups (cf. Bell, 1984; Giles et al., 2023; Allard and Holmstrom, 2023). The work by Fanton et al. (2023) is, to the best of our knowledge, the first computational approach for distinguishing audience-specific English instructional texts. One of the main findings of this work is that the examined texts are subjected to subtle biases. Fanton and Roth (2024) expand on this on a cross-linguistic level. The audience classifiers rely prominently on terms indicating group membership (group terms) and on various attributes reinforcing (gender) stereotypes. On top of these, we aim to both tackle the temporal dimension and to integrate human judgments, thus filling a gap in current research on audience-specific how-to guides.

3 Data

This section presents the data employed for the two studies we conduct. We base our studies on wikiHowAudiences (wHA-EN) (Fanton and Roth, 2024; Fanton et al., 2023), comprising guides tailored for specific audience groups, over gender and age. By opting for this data, we build upon previous findings concerning audience-specific instructional texts. We focus on the gender dimension only. Briefly, each guide comprises title and how-to instructional text. However, pursuing diachronic analyses require further data points, which wHA-EN does not offer. We detail how we proceed in the following.

3.1 Data Preparation: EwHA-EN

We examine whether the patterns learned by the training data changed over time (RQ1), by enhancing the gender subset from wHA-EN with corresponding *earlier* texts by means of revisions histories. For the guides in the scope of our interest in wHA-EN, by retrieving their early[3] versions, we obtain **Early-wikiHowAudiences-ENglish** (EwHA-EN, $N = 1202$). Table 2 reports the number of instances in the stratified partitions we employ to answer to RQ1.

RQ1 Data	Train	Dev	Test
EwHA-EN	961	120	121
wHA-EN	961	120	121

Table 2: The data partitions for answering RQ1.

RQ2 Data	Train	Dev	Test
EwHA-EN	1107	20	80
wHA-EN	1107	20	80

Table 3: The data partitions for answering RQ2. The instances pertaining 2PINS are underlined (they regard development and testing sets only).

3.2 Data Preparation: 2PINS

To investigate differences over time in terms of model vs. human predictions (RQ2), we manually curate a set of ($N = 200$) early and revised extracts. The *early extracts* are text passages from EwHA-EN ($N = 100$) and the *revised extracts* are the corresponding passages from wHA-EN ($N = 100$). We select passages by leveraging potentially relevant terms, including the most influential tokens resulted from previous work, and build **2PerceiveINStructions** (2PINS). For our second study, we split the selected data, henceforth 2PINS, into dev and test sets, following a $2 : 8$ ratio.

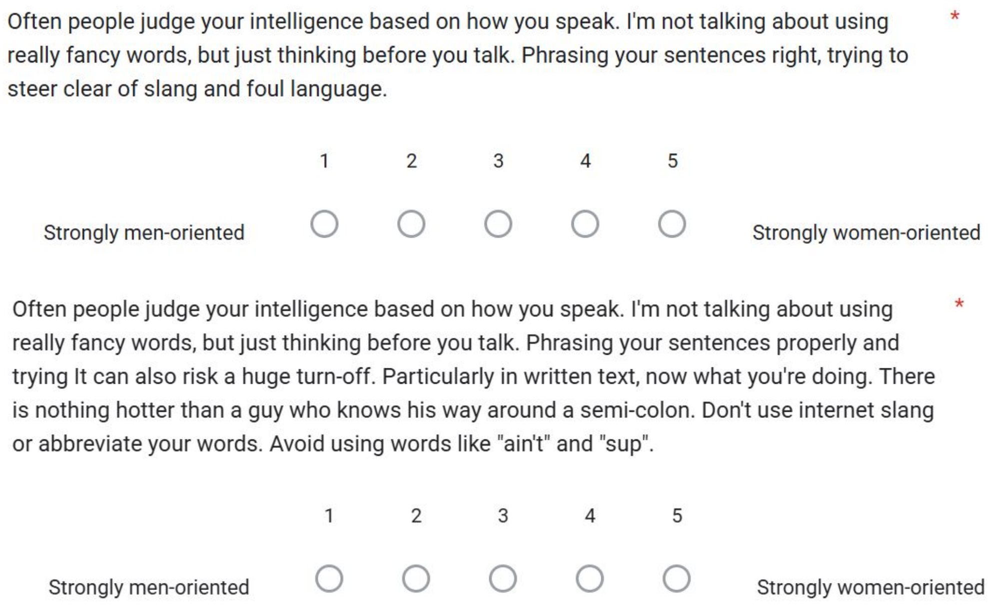
The development instances are 40 (20 early and the corresponding 20 revised) manually curated extracts, balanced with regard to the pertaining audience group. The testing instances are 160 (80 early and the corresponding 80 revised) balanced extracts as well. As training material, we use guides that are not present in the dev and test sets. The 1107 training instances[4] are instructional texts, either in their early version (EwHA-EN) or in their more recent version (wHA-EN). Table 3 displays the data partitions for this part of our work.

4 Human Ratings

This section describes our experimental setup for collecting human preference ratings for the 2PINS

[3] We do not use "*first* versions" (of the guides), when introducing EwHA-EN, because it is not always the case that the retrieved "early" version of a guide is the *very* first. See Appendix A.1 for further details about this.

[4] The dev partition originates from 19 distinct guides. The test partition originates from from 76 distinct guides.



Often people judge your intelligence based on how you speak. I'm not talking about using really fancy words, but just thinking before you talk. Phrasing your sentences right, trying to steer clear of slang and foul language. *

	1	2	3	4	5	
Strongly men-oriented	○	○	○	○	○	Strongly women-oriented

Often people judge your intelligence based on how you speak. I'm not talking about using really fancy words, but just thinking before you talk. Phrasing your sentences properly and trying It can also risk a huge turn-off. Particularly in written text, now what you're doing. There is nothing hotter than a guy who knows his way around a semi-colon. Don't use internet slang or abbreviate your words. Avoid using words like "ain't" and "sup". *

	1	2	3	4	5	
Strongly men-oriented	○	○	○	○	○	Strongly women-oriented

Figure 1: Two different versions of a text from 2PINS, as displayed to participants.

dataset. In this experiment, English speakers were asked to identify the intended audience of a given text on a 5-point scale, as shown in Fig. 1, with 1 indicating strongly (for) Men and 5 indicating strongly (for) Women (Likert, 1932). This finer granularity in scale over, say, the 3-point scale (men, women, unsure) increases the overall engagement of the respondents while taking the survey, as it captures deeper insights into what people are thinking and feeling (Obon et al., 2025). This is because under experimental conditions, people often lean towards skewed choices or may make choices that do not really reflect their thinking (Sullivan and Artino Jr, 2013; Jeong and Lee, 2016). Recently, Heo et al. (2022) found that a 5-point scale is an effective approach to study and compare group differences, such as gender differences.

Participants were shown either of the two text versions of the same text in Google form in a Latin square design (Fisher, 2006)[5]. We used the Prolific[6] platform to recruit our participants. The 2PINS dataset containing 100 pairs was divided into 4 sets and each set containing 50 sentences was rated by 21 different speakers. On average, participants took 15 minutes to complete each set and we paid £6 per participant, including platform service fees.

We now present statistics about the participants involved in our rating experiment. All of the 84 participants resided in the United States of America, and most indicated that the primary language spoken is English (only two participants stated that their primary language spoken is French, Tagalog). 42 participants identified themselves as female (avg. minutes taken 16.6, avg. age 37.5), 41 as male (avg. minutes taken 18.5, avg. age 34.7) and 1 as non-binary (minutes taken 10.1, age 23).

The reasons guiding us in deciding for collecting human ratings are not only in view of assessing humans' performance in the task of distinguishing audience-specific instructional texts, but also because of its importance in view of future work. These ratings can inform us about the challenges towards the evaluation of debiased systems for audience-specific instructional texts generation.

We inspect the instances whose average value of the ratings given by the participants is close to the middle rating (3), that means within $[2.9; 3.1]$. We obtain 3 instances whose gold label is Women, and 11 for Men. After examining them, we opt for keeping for the subsequent analyses the instances whose average rating is comprised by the upper and lower boundaries. However, for the last part of Study 2, we discard the 4 instances (1, development set; 3, testing set) whose average rating is exactly 3, the middle value[7]. We refer to 2PINS without

[5] This experimental design ensures that participants see either version of the text in a way that balances the diachronic initial and revised types.

[6] https://www.prolific.com/

[7] Closer individual ratings' inspections show a rather general (holding for 3 instances out of 4) trend of the middle rating as the modal value.

these 4 instances with 2PINS*.

5 RQ1: Influence of Training Data over Time

In this section, we inspect the effects of using different data in finetuning LMs. We address the following research question (RQ1):

> **Have the patterns learned by the training data changed over time in the task of distinguishing the gendered audience-specific instructional texts?**

5.1 Methodology

To answer, we make use of EwHA-EN, the data previously introduced (see §3.1). We finetune and test the different monolingual LMs from previous work[8]: RoBERTa (Liu et al., 2019) base and BERT (Devlin et al., 2019) base in the cased and uncased versions, accessing them from HuggingFace (Wolf et al., 2020). For comparability, we follow the setup by Fanton and Roth (2024) for our experiments. We use Optuna (Akiba et al., 2019) for 3 hyperparameters' optimization trials for each of the models and maximize the macro F_1 on the development set.[9] We chose macro F_1 to treat each class equally.

In order to compare the pattern learned by the training data over time with respect to our first research question, we need to extract the relevant snippets in view of the finetuned LMs from the instructional texts. We employ a variant of the Integrated Gradients method[10] (Sundararajan et al., 2017), with the instructional text, the finetuned LM and the tokenizer, as inputs. As outputs, we obtain tokens and corresponding scores, which we average in order to inspect the highly influential tokens for the models in the task.

5.2 Results

The performance of the three finetuned LMs on the development sets is always over 80% macro F_1. Surprisingly, the best performing LM from previous work, RoBERTa, ranks only second with its performance with EwHA-EN. The uncased version of BERT obtained 90% macro F_1, surpassing RoBERTa by 6% macro F_1, thus ranking at the top of the list. The held out testing set performance differs only by 1% between the two mentioned LMs (with RoBERTa 87% and uncased BERT 86%).[11]

Early unc. BERT	**Revised RoBERTa**
Women	
friends	Girls
skirt	you
mascara	she
woman	You
parents	Girl
lipstick	Make
yourself	pretty
make	it
friend	the
earring	.
female	pink
Men	
hair	him
people	He
person	male
can	gentleman
shirt	kid
music	Guy
for	Men
this	partner
the	teenager
skin	Boy
is	professional
who	nerd

Table 4: Comparison of audience-specific highly-influential tokens for uncased BERT (trained on Early) vs. RoBERTa (trained on Revised). Stereotypical tokens are highlighted and tokens common for both LMs are excluded, so only differences over time are visible here.

To answer our research question we inspect the 20-top ranking tokens. In a comparative manner, the obtained attributions' lists are set side by side. What we compare *across audiences* are the best-performing LMs on the development sets respectively: for EwHA-EN uncased BERT, and for wHA-EN RoBERTa. This is a subjective qualitative analysis by the authors. For the audience Women, the attributions' lists show that strongly stereotypical tokens tend to get more influential over time ("pretty", "pink"). The same trend holds for the Men audience: "gentleman" and "nerd" appear only for the model finetuned with wHA-EN. More stereotypical

[8] We leave out from our experimental setup the multilingual LMs because they are outperformed by the monolingual LMs.

[9] Please refer to the Appendix for further details (§A.2).

[10] Via Transformers Interpret and its SequenceClassificationExplainer: https://github.com/cdpierse/transformers-interpret.

[11] See Appendix A.3.

Early RoBERTa vs. **Revised RoBERTa**	
Women	
make	she
friends	her
,	makeup
and	Girl
to	pretty
parents	the
yourself	.
not	pink
are	skirt
Men	
hair	him
shirt	He
shirts	he
3	his
Guys	male
4	gentleman
Hair	kid
work	Guy
jeans	partner
we	teenage
2	Boy
wear	professional
Tips	nerd

Table 5: Comparison of audience-specific highly-influential tokens for RoBERTa (trained on Early vs. Revised). Stereotypical tokens are highlighted and tokens common for both LMs are excluded, so only differences over time are visible here.

tokens (e.g. “skirt”, “mascara”, “lipstick”) appear for the class Women from the model finetuned with EwHA-EN than for the class Men (same model finetuned with EwHA-EN). According to Table 5, showing the audience-specific highly-influential tokens for the *same* LM (RoBERTa) finetuned with either early or revised data, the trend is even clearer. The pattern learned by the training data have changed over time towards a more stereotypical direction as well.

6 RQ2: Model vs. Human Predictions

In this section, we study in how far the LMs vs. the participants to the human-subjects experiment correctly predict the gender of the audience groups of the instructional texts. We address the following research question (RQ2):

How do model and human predictions in the task of distinguishing the (gendered) audiences of instructional texts differ over time?

1. BERT finetuned with early data
   → **LM-Early**
2. BERT finetuned with revised data
   → **LM-Revised**
3. RoBERTa finetuned with early data
4. RoBERTa finetuned with revised data
5. DeBERTa finetuned with early data
6. DeBERTa finetuned with revised data

Table 6: The six models we finetune with respect to RQ2.

6.1 Methodology

To answers to RQ2 we study both the model and the human predictions, and we detail below how we obtain the predictions from each subject type. We now focus on 2PINS, assembled for this purpose. With regard to it, the reference points for the human predictions are the human ratings. However, we also need reference points from the models’ perspective, the models’ predictions, in order to compare them with those of the humans.

Models. We finetune a set of LMs to the task of distinguishing the gender groups of the audience-specific passages. Next, we describe our finetuning setup. Moreover, since RQ2 includes time as a further dimension, we finetune each LM with respect to either the early or to the revised data. For this study, as training data for the finetuning we employ either EwHA-EN (§3.1) or wHA-EN from previous work. As development and testing partitions, we make use of 2PINS described in §3.2. Generally, the finetuning framework reflects the one of the previous study (§5.1). We test three different LMs in their base versions: uncased BERT (Devlin et al., 2019); RoBERTa (Liu et al., 2019); DeBERTa (He et al., 2020). However, since we finetune each of them in two flavors, i. e. with two distinct datasets, we prepare altogether 6 models. Moreover, we provide different shallow baselines in the Appendix (§A.4).

To get the LMs predictions, we use the finetuned best-performing models to score 2PINS. With the same approach of the previous study, for the influential tokens’ extraction we obtain attribution lists (influential tokens, their scores and ranking) with respect to the development sets only.

LMs	**Dev**		**Test**	
	Early	**Rev.**	**Early**	**Rev.**
Finetuning with **early** instances				
BERT (unc.)	0.80	0.79	0.68	0.66
RoBERTa	0.74	0.80	0.64	0.69
DeBERTa	0.67	0.60	0.58	0.53
Finetuning with **revised** instances				
BERT (unc.)	0.75	0.90	0.69	0.66
RoBERTa	0.67	0.73	0.64	0.59
DeBERTa	0.80	0.74	0.65	0.67

Table 7: The LMs finetuned with either EwHA-EN or wHA-EN scored on the training and development sets in terms of macro F_1.

Humans. We calculate the average of the ratings by the human participants and convert the results into label predictions l, according to:

$$l = \begin{cases} \texttt{Men}, & \text{if } \bar{r} < 3 \\ \texttt{Women}, & \text{if } \bar{r} > 3, \end{cases}$$

where $\bar{r}$ stands for the average of the ratings for each instance in 2PINS, thus obtaining one human prediction per instance.

6.2 Results

This subsection discusses the results concerning the second research question.

Models. The LMs finetuned with the revised versions tend to perform better on the development set than the models finetuned with the early versions (see Table 7). RoBERTa is an exception: the performance of RoBERTa finetuned with early instances gets +1% macro F_1 with respect to RoBERTa finetuned with revised instances. Nonetheless, for uncased BERT and for DeBERTa, the gain in the opposite direction is substantially larger, respectively: +10% macro F_1 and +7% macro F_1.

For the next steps we focus now only on **LM-Early** and on **LM-Revised**, that are the uncased BERT LMs, finetuned with either EwHA-EN or wHA-EN.

Which different tokens from the same (early/revised) data are highly influential for the two LMs, finetuned with respectively early or revised data? We compare the attributions list model-wise: this means that we inspect what are the differences and the similarities among the highly-influential tokens between the same LM (uncased BERT) finetuned either with EwHA-EN or wHA-EN. What follows concerns the top-20 most influential tokens, respectively.

LM-Early	**LM-Revised**	**Humans**
0.704	0.702	0.815

Table 8: Macro F_1 on 2PINS* ($N = 196$).

Audience	**(for) Women**	**(for) Men**
LM-Early	72	28
LM-Revised	58	42
Humans	59	41

Table 9: Percentage distributions of the predictions by the different subject types on 2PINS* ($N = 196$).

Early data, tokens common to LM-Early and LM-Revised. **Women**: “woman”; “makeup”; “girls”; “!”; “boyfriend”; “school”; “girl”; “she”. **Men**: “jacket”; “took”; “her”; “if”; “with”; “for”; “girlfriend”.

Revised data, tokens common to LM-Early and LM-Revised. **Women**: “.”; “she”; “girls”; “other”; “school”; “,”; “!”. **Men**: “girl”; “this”; “girlfriend”.

In general, we notice how some group terms and pronouns seem to be constant over the two different finetuning datasets (i.e. over time).

Models and humans. We split the comparison between LM and human predictions into two sub-questions:

a. **To what extent do model and human predictions match the gold-standard labels, as in the actual audience group *to whom* the instructional texts are written?**

b. **Are the predictions of LMs and human-subjects leaning towards one audience group?**

With regard to RQ2a, Table 8 reports the macro F_1 on 2PINS*. To compute the extent to which the models predictions and the predictions from the human ratings match the gold-standard labels we choose the macro F_1 score. The perception of human participants scores slightly higher than 0.80.

How then do the human predictions compare to predictions by language models? If we compare both human and model predictions against the indicators given in wikiHow, we find that humans perform better than the LMs, which only score 0.70 in

#	**“Passage” – Title – *Version***
1. HC	“If you pass by them somewhere (School Cafeteria, Hallway, or simply your Classroom), and they suddenly make fun of you out of the blue, don’t cry and run away! Deliver his order of embarrassment! ” Embarrass Your Arch Enemy (Guys) *E.*
2. HC	“dealing with your sexuality is dificalt at some stage in our live we will all want to experimant with our sexuality but if you have made up your mind and know that you want to be a woman then this is for you” Deal With Wanting to Be a Woman (for Men) *E.*
3. HC	“Smile, be the person that you would want to be friends with. Be friendly, outgoing, but not obnoxious! Say hi to her, wave when she comes in the room, or start up a conversation with her.” Get a New Friend (Girls) *E.*
4. HC	“So gender-segregated activities - when you got lumped in with the boys - always made you feel horrible? Or when somebody referred to you as a boy, man, or "he", you always felt a slight tugging feeling that something wasn’t right. You feel different from those around you - you know that you want to be a woman. If you’re having trouble dealing with this, it’s a good idea to take a look at yourself and understand you.” Deal With Wanting to Be a Woman (for Men) *R.*
5. HO	“For pants, you can go with simple dark-washed jeans Skinny jeans go well,but if you are uncomfortable, than loose but not baggy pants are okay” Dress Nerd Chic (for Boys) *E.*
6. HO	“Nobody will think you’re any less weird if you stink! Take a shower either the morning of the first day of school, or the night before. You can try using shower gels if you want to smell nice.” Get a Good Reputation on the First Day of School (for Girls) *R.*

Table 10: The instances mispredicted both by models and humans: “hard cases” (HC), and by humans only (HO).

terms of F_1-score. However, human performance is still far from perfect, which indicates that distinguishing the audience-specific instructional texts is challenging also for human subjects. Table 16 of the Appendix provides the macro F_1 scores for the development and testing partitions of 2PINS* separately as well.

With regard to RQ2b: Both LMs’ and humans’ predictions lean towards the audience group “(for) Women”. The predictions in percentages over early and revised texts can be found in detail in the Appendix A.5. By comparing LM Early and LM Revised on 2PINS*, we note how the former tends to predict (for) Women more than the latter, namely: LM-Early 72% and and LM-Revised 58%. Also the human subjects tend to predict in percentage more the (for) Women class (59%): even more than LM-Revised, but not as much as LM-Early. To draw a ranking, for the tendency of predicting in percentage more (for) Women: LM-Early, followed by human subjects, followed by LM-revised. This result is interesting because it shows how also humans – within the context of this task – are not free to this form of gender bias, which highlights the need for further assessment.

7 Error Analysis

In a qualitative analysis, we found two sets of interesting errors: “hard cases”, namely errors by both LMs and human participants; and human errors, made by human participants only (Table 10). Regarding hard cases, 3 out of 4 pertain to early texts and 3 out of 4 pertain texts that are written for Men, 2 extracted from “Deal With Wanting to Be a Woman (for Men)”, which seems like a particularly challenging guide regarding gender identity.

Among the human errors, one instance may have been misclassified due to the emotional term ‘cry’, which could have biased perception toward Women. In two cases, the phrase “(you want to be a) woman” may have led annotators to incorrectly infer the intended audience as Women, despite the guide being directed at Men. Another misprediction may stem from ambiguous or misleading use of pronouns, further complicating accurate audience prediction.

8 Conclusion

We studied audience-specific how-to guides from a diachronic perspective and compared human judgments and predictions by fine-tuned language models. Our findings indicate that language models over time increasingly rely on stereotypical tokens, with earlier models additionally being heavily biased towards predicting women as a target audience. In our second study, we found that such bias decreases for more highly edited texts (from 72% to 58%) but that even humans are biased in their judgments, favoring women as a target audience in 59%. Our error analysis revealed that stereotypical beliefs, such that only women would want to be (like) woman, could be a potential source for misjudgments.

Audience-unspecific, as in instructional text tailored for a general audience, can inform future analyses, especially with regard to the human ratings and their gender leaning. More data, also beyond the instructional domain, is necessary for broader generalizations. We encourage future work concerning this, and we point to movies' transcripts as well as to advertisement texts as starting points for further research. For more robust gender bias analyses, systematic approaches are needed. A more comprehensive understanding of texts that are written for specific audience groups will definitely benefit from this kind of assessments over different data source.

Our findings underscore the complexity of evaluating educational material and call attention to underlying challenges. Future work should also explore broader demographic attributes and develop methodologies for mitigating representational biases in educational content.

Acknowledgements

This work is supported by the Ministry of Science, Research, and the Arts, Baden-Württemberg through the project IRIS3D (Reflecting Intelligent Systems for Diversity, Demography, and Democracy, Az. 33-7533-9-19/54/5). Work by the first and last authors was in parts funded by the DFG Emmy Noether program (RO 4848/2-1). We would like to thank Aidan Combs and three anonymous reviewers for their valuable feedback.

Limitations

We acknowledge that any perspective represents specific viewpoints. The current work is limited to the English language and to western culture.

While instructional texts are a relevant starting point for assessing educational resources, other data sources are required for further generalizations, also and especially over domains.

Moreover, the present study comprises the attribute of gender of the intended audience groups *to whom* how-to guides are tailored – only. Straightforwardly, the possible (demographic) attributes of audiences are beyond gender (e. g. age, as for example previous work explored).

We remark here the representational bias severely affecting multiple queer identities, beyond the binary of "(for) Women" and "(for) Men", which for the moment are not included in this research.

References

Anjali Adukia, Patricia Chiril, Callista Christ, Anjali Das, Alex Eble, Emileigh Harrison, and Hakizumwami Birali Runesha. 2022. Tales and tropes: Gender roles from word embeddings in a century of children's books. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3086–3097, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anjali Adukia, Alex Eble, Emileigh Harrison, Hakizumwami Birali Runesha, and Teodora Szasz. 2023. What we teach about race and gender: Representation in images and text of children's books. *The Quarterly Journal of Economics*, 138(4):2225–2285.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Amanda Allard and Amanda J. Holmstrom. 2023. Students' perception of an instructor: The effects of instructor accommodation to student swearing. *Language Sciences*, 99:101562.

April H. Bailey, Marianne LaFrance, and John F. Dovidio. 2019. Is man the measure of all things? a social cognitive account of androcentrism. *Personality and Social Psychology Review*, 23(4):307–331. PMID: 30015551.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Martin Borquez, Mikaela Keller, Michael Perrot, and Damien Sileo. 2024. Recipient profiling: Predicting characteristics from messages. *Preprint*, arXiv:2412.12954.

Judith Butler. 1989. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

Hongyu Chen, Michael Roth, and Agnieszka Falenska. 2024. What can go wrong in authorship profiling: Cross-domain analysis of gender and age prediction. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 150–166, Bangkok, Thailand. Association for Computational Linguistics.

Lee Crawfurd, Christelle Saintis-Miller, and Rory Todd. 2024. Sexist textbooks: Automated analysis of gender bias in 1,255 books from 34 countries. *PLOS ONE*, 19(10):1–27.

Xiao Lan Curdt-Christiansen. 2017. *Language Socialization Through Textbooks*, pages 1–16. Springer International Publishing, Cham.

Freya De Keyzer, Nathalie Dens, and Patrick De Pelsmacker. 2015. Is this for me? how consumers respond to personalized advertising on social network sites. *Journal of Interactive Advertising*, 15(2):124–134.

Freya De Keyzer, Nathalie Dens, and Patrick De Pelsmacker. 2022. Let's get personal: Which elements elicit perceived personalization in social media advertising? *Electronic Commerce Research and Applications*, 55:101183.

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.

Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors. 2024. *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Bangkok, Thailand.

Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing gender bias in Wikipedia: Inequalities in article titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 75–85, Online. Association for Computational Linguistics.

Nicola Fanton, Agnieszka Falenska, and Michael Roth. 2023. How-to guides for specific audiences: A corpus and initial findings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 321–333, Toronto, Canada. Association for Computational Linguistics.

Nicola Fanton and Michael Roth. 2024. On shortcuts and biases: How finetuned language models distinguish audience-specific instructions in Italian and English. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 78–93, Bangkok, Thailand. Association for Computational Linguistics.

R.A. Fisher. 2006. *Statistical Methods For Research Workers*. Cosmo study guides. Cosmo Publications.

Howard Giles. 1973. Accent mobility: A model and some data. *Anthropological linguistics*, pages 87–105.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. *Accommodation theory: Communication, context, and consequence*, page 1–68. Cambridge University Press.

Howard Giles, America L. Edwards, and Joseph B. Walther. 2023. Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, 99:101571.

Charlotte Perkins Gilman. 1911. Our androcentric culture: Or, the man-made world.

Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Cindy Yoonjoung Heo, Bona Kim, Kwangsoo Park, and Robin M Back. 2022. A comparison of best-worst scaling and likert scale methods on peer-to-peer accommodation attributes. *Journal of business research*, 148:368–377.

Fernando Hsieh, Rafael Dias, and Ivandré Paraboni. 2018. Author profiling from Facebook corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

H Jae Jeong and Wui Chiang Lee. 2016. The level of collapse we are allowed: Comparison of different response scales in safety attitudes questionnaire. *Biom Biostat Int J*, 4(4):00100.

Donna Kernaghan and Jannette Elwood. 2013. All the (cyber) world's a stage: Framing cyberbullying as a performance. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 7(1):Article 5.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Moshe Koppel, Shlomo Engelson Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Lit. Linguistic Comput.*, 17:401–412.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in West Slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154, Dubrovnik, Croatia. Association for Computational Linguistics.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1722–1742, St. Julian's, Malta. Association for Computational Linguistics.

April M. Obon, Jolly S. Balila, and Edwin A. Balila. 2025. Factor analysis of research culture: A comparative study of 3-point and 5-point likert scales. *International journal of health sciences*, 9(1):26–51.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

Polina Panicheva, John Cardiff, and Paolo Rosso. 2010. Personal sense and idiolect: Combining authorship attribution and opinion analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Hansen Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Katharina Suhr and Michael Roth. 2024. A diachronic analysis of gender-neutral language on wikiHow. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 118–123, St. Julian's, Malta. Association for Computational Linguistics.

Gail M Sullivan and Anthony R Artino Jr. 2013. Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, 5(4):541.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Janet K. Swim, Kathryn J. Aikin, Wayne S. Hall, and Barbara A. Hunter. 1995. Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology*, 68(2):199–214.

Janet K. Swim, Robyn Mallett, and Charles Stangor. 2004. Understanding subtle sexism: Detection and use of sexist language. *Sex Roles*, 51(3/4):117–128.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating online misogyny. 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 3181–3197, Online. Association for Computational Linguistics.

A Appendix

A.1 On the Data Preparation of EwHA-EN

To verify the retrieval of the very first versions, under the intuition that some guides in their very first version might be just "initiated" and not filled with actual content, we sort the texts by length (by whitespace splitting) and explore their distribution. We experimentally set a minimum of 20 tokens as the minimum allowed length of the retrieved versions.

This leaves us with 15 versions affected by not achieving the minimum length. To overcome the issue, we then select the second version, for the affected versions, instead of the very first one. This (selecting the second version) is possible for 13 out of the 15 affected versions. For the other 2, the third and the fourth versions needed to be selected.

The resulting data collection, EwHA-EN comprises 1202 instances, corresponding to the gender-subset of wHA-EN. In EwHA-EN the large majority of the retrieved instances are from the very first version. If not, most of them from the second existing version. For the few remainder: from later versions (third and fourth).

A.2 RQ1: Hyperparameters

Seed: [22, 17, 4]
Learning Rate: [2e-5, 2e-6]
Batch Size: [4, 8]
Epochs: [5]

A.3 RQ1: Performance

LMs EwHA-EN	Train	Dev	Test
RoBERTa	0.98	0.84	0.87
cased BERT	0.98	0.82	0.90
uncased BERT	1.00	0.90	0.86

Table 11: The LMs finetuned with EwHA-EN scored on the partitions in terms of macro F_1.

LMs wHA-EN	Train	Dev	Test
RoBERTa	0.99	0.85	0.86
cased BERT	0.97	0.83	0.84
uncased BERT	0.99	0.80	0.84

Table 12: The LMs finetuned with wHA-EN scored on the partitions in terms of macro F_1. Cf. Fanton and Roth (2024).

A.4 RQ2: Baselines

	Train	Dev	Test
Baselines – early			
Most Frequent	0.46	0.33	0.33
Tf-idf LR	0.54	0.52	0.35
Group Terms LR	0.61	0.33	0.39
Baselines – revised			
Most Frequent	0.46	0.33	0.33
Tf-idf LR	0.53	0.52	0.33
Group Terms LR	0.63	0.33	0.35

Table 13: Baselines, macro F_1 scores. We experiment with 3 different baselines types: one dummy classifier predicting always the most frequent class (Most Frequent); one Logistic Regression baseline with tf-idf (Tf-idf LR) and one Logistic Regression using as features the counts of group terms (Group Terms LR). All baselines are implemented with scikit-learn (Pedregosa et al., 2011) and default values.

Group Terms employed: boy; boys; female; females; girl; girls; guy; guys; male; males; man; men; woman; women; mom; moms; mother; mothers; dad; dads; father; fathers; girlfriend; girlfriends; boyfriend; boyfriends; wife; wives; husband; husbands; dude; dudes; lady; ladies; gentleman; gentlemen.

A.5 RQ2: Results

LMs	Training	
	Early	**Revised**
Finetuning with **early** instances		
uncased BERT	0.96	0.90
RoBERTa	0.96	0.92
DeBERTa	0.95	0.88
Finetuning with **revised** instances		
uncased BERT	0.91	0.99
RoBERTa	0.85	0.94
DeBERTa	0.87	0.95

Table 14: The LMs finetuned with either EwHA-EN or wHA-EN scored on the testing set in terms of macro F_1.

2PINS*	**∼ Dev**	**∼∼ early**	**∼∼ revised**
LM-Early	64 – 36	60 – 40	68 – 32
LM-Revised	54 – 46	45 – 55	63 – 37
Humans	54 – 46	55 – 45	53 – 47
(N)	(39)	(20)	(19)
2PINS*	**∼ Test**	**∼∼ early**	**∼∼ revised**
LM-Early	75 – 25	77 – 23	72 – 28
LM-Revised	59 – 41	60 – 40	57 – 43
Humans	60 – 40	62 – 38	58 – 42
(N)	(157)	(78)	(79)

Table 15: *W–M* percentages of the predictions by models and humans over the gender of the audiences.
W: (for) Women – M: (for) Men.

$\mathbf{F_1}$	**LM Early**	**LM Revised**	**Human Subjects**
2PINS*	0.704	0.702	0.815
∼ Dev	0.816	0.820	0.820
∼ Test	0.674	0.673	0.813

Table 16: To what extent LMs and human predictions match the gold-standard labels.

A.6 RQ2: Errors by the Models only

By LM Early: “Do you wanna look cool? whether you’re going for a skater, or a skinny jeans and jacket look, here’s a few tips to look nice, but cool at the same time.” – Dress Cool (Guys) – Early
“Do you wanna look cool? whether you’re going for a skater, or a skinny jeans and jacket look, here’s a few tips to look nice, but cool at the same time.” – Dress Cool (Guys) – Revised

By LM Revised: “Accept your body. Everybody is different, and that is what makes you so special and unique. Many would envy having such a small butt, as it can be a problem to some girls. Take some pride in yourself and feel good!” – Deal With Having a Small Butt (Teen Girls) – Early

Errors by both LMs: “If you pass by your enemy somewhere (school cafeteria, hallway or simply your classroom) and he suddenly makes fun of you out of the blue, don’t cry and run away! Deliver his order of embarrassment!” – Embarrass Your Arch Enemy (Guys) – Revised