What Makes a Good Summary? Reconsidering the Focus of Automatic Summarization

ter Hoeve, M. ; Kiseleva, Y.; de Rijke, M.

DOI
https://arxiv.org/abs/2012.07619v1

Publication date
2020

Document Version
Submitted manuscript

License
Unspecified

What Makes a Good Summary? Reconsidering the Focus of Automatic Summarization

Maartje ter Hoeve
University of Amsterdam
m.a.terhoeve@uva.nl

Julia Kiseleva
Microsoft Research
julia.kiseleva@microsoft.com

Maarten de Rijke
University of Amsterdam & Ahold Delhaize
m.derijke@uva.nl

ABSTRACT

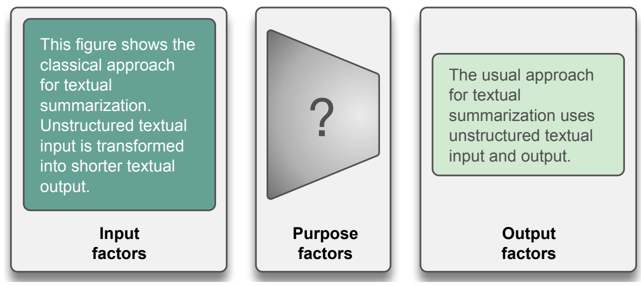
Automatic text summarization has enjoyed great progress over the last years. Now is the time to re-assess its focus and objectives. Does the current focus fully adhere to users' desires or should we expand or change our focus? We investigate this question empirically by conducting a survey amongst heavy users of pre-made summaries. We find that the current focus of the field does not fully align with participants' wishes. In response, we identify three groups of implications. First, we argue that it is important to adopt a broader perspective on automatic summarization. Based on our findings, we illustrate how we can expand our view when it comes to the types of input material that is to be summarized, the purpose of the summaries and their potential formats. Second, we define requirements for datasets that can facilitate these research directions. Third, *usefulness* is an important aspect of summarization that should be included in our evaluation methodology; we propose a methodology to evaluate the usefulness of a summary. With this work we unlock important research directions for future work on automatic summarization and we hope to initiate the development of methods in these directions.

1 INTRODUCTION

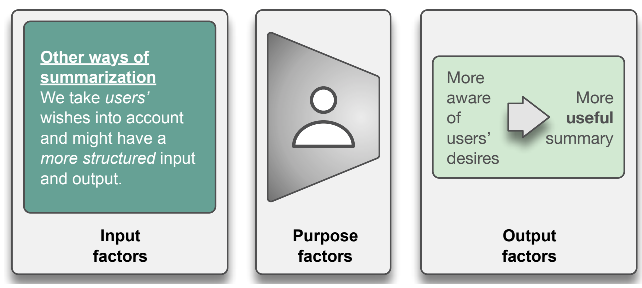
Automatic text summarization has been an important research direction since the early days of the IR and NLP community [26]. The – often *implicit* – goal of the work on automatic text summarization is to generate a condensed textual version of the original input document(s), while preserving the main message of the original source(s). This notion is embedded in today's most common evaluation metrics for the summarization task. These metrics, computed either automatically or by performing a human evaluation, focus on characteristics such as informativeness, fluency, succinctness and, especially recently, factuality [e.g., 16, 22, 32, 34, 36, 48].

In recent years the quality of automatically generated textual summaries has increased tremendously with the rise of neural sequence to sequence models [e.g., 6, 41]. The introduction of Transformers [45] and self-supervised language representation models like BERT [10] have given the summarization quality an additional boost [e.g., 24, 25, 51].

Given these positive developments, it is important to ask ourselves what the future directions of automatic summarization should be and whether the current form of automatic summarization aligns with users' wishes – an important aspect in *explicit* definitions of (the goal of) automatic summarization [e.g., 19, 27]. For example, Mani [27] defines this goal as: "*to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs.*" In this paper we empirically explore users' needs



(a) Most current automatic text summarization techniques.



(b) Example of automatic text summarization with users' desires.

Figure 1: Summarization methods that are currently the standard vs. example of summarizing while taking users' wishes and desires into account.

and compare them with current efforts for automatic text summarization. We also examine how we can evaluate the *usefulness* of a summary in a feasible and comprehensive manner.

We start our investigation by conducting a survey amongst heavy users of *pre-made* summaries. We use the word *pre-made* to differentiate these summaries from summaries that people write themselves, for example to help them to understand a text [e.g., 37]. Instead, pre-made summaries are made by someone else, e.g., a teacher who writes a summary to help students studying for their exams. As automatically generated summaries also fall in this category of pre-made summaries, they should have the same characteristics as users desire for this category. In this research we identify and investigate these characteristics.

During our investigation we focus on the three classes of context factors defined by Jones [19]: *input factors*, *purpose factors* and *output factors*, which describe the input material, the purpose of the summary, and what the summary should look like, respectively. Figure 1 gives an example. Figure 1a shows the textual, unstructured input and output of most current textual summarization techniques. Purpose factors are often ignored. Figure 1b shows an example of how the interpretation of these factors could differ, keeping the purpose of the summary for the user in mind.

We conclude our investigation by carefully examining the implications of our findings. By doing so, we contribute the following:

- **C1** We unveil important and currently underexposed research directions for research on automatic text summarization, concerning the input, purpose and output factors;
- **C2** We define priorities for efforts on dataset collection to be able to study these directions; and
- **C3** We propose a new, feasible and comprehensive evaluation methodology to evaluate the usefulness of a generated summary.

The rest of this paper is structured as follows. We discuss related work in Section 2. In Section 3 we describe our survey. We present our results in Section 4. In Section 5 we discuss the implications of our research for future research on automatic text summarization and we conclude in Section 6.

2 RELATED WORK

Below we introduce the context factors [19] in greater detail and use them to give an overview of methods for automatic text summarization. We then give an overview of evaluation schemes for automatic summarization. We conclude by framing our own position.

2.1 Requirements of an automated summary

Jones [19] argues that one should take the context of a summary into account in order to generate useful summaries – a statement that has been repeated by others [e.g., 2, 27]. To do this in a structured manner, Jones defines three classes of *context factors*: (1) *input factors*, (2) *purpose factors*, and (3) *output factors*. Each of these classes is concerned with a step in the summarization process. Input factors describe the input material that is to be summarized. Output factors describe what the generated summary looks like. Purpose factors are the most important context factors according to Jones; they describe the purpose of the generated summary. Jones argues that the purpose factors are often not fully recognized – a statement that is still timely at present, as we will show in Section 2.2. Each context factor class can be divided into more fine-grained classes. We refer to Table 2 in Appendix A for an extensive overview.

2.2 Automatic text summarization

Now we discuss recent work on automatic summarization, structured around the three context factor classes. Specifically, we connect this work to the fine-grained factors that each of the context factor classes can be divided into.

Input factors. We start with the factor *unit*, which describes how many sources are to be summarized at once, and the factor *scale*, which describes the length of the input data that we are summarizing. These factors are related to the difference between single and multi-document summarization [e.g., 6, 7, 25, 30, 32, 49, 52]. *Scale* plays an important role when material shorter than a single document is summarized, such as in sentence summarization [e.g., 39]. Regarding the *genre* of the input material, we see that most current work on automatic text summarization focuses on the news domain or Wikipedia [e.g., 17, 20, 31, 40]. A smaller body of work addresses different input genres, such as scientific articles [e.g., 8], forum data [e.g., 46], or opinions [e.g., 1]. The aforementioned differences are closely related to the input factor *subject type*, which describes the difficulty level of the input material. The factor *medium* refers to the input language. Most research on automatic text summarization is concerned with English as language input, although there are exceptions, such as Chinese [e.g., 18] or multi-lingual input [21]. The last input factor is *structure*. Especially in recent neural approaches, explicit structure of the input text is often ignored. Exceptions include graph based approaches, where implicit structure is used to summarize a document [e.g., 42, 52], and summarization of tabular data [e.g., 53] or screenplays [e.g., 35].

Purpose factors. Although identified as the most important context factor class by Jones [19] – and followed by, for example, Mani [27] – purpose factors do not receive a substantial amount of attention in work on automatic text summarization. There are some exceptions, such as query-based summarization [e.g., 23, 33], question-driven summarization [e.g., 9] and personalized summarization [e.g., 29]. They take the *situation* and the *audience* into account. The *use*-cases of the generated summaries are also clearer in these approaches than in typical work on automatic text summarization.

Output factors. We start with the output factors *style* and *material*. The latter is concerned with the degree of coverage of the summary. Most generated summaries have an *informative* style and cover most of the input material. There are exceptions. For example, the XSum dataset [31] constructs summaries of a single sentence and is therefore more *indicative* in terms of style and inevitably less of the input material is covered. Not many summaries have a *critical* or *aggregative* style. Aggregative summaries put different source texts in relation to one another, to give an overview of a topic. Currently, most popular summarization techniques focus on a *running format*. Work on template based summarization follows a more *headed* (structured) *format* [e.g., 5]. Falke and Gurevych [14] introduce a more structured format in the form of concept maps and Wu et al. [50] make knowledge graphs. There is also a small body of work on multi-modal summarization, which has a more structured output [e.g., 44, 55]. The difference between *abstractive* and *extractive* summarization is likely the best known distinction in output type [e.g., 15, 25, 30, 32, 41], although it is not entirely clear which output factor best describes the difference.

2.3 Evaluation

Evaluation methods for automatic text summarization can be grouped in different ways. One way is in *intrinsic* vs. *extrinsic* evaluation methods [28]. *Intrinsic* methods evaluate the model itself, for example on informativeness or fluency [e.g., 25, 36]. *Extrinsic* methods target how well the summary performs when used for a certain task [e.g., 11, 48]. Extrinsic methods require a lot of resources, which explains the popularity of intrinsic methods.

Another popular way to distinguish different types of evaluation metrics is between *automatic* and *human* evaluation. Over the years, different *automatic* metrics have been proposed. Rouge [22], which is most popular, evaluates on lexical similarity. The recently proposed BERTScore [54] evaluates on semantic similarity. Wang et al. [48] introduce an automatic extrinsic way of evaluating generated summaries, by automatically generating questions about an input document and answering these questions based on the summary. A similar approach is proposed by Durmus et al. [13]. Most *human* evaluation approaches evaluate intrinsic factors such as informativeness, readability and conciseness [12, 25, 30, 36] – factors that are difficult to evaluate automatically. There are some examples of extrinsic human evaluation methods, where judges are asked to perform a certain task based on the summary. Examples are relevance assessment, where the relevance of a document for a certain topic is judged based on its summary [e.g., 11], and reading comprehension, such as question answering [e.g., 32].

2.4 Our position

We conclude this related work section by stating how our work relates to the context factors and evaluation metrics.

Context factors. In this work we empirically investigate the needs and desires of heavy users of pre-made summaries when it comes to the input, purpose and output factors. We do this by conducting a survey amongst this group of people. The outcomes of this survey allow us to identify underexposed context factors and by doing so to reveal important and exciting research directions for future work on automatic summarization.

Evaluation. We return the observation by Jones [19] that the purpose factors are not explicitly addressed in most work on automatic summarization. In some cases one can justify this, when the usefulness of a generated summary is less important – for example when the objective is to test a certain model architecture or when it is not precisely known who will use the summary (defined by the purpose factor *situation*). However, we agree with Jones that the purpose factors deserve more attention. As a next step, we should also explicitly evaluate generated summaries on their *usefulness* for the intended use-cases. So far, usefulness is not evaluated in a feasible and comprehensive manner. The few existing metrics are often either very resource demanding and too task specific [e.g., 11, 38] or too little specific and hence ignoring the purpose factors [e.g., 12]. Moreover, these metrics are ignored by most current work on automatic summarization. In this paper we aim to bridge the gap by introducing a feasible and comprehensive evaluation methodology to evaluate usefulness.

3 METHOD

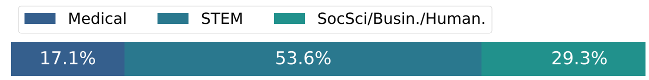
Here we describe the participants of our survey (Section 3.1) and our survey procedure (Section 3.2).

3.1 Participants

We recruited our participants among university students. This group is particularly well suited for our investigation as university students are heavy users of pre-made summaries, for example during exam preparations. Their extensive experience with pre-made summaries makes them experts in this area and therefore they can be expected to have strong and grounded opinions on this topic. Because of their expertise we can use their identified requirements as a reliable starting point to broaden our focus on automatic summarization.

We also considered two different strategies for choosing the participant pool: (1) no restrictions on background / summarization usage, and (2) investigating a number of different target groups.

(a) Study levels.



(b) Study backgrounds.

Figure 2: Participant details.

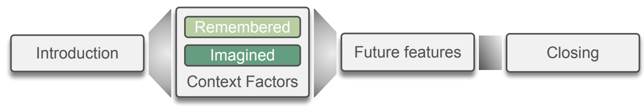


Figure 3: Overview of survey procedure.

The first would not work, as it would be unclear how to compare experiences from different, potentially unexperienced, participants and hence it would not lead to reliable or actionable conclusions. Although the second setup would give us more data points, it comes at the cost of making our study unnecessarily cluttered, whereas not adding much generalizability.

We recruited participants by contacting ongoing courses and student associations, and through advertisements on internal student websites. As an incentive, we offered a ten euro shopping voucher to ten randomly selected participants.

A total of 118 participants started the survey and 82 completed the full survey, resulting in a 69.5% completion rate. We only include participants who completed the study in our analysis. Participants spent 10 minutes on average on the survey. In the final part of our survey we ask participants to indicate their current level of education and main field of study. The details are given in Figure 2.

3.2 Survey procedure

Figure 3 shows a brief, schematic overview of our survey procedure. A detailed account is given in Appendix B, Figure 8. We arrived at this version of the survey after a number of initial pilot runs where we ensured participants understood their task and all questions. We ran the survey with surveymonkey.com. The entire survey – with the exact formulation of the instructions, questions and answer options – is attached in Appendix C to ensure reproducibility. Our survey can also be re-used to inquire different target groups, with slight modifications to match the target group. For example, the current framing around study activities can easily be adapted to activities representative for another target group.

Introduction. The survey starts with an introduction in which we explain to participants what to expect, how we process the data and that participation is voluntarily. After participants agree with this, an explanation of the term *pre-made summary* follows. As we do not want to bias participants by stating that the summary was automatically generated, we explain that this summary can be made by anyone, e.g., a teacher, a good performing fellow student, the authors of the original material, or a computer. Recall from Section 1 that an automatically generated summary is also a pre-made summary and for this reason our survey identifies the characteristics a good, automatically generated summary should have. We also give some examples of types of pre-made summaries, as based on the feedback from our initial pilot experiments we noticed that participants were missing this information. We explicitly state that these are just examples and that participants can come up with any type of summary themselves.

Context factors. In the main part of our survey we focus on the context factors. First, we ask participants whether they have made use of a pre-made summary in one of their recent study activities. If so, we ask them to choose the study activity where a summary was most useful. We call this group of participants the *remembered* group, as they describe an existing summary from memory. If participants indicate that they have not used a pre-made summary in one of their recent study activities, we ask them whether they can imagine a situation where a pre-made summary would have been helpful. If not, we ask them to explain their answer and lead them to the final background questions and closing page. If yes, we ask them to keep this imaginary situation in mind for the rest of the survey. We call this group the *imagined* group.

Now we ask the *remembered* and *imagined* groups about the input, purpose and output factors of the summary they have in mind. We ask questions for each of the subcategories of the context factors that we discussed in Section 2 and that can be found in the overview in Appendix A as well. At this point, the two groups are in different branches of the survey. The difference between the branches is mainly linguistically motivated: in the *imagined* group we use verbs of probability instead of asking them to describe an existing situation. A number of questions can only be asked in the *remembered* group, e.g., how helpful the summary was.

For the first question of the context factors part, we ask participants what the study material consisted of. We give them a number of options, as well as an 'other' checkbox. To avoid position bias, all answers options for multiple choice and multiple response questions in the survey are randomized, with the 'other' checkbox always as the last option. If participants do not choose the 'mainly text' option for this first question, we tell them that we focus on textual input in the current study[1] and we ask them whether they can think of a situation where the input material consisted of text. If not, we lead participants to the background questions and closing page. If yes, they proceed to the remaining questions that give us a full overview of the input, purpose and output factors of the situation that participants have in mind. Finally, we ask the *remembered* group to suggest how their described summary could be turned into their ideal summary. We then ask both groups for any final remarks about the summary or input material.

Trustworthiness and future features questions. Both groups are then led to some exploratory questions. We add these questions to get some initial understanding of the trust users would have in machine-generated summaries and to get some preliminary ideas for the interpretation of the context factors in a less standard setting, but these questions are not the main focus of this research. For the first set of questions we tell participants to imagine that the summary was made by a computer, but contained all the needs that were identified in the previous part of the survey. We then ask them questions about trust in computer versus human generated summaries. As a next step we ask participants to imagine that they could interact with the computer program that made the summary in the form of a digital assistant. We tell them not to feel restricted by the capabilities of today's digital assistants. The full scenario sketch can be found in Appendix C. We then ask participants to select the three most useful and the three least useful features for the digital assistant to have in this scenario, in a similar fashion as in ter Hoeve et al. [43].

[1] We acknowledge that other modalities as well as a mixture of modalities are important to investigate, but leave this for future work to ensure clarity in our results.

Table 1: Different levels of investigation. We did not find significant differences for (4), but add it here for completeness.

1	All respondents together
2	*Remembered* branch vs *imagined* branch
3	Different study fields
4	Different study levels
5	Different levels of how helpful the summary was according to participants, rated on a 5-point Likert Scale (note that only the *remembered* group answered this question)

4 RESULTS

Here we present the outcomes for each survey question and examine them at different levels, summarized in Table 1. For space and clarity reasons, we present the results on a *per group* level when interesting differences are found, otherwise we present the results of all respondents together. We use the question formulation as used for the *remembered* group and abbreviate the answer options.

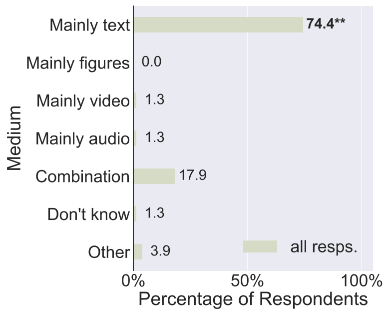
4.1 Identifying branches

Of our participants, 78.0% indicated that they had used a pre-made summary before and hence they were led to the *remembered* branch. Of the remaining 22.0%, 78.2% responded that they could think of a situation where a pre-made summary would be useful for them. They were led to the *imagined* branch. We asked the few remaining participants why they could not think of a situation. People answered that they would not trust a pre-made summary and that making a summary themselves helped them with their study activities. Previous work has indeed found that writing summaries can help with tasks such as reading comprehension [e.g., 37].

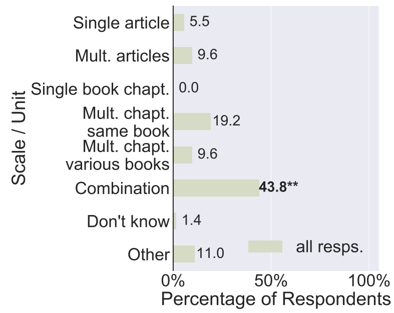
4.2 Input factors

Figure 4 shows the results for the input factor questions. Here we highlight some particularly noteworthy results. First, we see that *textual input* is significantly more popular than the other input types (Figure 4a). This result is based on participants' initial responses and not on the follow up question if they selected another option than 'text'. This stresses the relevance of automatic text summarization. Furthermore, participants described a very diverse input for the factors scale and unit (Figure 4b) – much more diverse than the classical focus of automatic text summarization. Figure 4e shows that most input material had a considerable amount of structure. Typically, this structure is discarded in work on automatic summarization, even though it can be an important source of information. We discuss the implications of these findings in Section 5.1.

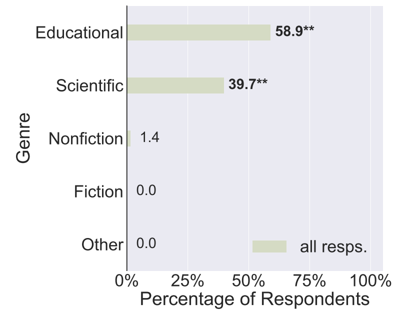
(a) *Medium:* The study material consisted of (MC)



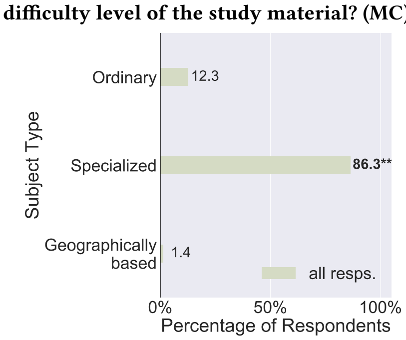
(b) *Scale / Unit:* What was the length of the study material? (MC)



(c) *Genre:* What was the genre of the study material? (MC)



(d) *Subject Type:* How would you classify the difficulty level of the study material? (MC)



(e) *Structure:* How was the study material structured? (MR)

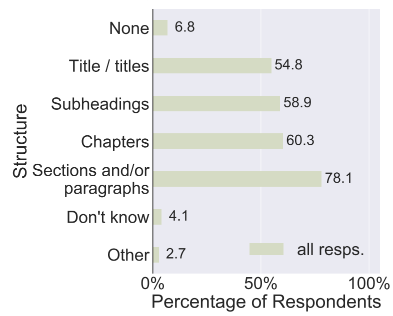


Figure 4: Results for the *input factor* questions. Specific input factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response. ** indicates significance (χ^2), after Bonferroni correction, with $p \ll 0.001$. If two options are flagged with **, these options are not significantly different from each other, yet both have been chosen significantly more often than the other options.

4.3 Purpose factors

Figure 5 shows the results for the purpose factor questions. Again, we highlight a number of particularly interesting results. First, for the purpose factor *audience* (Figure 5d) we asked how much domain knowledge was (or should be) expected from the readers of the summary. We find that people selected the targeted level (4) and targeted level (5) option – i.e. a lot or full domain knowledge – significantly more often than the other options. This corresponds to the result we found in the previous section for the difficulty level of the input – most participants described *specialized* input material (Figure 4d). As our participant pool consists of students, this is rather unsurprising. However, for other target groups the objective could be to rather make a summary for people without a lot of domain knowledge, based on very specialized input material. The main takeaway here is that one should make sure that the expected difficulty level of the summary is aligned with the users' expectations. In our example we can see this was the case. In Figure 5d we show the results split based on how helpful participants indicated the summary was. We can see that targeted level (2) (i.e., almost no domain knowledge) was mostly perceived as not helpful. Although this result is not significant with a Fisher's exact test,[2] it denotes a trend that is worth paying attention to when designing future summarization models.

[2]This is more suitable in this case than a standard χ^2-test due to the small sample size for some options.

In Figure 5e we find how the summary helped participants with their task. We show the results for the *remembered* and the *imagined* group. A first interesting observation from our data (i.e., not denoted in Figure 5e) is that participants in the *imagined* group ticked more boxes than participants in the *remembered* group: 3.33 vs. 2.57 per participant on average. This is a first observation that shows the wide variety of potential use-cases of pre-made summaries. Secondly, we find that the *imagined* group chose the option *refresh memory* and *overview* more often than the *remembered* group (Fisher's exact test, $p < 0.05$). Although this result is not significant after a Bonferroni correction (to correct for the number of tests), we think this can inspire interesting future research directions when it comes to defining the purpose factors for a generated summary. In general, we can see that many different use-cases were very popular, whereas current research on automatic summarization is mostly concerned with simply giving an overview of the input material. This is an important observation that we should use to broaden our vision of the automatic summarization research.

Participants reported that the summary was *helpful* or *very helpful* (Figure 5f), which allows us to draw valid conclusions from the results of this survey.[3] In Section 5.1 we discuss the implications of

[3]Because we do not find any significant differences in the overall results when we exclude these few participants who did not find their summary helpful and we do not find many correlations w.r.t. how helpful a summary was and a particular context factor, we choose to include all participants in the analysis, regardless of how helpful they found their summary.

(a) ***Situation (1):*** **What was the goal of this study activity? (MC)**

(b) ***Situation (2):*** **Who made this pre-made summary? (MC, Only if *remembered*)**

(c) ***Situation (3):*** **The summary was made specifically to help me (and potentially my fellow students) with my study activity (LS, Only if *remembered*)**

(d) ***Audience:*** **For what type of people was the summary intended? (LS)**

(e) ***Use (1):*** **How did this summary help you with your task? (MR)**

(f) ***Use (2):*** **Overall, how helpful was the pre-made summary for you? (LS, Only if *remembered*)**

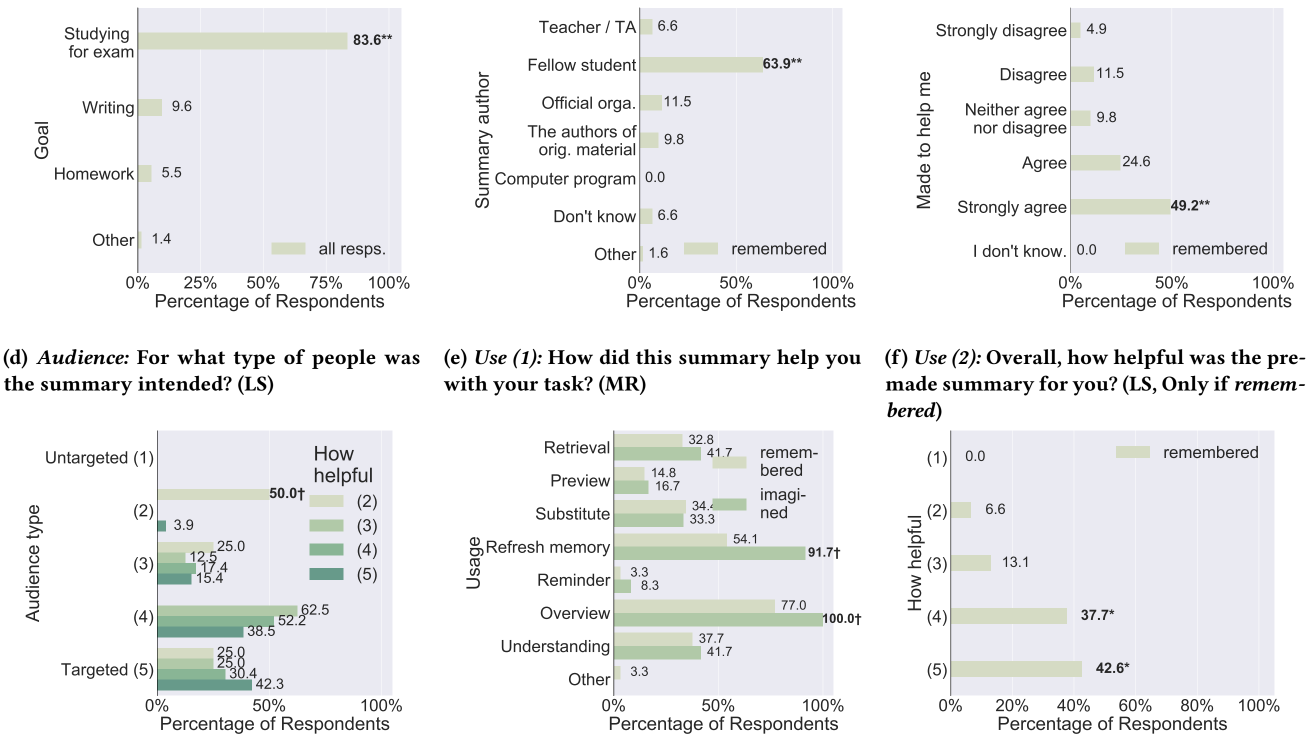


Figure 5: Results for the *purpose factor* questions. Specific purpose factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response, LS = Likert Scale. ** indicates significance (χ^2), after Bonferroni correction, with $p \ll 0.001$, * with $p < 0.05$. † indicates noteworthy results where significance was lost after correction for the number of tests. If two options are flagged, these options are not significantly different from each other, yet both were chosen significantly more often than the other options.

our findings for the research on automatic summarization regarding the purpose factors in more detail.

4.4 Output factors

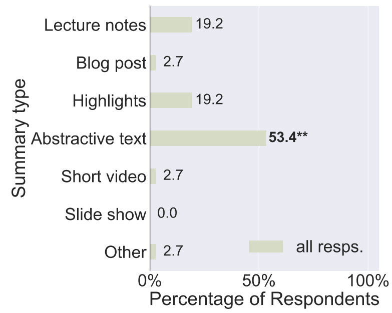
Figure 6 shows the results for the output factor questions. Textual summaries were significantly more popular than the other summary types (Figure 6a). This again stresses the importance of automatic text summarization.

When we investigate the results for the described coverage of the summary (Figure 6c), we find that most participants indicated that the summary covered (or should cover) most of the input material. We split these results based on how helpful the summary was for participants and find that summaries that only covered some of the input material were significantly less helpful for participants than summaries with more coverage. This is in line with the findings that most participants used the summary to study for an exam (Section 4.3, Purpose Factors, Figure 5a). Studying for an exam likely requires an overview of the full study material. This result shows that, in agreement with Jones [19], the purpose factors are extremely important in order to define the output factors.

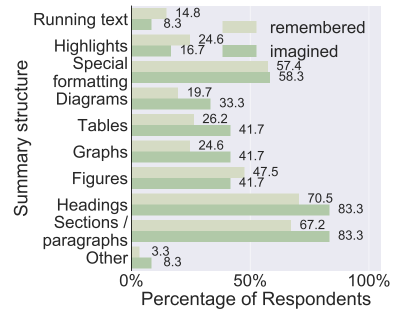
For the output factor *style* we find a fascinating difference between the *remembered* and *imagined* group (Figure 6d). Whereas the *remembered* group described significantly more often an *informative* summary, the *imagined* group opted significantly more often for a *critical* or *aggregative* summary. Most research on automatic summarization focusses on informative summaries only – this result opens up very exciting directions for future research.

The results for the described output structure of the summary (Figure 6b) are also very important and insightful. Participants described a substantially richer format of the pre-made summaries than is adopted in most research on automatic summarization. Instead of consisting of just a running text, the vast majority of participants indicated that the summary contained (or should contain) all kinds of structural elements such as special formatting, diagrams, headings, etc. Moreover, we find that participants in the *imagined* group ticked more boxes on average than participants in the *remembered* group: 4.17 vs. 3.56 per participant, indicating a desire for structure in the generated summaries. This is supported by the answers to the open-ended question where we asked participants in the *remembered* group what would be needed to optimize the described summary. We discuss these results in the next paragraph.

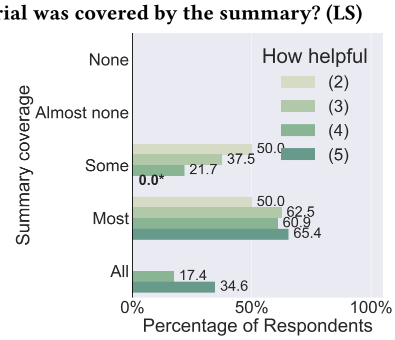
(a) *Format (1):* What was the type of the summary? (MC)



(b) *Format (2):* How was the summary structured? (MR)



(c) *Material:* How much of the study material was covered by the summary? (LS)



(d) *Style:* What was the style of this summary? (MC)

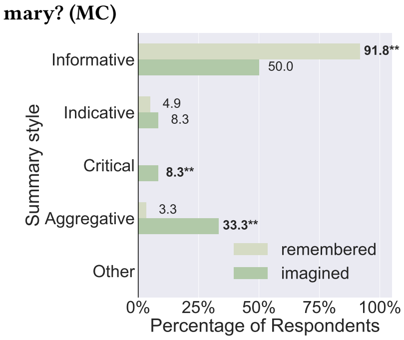


Figure 6: Results for the *output factor* questions. Specific output factor in italics. Answer type in brackets: MC = Multiple Choice, MR = Multiple Response, LS = Likert Scale. ** indicates significance (χ^2 or Fisher's exact test), after Bonferroni correction, with $p \ll 0.001$, * with $p < 0.05$.

The results we found for the output factors unlock many future research directions, which again indicates that we should widen our focus on the automatic summarization research. We discuss this in more detail in Section 5.1.

4.4.1 Open answer questions. We asked the participants who described an existing summary how this summary could be transformed into their ideal summary. Of the participants who filled out this question, 86.9% made suggestions. Many of these suggestions are centered around adding additional structural elements to the summary, like figures, diagrams or tables. One of the participants wrote: *"An ideal summary is good enough to fully replace the original (often longer) texts contained in articles that need to be read for exams. The main purpose behind this is speed of learning from my experience. More tables, graphs and visual representations of the study material and key concepts / links would improve the summary, as I would faster comprehend the study material."* Another participant wrote: *"More images/graphs, to have some changes from just studying from text. Perhaps some online video material about the most difficult parts"*.

Participants also indicated a desire for more structure in the summary text itself, for example by adding headings or color codings. One participant wrote: *"– colors and a key for color-coding – different sections, such as definitions on the left maybe and then the rest of the page reflects the structure of the course material with notes on the readings that have many headings and subheadings."*

Another theme that can be distilled from participants' answers is the desire to have more examples in the summary. One participant wrote: *"More examples i think. For me personally i need examples to understand the material. Now i needed to imagine them myself"*.

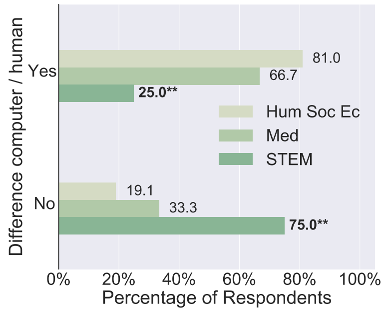
Some participants wrote that they would like to have a more personalized summary: *"I'd highlight some things I find difficult. So I'd personalise the summary more."* Another participant wrote: *"Make it more personalized may be. These notes were by another student. I might have focussed more on some parts and less on others."*

4.5 Trustworthiness and future features

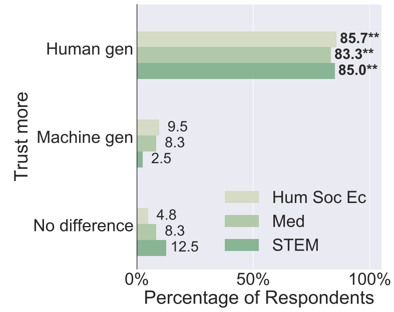
In this section we report the results for the exploratory questions that we asked about the trustworthiness of a summary generated by a machine versus a human, as well as the results for the questions about features for summarization with a digital voice assistant.

We find that participants are divided on the question whether it would make a difference to them whether the summary was generated by a machine or a computer. If we look at all participants together, we find that 48.0.% of the participants answered that it would make a difference, whereas 52.0% answered that it would not. However, if we split the participants based on study background, an interesting difference emerges (Figure 7a). Participants with a background in STEM indicated significantly more often that it would not make a difference to them, whereas the other groups of students indicated the opposite. Almost all participants who

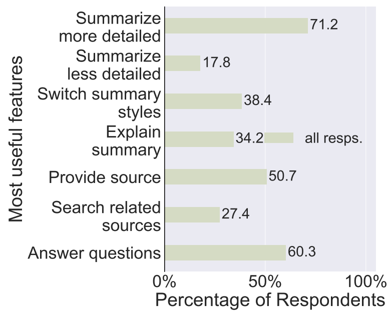
(a) Would it make a difference to you whether the summary was generated by a computer program or by a human? (MC)



(b) Which type of summary would you trust more: (MC)



(c) Please choose the three *most* useful features for a digital assistant to have in this scenario. (MR)



(d) Please choose the three *least* useful features for a digital assistant to have in this scenario. (MR)

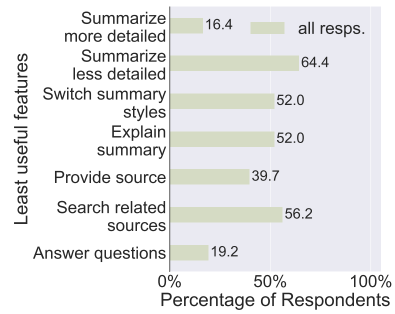


Figure 7: Results for the future feature questions. Answer type in brackets. MC = Multiple Choice, MR = Multiple Response. ** indicates significance (χ^2 or Fisher's exact test), after Bonferroni correction, with $p \ll 0.001$.

answered that it would make a difference said that they would not trust a computer on being able to find the relevant information, i.e., all seemed to favor the human generated summary. Only one participant advocated for the computer generated summary as a *"computer is more objective."* Almost all participants who said it would not matter to them did add the condition that the quality of the generated summary should be as good as if a human had generated it, similar to the observation reported in Vtyurina et al. [47] for automatic systems for conversational search. One person wrote: *"If the summary captures all previously discussed elements it is effectively good for the same purpose. So then it does not matter who generated it."* This comment exactly captures the motivation of the setup of our survey.

This caution regarding automatically generated summaries is confirmed by the question where we asked which type of summary participants would trust more – a human generated one or a machine generated one. People chose the human generated summary significantly more often (Figure 7b). This also holds for the participants with a STEM background, which aligns with the responses to the open questions we reported earlier – apparently participants do not fully trust that the condition they raised earlier would be satisfied, namely that only if the machine was just as good as the human, it would not matter for them whether the summary was generated by a machine or a computer.

The results for the most and least useful features for a digital assistant in a summarization scenario are given in Figure 7c and 7d. Adding more details to the summary and answering specific questions based on the content of the summary are very popular features, whereas summarizing parts of the input material with less detail is not very popular.

Lastly, we asked participants whether they could think of any other features that they would like their digital assistant to have in the outlined scenario. A number of participants answered that they would like the digital assistant to generate questions based on the summary, so that they could test their own understanding. E.g., one participant said: *"Make questions for me (to test me)"* and another participant had a related comment: *"Maybe the the digital assistant could find old exam questions to link to parts of the summary where the question is related to, so that there is a function to test if you've understood the summary."* Another line of answers pointed towards giving explicit relations between the input material and summary, for example: *"Show links between subject materials and what their relation is"* and another person wrote: *"Dynamic linking from summary to original source is a great added value of generating a summary"*.

5 IMPLICATIONS

In the previous section we have presented the results of our survey and we have discussed our interpretation of these findings on a per question basis. Here we want to take the opportunity to summarize our findings as general implications for future research efforts.

We explicitly *do not* argue that current research efforts on automatic summarization are invalid. On the contrary, we are excited to see the great progress of the last years. Instead, we argue that our results help to spark the development of methods and datasets that facilitate a more encompassing range of summaries.

Here, we first discuss the implications of our empirical findings for the *context factors* of summaries and automatic summarization methods. This will then lead us to answer what this means for *dataset collection efforts*. Finally, we move our focus to *evaluation* and propose a new evaluation methodology to evaluate usefulness.

5.1 Context factors

Input factors. As noted in Section 2, the input factors are best represented from all context factors in current research efforts on automatic summarization. Despite this, the vast majority of research focuses on English news or Wikipedia summarization, while stripping away all cues other than raw text. In order to serve a wider audience and to test the generalization capabilities of current approaches, we argue that a wider range of input factors is necessary. Our survey results lead to concrete suggestions: different styles, difficulty levels and including structure. The feasibility of this is strongly dependent on available datasets. Therefore, we focus on the implications for dataset collection efforts later in this section.

Purpose factors. Automatically generated summaries can serve a wide variety of use-cases. Current research efforts often serve the purpose of improving model performance on evaluation metrics such as Rouge [22], leaving the intended use-cases implicit. We see great value in this type of research, as the precise users are not always known, and believe that our research can inspire new directions for summarization that can contribute to the challenging goal of making strong performing models with a good understanding of the input, that generalize to a wide variety of situations.

Nevertheless, we also echo Jones [19]: the purpose factors deserve more attention, as the ideal format of the summary highly depends on them. Who will use the summary, why and in what scenario? Our results inspire many ways of taking the purpose factors into account. We need to evaluate whether our generated summaries are indeed useful for the intended users. As there is no comprehensive way of evaluating this yet, we propose an evaluation methodology to evaluate usefulness at the end of this section.

Output factors. The results of our survey inspire the generation of different summary types that are different than the current standard. We are particularly excited about the implications for the summary format and style. Including more structure in the summaries, for example by means of diagrams or figures, as well as by explicitly adding relations between input text and summary or between parts of the summary as requested by users, requires a thorough understanding of the input text. The many recent publications on model hallucinations [e.g., 13, 48] show that there are still many challenging and exciting research questions in this area.

5.2 Dataset requirements

One important requirement for the execution of the defined research directions is the availability of appropriate datasets. As mentioned in Section 2, there are some datasets available that are unlike most of the existing datasets, especially in terms of input factors. These datasets deserve a stronger recognition and research focus. On top of that, we hope to inspire dataset collection efforts that include a more encompassing range of context factors and especially output factors, e.g., more datasets with an explicit use of structural elements such as diagrams, headers and explicit formatting. We strongly encourage keeping the purpose factors closely in mind during these collection efforts.

5.3 Usefulness as evaluation methodology

Following Jones [19] and Mani [27] we argue that only a correct choice of context factors will result in a useful summary for users. It is important to explicitly evaluate this *usefulness*. In our survey we found that participants mostly found their described summaries very helpful, yet it was hard to define a single factor that makes a summary particularly helpful. Instead, it is the combination that counts. Therefore, *usefulness* can best be evaluated with a human evaluation. As existing metrics are very resource demanding [e.g., 11, 38] or not comprehensive enough [e.g., 12], here we propose a feasible and comprehensive method to evaluate usefulness.

First, from the purpose factors, the intended use factors of the summary need to be identified. Next, the output factors need to be evaluated on these use factors. For this, we take inspiration from research on *simulated work tasks* [3]. Evaluators should be given a specific task to imagine, e.g., *writing a news article*, or *studying for an exam*. This task should be relatable to the evaluators, so that reliable answers can be obtained [4]. With this task in mind, evaluators should be asked to judge two summaries in a pairwise manner on their usefulness, in the following format: *The [output factor] of which of these two summaries is most useful to you to [use factor]?* An example of such a question would be: *The style of which of these two summaries is most useful to you to substitute a chapter that you need to learn for your exam preparation?* As with all human evaluations, it is important to ensure that judges understand the meaning of each of the evaluation criteria, e.g. *style* and *substitute* in the example. We give example questions for each of the remaining output and use factors in Appendix D.

6 CONCLUSION

In this paper we have empirically investigated the desiderata of users of automatic summaries, by means of a survey amongst heavy users of pre-made summaries. We focused on three classes of essential context factors: input, purpose and output factors. We identified that purpose factors are often under-addressed and found that users' desiderata deviate especially from the current focus of automatic summarization research when it comes to the output factors.

Based on these findings we identified important future research directions and requirements for efforts on dataset collection. We also proposed a new methodology to evaluate the usefulness of automatically generated summaries. Our study opens new important future directions to enhance further research on automatic summarization.

ACKNOWLEDGMENTS

We thank Ana Lucic for her helpful feedback. This research was supported by the Nationale Politie. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised Opinion Summarization with Noising and Denoising. *arXiv preprint arXiv:2004.10150* (2020).

[2] Abdelkrime Aries, Walid Khaled Hidouci, et al. 2019. Automatic text summarization: What has been done and what has to be done. *arXiv preprint arXiv:1904.00688* (2019).

[3] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research* 8, 3 (2003), 8–3.

[4] Pia Borlund. 2016. A study of the use of simulated work task situations in interactive information retrieval evaluations. *Journal of Documentation* 72, 3 (2016), 394–413.

[5] Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 152–161.

[6] Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252* (2016).

[7] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 93–98.

[8] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685* (2018).

[9] Yang Deng, Wenxuan Zhang, and Wai Lam. 2020. Multi-hop Inference for Question-driven Summarization. *arXiv preprint arXiv:2010.03738* (2020).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Bonnie Dorr, Christof Monz, Stacy President, Richard Schwartz, and David Zajic. 2005. A methodology for extrinsic evaluation of text summarization: does ROUGE correlate?. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 1–8.

[12] DUC. 2003. DUC 2003: Documents, Tasks, and Measures. (2003).

[13] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. *arXiv preprint arXiv:2005.03754* (2020).

[14] Tobias Falke and Iryna Gurevych. 2017. Bringing structure into summaries: Crowdsourcing a benchmark corpus of concept maps. *arXiv preprint arXiv:1704.04452* (2017).

[15] Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-Up Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4098–4109.

[16] Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 166–175.

[17] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*. 1693–1701.

[18] Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865* (2015).

[19] K Sparck Jones. 1998. Automatic summarizing: factors and directions. In *Advances in automatic text summarization*. Number 1. MIT press Cambridge, Mass, USA, 1–12.

[20] Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305* (2018).

[21] Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A New Benchmark Dataset for Cross-Lingual Abstractive Summarization. *arXiv preprint arXiv:2010.03093* (2020).

[22] Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*. 74–81.

[23] Marina Litvak and Natalia Vanetik. 2017. Query-based summarization using MDL principle. In *Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres*. 22–31.

[24] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198* (2018).

[25] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* (2019).

[26] Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development* 2, 2 (1958), 159–165.

[27] Inderjeet Mani. 2001. *Automatic summarization*. Vol. 3. John Benjamins Publishing.

[28] Inderjeet Mani. 2001. Summarization evaluation: An overview. (2001).

[29] Róbert Móro et al. 2012. Personalized text summarization based on important terms identification. In *2012 23rd International Workshop on Database and Expert Systems Applications*. IEEE, 131–135.

[30] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. In *AAAI*. 3075–3081.

[31] Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745* (2018).

[32] Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking Sentences for Extractive Summarization with Reinforcement Learning. *arXiv preprint arXiv:1802.08636* (2018).

[33] Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1063–1072.

[34] Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*. 145–152.

[35] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020. Screenplay Summarization Using Latent Narrative Structure. *arXiv preprint arXiv:2004.12727* (2020).

[36] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* (2017).

[37] P David Pearson, Linda Fielding, et al. 1991. Comprehension instruction. *Handbook of reading research* 2, 815-860 (1991).

[38] Giuseppe Riccardi, Frederic Bechet, Morena Danieli, Benoit Favre, Robert Gaizauskas, Udo Kruschwitz, and Massimo Poesio. 2015. The sensei project: Making sense of human conversations. In *International Workshop on Future and Emerging Trends in Language Technology*. Springer, 10–33.

[39] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).

[40] Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia* 6, 12 (2008), e26752.

[41] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[42] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1171–1181.

[43] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W White. 2020. Conversations with Documents: An Exploration of Document-Centered Assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 43–52.

[44] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*. 43–52.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[46] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*. 59–63.

[47] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2187–2193.

[48] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. *arXiv preprint arXiv:2004.04228* (2020).

[49] Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2016. A sentence compression based framework to query-focused multi-document summarization. *arXiv preprint arXiv:1606.07548* (2016).

[50] Zeqiu Wu, Rik Koncel-Kedziorski, Mari Ostendorf, and Hannaneh Hajishirzi. 2020. Extracting Summary Knowledge Graphs from Long Documents. *arXiv preprint arXiv:2009.09162* (2020).

[51] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics. Association for Computational Linguistics.

[52] Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based Neural Multi-Document Summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 452–462.

[53] Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020. Summarizing and Exploring Tabular Data in Conversational Search. *arXiv preprint arXiv:2005:11490* (2020).

[54] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[55] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, Chengqing Zong, et al. 2018. MSMO: multimodal summarization with multimodal output. (2018).

A OVERVIEW CONTEXT FACTORS

Table 2: Overview of different context factors classes defined by Jones [19], with descriptions of the factors within these classes.

Input Factors	Purpose Factors	Output Factors
Form	***Situation***	***Material***
Structure: How is the input text structured? E.g. subheadings, rhetorical patterns, etc.	*Tied:* It is known who will use the summary, for what purpose and when.	*Covering:* The summary covers all of the important information in the source text.
Scale: How large is the input data that we are summarizing? E.g. a book, a chapter, a single article, etc.	*Floating:* It is not (exactly) known who will use the summary, for what purpose or when.	*Partial:* The summary (intentionally) covers only parts of the important information in the source text.
Medium: What is the input language type? E.g. full text, telegraphese style, etc. This also refers to which natural language is used.	***Audience***	***Format***
Genre: What type of literacy does the input text have? E.g. description, narrative, etc.	*Targetted:* A lot of domain knowledge is expected from the readers of the summary.	*Running:* The summary is formatted as an abstract like text.
Subject Type	*Untargetted:* No domain knowledge is expected from the readers of the summary.	*Headed:* The summary is structured following a certain standardised format with headings and other explicit structure.
Ordinary: Everyone could understand this input type.	***Use***	***Style***
Specialized: You need to speak the jargon to understand this input type.	*Retrieving:* Use the summary to retrieve source text.	*Informative:* The summary conveys the raw information that is in the source text.
Restricted: The input type text is only understandable for people familiar with a certain area, for example because it contains local names.	*Previewing:* Use the summary preview a text that one is about to read.	*Indicative:* The summary just states the topic of the source text, nothing more.
Unit	*Substitutes:* Use the summary as a substitute for the source text.	*Critical:* The summary gives a critical review of the merits of the source text.
Single: Only one input source is given.	*Refreshing:* Use the summary to refresh ones memory of the source text.	*Aggregative:* Different source texts are put in relation to one another to give an overview of a certain topic.
Multi: Multiple input sources are given	*Prompts:* Use the summary as action prompt to read the source text.	

B SURVEY OVERVIEW

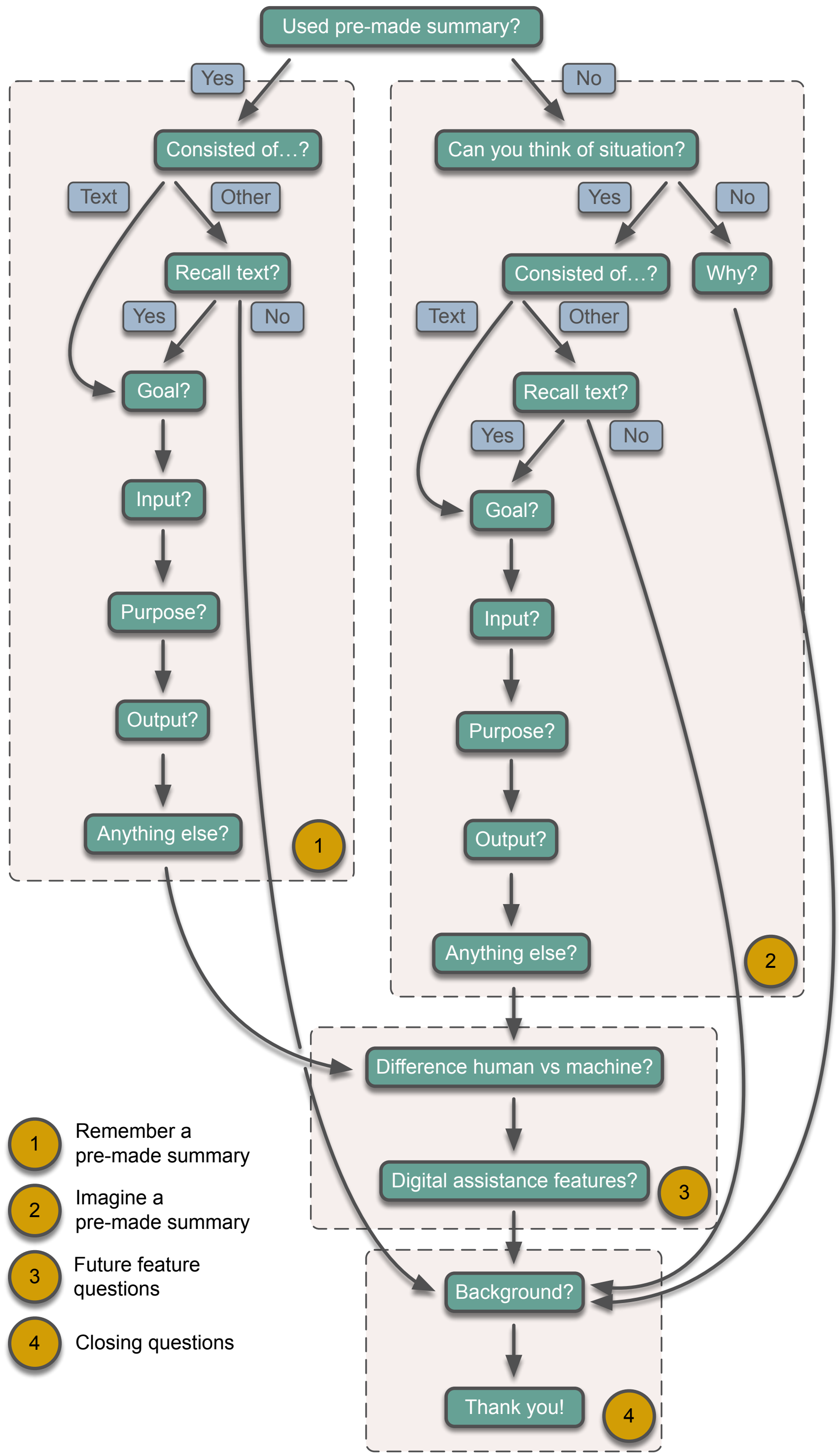


Figure 8: Overview survey design.

C VERBATIM SURVEY OVERVIEW

Table 3: A complete overview of the survey. This table includes the explanation that participants received, as well as all the questions and the answer options. If a question was the start of a branch, the direction of the branch has been written behind the answer options in italic. (This was never shown to the participants.) Note that the survey was performed in SurveyMonkey.[4] The survey had a lay-out as provided by SurveyMonkey, i.e., it consisted of different pages and colors were used to highlight certain important parts in texts.

Question Nr.	Question and Answer Options
Q1	**Introduction and Instructions** Thank you for taking the time to fill out this survey! Before you start, please take the time to read these instructions carefully. If you still have any questions after reading the instructions, please send them to [anonymized]. We will give away 10 [anonymized] vouchers of 10 [anonymized currency] each among the participants. If you would like to take part in the raffle, you can leave your email address at the end of this survey. **Goal of the study** The goal of this survey is to get insight in how summaries help or can help you when studying. **What the survey will look like** In what follows you will get questions that aim to develop an understanding for: • For which types of study material it is useful to have summaries • How these summaries can help you with your task • What these summaries should look like We expect this survey to take approximately 10 minutes of your time. Use the next button to go to the next page once you have filled out all the questions on the page. Use the prev button to go back one page. **About your privacy** We value your privacy and will process your answers anonymously. The answers of all participants in this survey will be used to gain insight in how pre-made summaries can be helpful for different types of studying activities. The answers will be presented in a research paper about this topic. This will be done either in an aggregated manner, or by citing verbatim examples of the answers. Again, this will all be done anonymously. **I agree that I have read and understood the instructions. I also understand that my participation in this survey is voluntarily.** ☐ I agree
Q2	**Important! Some background knowledge you need to know** Throughout this survey we make use of the term pre-made summary. It is very important that you understand what this means. On this page we explain this term, so please make sure to read this carefully.

[4] http://surveymonkey.com

Question Nr.	Question and Answer Options

Definition pre-made summary

One type of summary is one that you make yourself. Another type of summary is one that has been made for you. In this survey, we focus on this latter type and we call them pre-made summaries.

Who makes these pre-made summaries?

These pre-made summaries can be made by a person, for example your teacher, your friend, a fellow student or someone at some official organisation, etc. The pre-made summaries can also be made by a computer.

What kinds of summaries are we talking about?

There are no restrictions on what these pre-made summaries can look like. On the contrary, that is one of the things we aim to find out with this survey! But, to give some examples, you could think of a written overview of a text book, highlights in text to draw your attention to important bits, blog posts, etc. These are really just examples and don't let them limit your creativity! You can come up with any example of a pre-made summary that is helpful for you.

Yes, I understand what a pre-made summary is!
☐ Yes

Q3 Please think back to your recent study activities. Examples of study activities can be: studying for an exam, writing a paper, doing homework exercises, etc. Note that these are just examples, any other study activity is fine too.

Did you use a pre-made summary in any of these study activities?
☐ Yes – *participants are led to Q6*
☐ No – *participants are led to Q4*

Q4 **Can you think of one of your recent study activities where a pre-made summary would have been useful for you?**
☐ Yes – *participants are led to Q25*
☐ No – *participants are led to Q5*

Q5 **Why do you think a pre-made summary would not have helped you with any of your recent study activities?**

Open response – *participants are led to Q48*

Start branch of participants who described an existing summary

If you have multiple study activities where you used a pre-made summary, please take the one where you found the pre-made summary most useful.

Q6 **The original study material consisted of**
☐ Mainly text – *participants are led to Q8*
☐ Mainly figures – *participants are led to Q7*
☐ Mainly video – *participants are led to Q7*
☐ Mainly audio – *participants are led to Q7*
☐ A combination of some or all of the above – *participants are led to Q7*
☐ I do not know, because I have not seen the study material – *participants are led to Q7*
☐ Other (please specify) – *participants are led to Q7*

Question Nr.	Question and Answer Options
Q7	**For now we narrow down our survey to study material that is mostly textual. Do you recall any other recent study activity where you made use of a pre-made summary and where the original study material mainly consisted of text?** □ Yes – *participants are led to Q8* □ No – *participants are led to Q48*
Q8	**What was the goal of this study activity?** □ Studying for an exam □ Writing a paper / essay / report / etc. □ Doing homework exercises □ Other (please specify)
Q9	**Who made this pre-made summary?** □ A teacher or teaching assistant □ A fellow student □ An official organisation □ The authors of the original study material □ A computer program □ I am not sure, I found it online □ Other (please specify)
	Now some questions will follow about what the study material that was summarized looked like.
Q10	**What was the length of the study material?** □ A single article □ Multiple articles □ A single book chapter □ Multiple book chapters from the same book □ Multiple book chapters from various books □ A combination of the above □ I do not know because I have not seen the study material, only the summary □ Other (please specify)
Q11	**How was the study material structured? (Multiple answers possible)** □ There was no particular structure - e.g. just one large text □ The text contained a title or titles □ The text contained subheadings □ The text consisted of different chapters □ The text consisted of different sections and / or paragraphs □ I do not know because I have not seen the study material, only the summary □ Other (please specify)

Question Nr.	Question and Answer Options
Q12	**What was the genre of the study material?** ☐ Mainly educational (such as a text book (chapter)) ☐ Mainly scientific (such as an academic article, publication, report, etc) ☐ Mainly nonfiction writing (such as (auto)biographies, history books, etc) ☐ Mainly fiction writing (such as novels, short fictional stories, etc) ☐ Other (please specify)
Q13	**How would you classify the difficulty level of the study material?** ☐ Ordinary: most people would be able to understand it ☐ Specialized: you need to know the jargon of the field to be able to understand it ☐ Geographically based: you can only understand it if you are familiar with a certain area, for example because it contains local names
	Now we will ask some questions about the purpose of the pre-made summary that you used.
Q14	**The summary was made specifically to help me (and potentially fellow students) with my study activity.** Strongly disagree ☐ · Disagree ☐ · Neither agree nor disagree ☐ · Agree ☐ · Strongly agree ☐ · I don't know ☐
Q15	**For what type of people was the summary intended? Your score can range from (1) Untargetted, to (5) Targetted.** Untargetted: No domain knowledge is expected from the users of the summmary. — Targetted: Full domain knowledge is expected from the users of the summmary. (1) ☐ · (2) ☐ · (3) ☐ · (4) ☐ · (5) ☐

Question Nr.	Question and Answer Options
Q16	**How did this summary help you with your task? (Multiple answers possible)** ☐ The summary helped to retrieve parts of the original study material ☐ I used the summary to preview the text that I was about to read ☐ I used the summary as a substitute for the original study material ☐ I used the summary to refresh my memory of the original study material ☐ I used the summary as a reminder that I had to read the original study material ☐ The summary helped to get an overview of the original study material ☐ The summary helped to understand the original study material ☐ Other (please specify)
Q17	**What was the type of the summary?** ☐ Lecture notes ☐ Blog post ☐ Highlights of some kind in the original study material ☐ Abstractive piece of text, such as a written overview of a text book, an abstract of a paper, etc. ☐ Short video ☐ A slide show ☐ Other (please specify)
Q18	**How was the summary structured? (Multiple answers possible)** ☐ The summary was a running text, without particular structure ☐ The summary consisted of highlights in the original study material, without particular structure ☐ The summary itself contained special formatting, such as bold or cursive text, highlights, etc ☐ The summary contained diagrams ☐ The summary contained tables ☐ The summary contained graphs ☐ The summary contained figures ☐ The summary contained headings ☐ The summary consisted of different sections / paragraphs ☐ Other (please specify)
Q19	**How much of the study material was covered by the summary?** None of the study material was covered (1) ☐ | Almost none of the study material was covered (2) ☐ | Some of the study material was covered (3) ☐ | Most of the study material was covered (4) ☐ | All of the study material was covered (5) ☐

Question Nr.	Question and Answer Options
Q20	**What was the style of this summary?** ☐ Informative: the summary simply conveyed the information that was in the original study material ☐ Indicative: the summary gave an idea of the topic of the study material, but not more ☐ Critical: the summary gave a critical review of the study material ☐ Aggregative: the summary put different source texts in relation to one another and by doing this gave an overview of a certain topic ☐ Other (please specify)
Q21	**Overall, how helpful was the pre-made summary for you? Your score can range from (1) Not helpful at all, to (5) Very helpful.** Not helpful at all … Very helpful (1) ☐ (2) ☐ (3) ☐ (4) ☐ (5) ☐
Q22	**Imagine you could turn this summary into your ideal summary. What would you change?** Open response
Q23	**Is there anything else you want us to know about the summary that we have not covered yet?** Open response
Q24	**Is there anything else you want us to know about the original study material that we have not covered yet?** Open response – *participants are led to Q40*
Start branch of participants who described an imagined summary	
	Please take one of these study activities in mind and imagine you would have had a pre-made summary.
Q25	**The original study material consisted of** ☐ Mainly text – *participants are led to Q27* ☐ Mainly figures – *participants are led to Q26* ☐ Mainly video – *participants are led to Q26* ☐ Mainly audio – *participants are led to Q26* ☐ A combination of some or all of the above – *participants are led to Q26* ☐ Other (please specify) – *participants are led to Q26*

Question Nr.	Question and Answer Options
Q26	**For now we narrow down our survey to study material that is mostly textual. Do you recall any other recent study activity where you could have used a pre-made summary and where the original study material mainly consisted of text?** ☐ Yes – *participants are led to Q27* ☐ No – *participants are led to Q48*
Q27	**What was the goal of this study activity?** ☐ Studying for an exam ☐ Writing a paper / essay / report / etc. ☐ Doing homework exercises ☐ Other (please specify)
	Now some questions will follow about what the study material that could be summarized looked like.
Q28	**What was the length of the study material?** ☐ A single article ☐ Multiple articles ☐ A single book chapter ☐ Multiple book chapters from the same book ☐ Multiple book chapters from various books ☐ A combination of the above ☐ Other (please specify)
Q29	**How was the study material structured? (Multiple answers possible)** ☐ There was no particular structure - e.g. just one large text ☐ The text contained a title or titles ☐ The text contained subheadings ☐ The text consisted of different chapters ☐ The text consisted of different sections and / or paragraphs ☐ Other (please specify)
Q30	**What was the genre of the study material?** ☐ Mainly educational (such as a text book (chapter)) ☐ Mainly scientific (such as an academic article, publication, report, etc) ☐ Mainly nonfiction writing (such as (auto)biographies, history books, etc) ☐ Mainly fiction writing (such as novels, short fictional stories, etc) ☐ Other (please specify)

Question Nr.	Question and Answer Options
Q31	**How would you classify the difficulty level of the study material?** ☐ Ordinary: most people would be able to understand it ☐ Specialized: you need to know the jargon of the field to be able to understand it ☐ Geographically based: you can only understand it if you are familiar with a certain area, for example because it contains local names
	Now we will ask some questions about the purpose of the pre-made summary that would have been helpful.
Q32	**For what type of people should the summary ideally be intended? Your score can range from (1) Untargetted, to (5) Targetted.** Untargetted: No domain knowledge is expected from the users of the summmary. … Targetted: Full domain knowledge is expected from the users of the summmary. (1) ☐ (2) ☐ (3) ☐ (4) ☐ (5) ☐
Q33	**How would this summary help you with your task? (Multiple answers possible)** ☐ The summary would help to retrieve parts of the original study material ☐ I would use the summary to preview the text that I was about to read ☐ I would use the summary as a substitute for the original study material ☐ I would use the summary to refresh my memory of the original study material ☐ I would use the summary as a reminder that I had to read the original study material ☐ The summary would help to get an overview of the original study material ☐ The summary would help to understand the original study material', ☐ Other (please specify)
	Now we will ask some questions about what the summary should look like and cover.
Q34	**What would be the ideal type of the summary?** ☐ Lecture notes ☐ Blog post ☐ Highlights of some kind in the original study material ☐ Abstractive piece of text, such as a written overview of a text book, an abstract of a paper, etc. ☐ Short video ☐ A slide show ☐ Other (please specify)

Question Nr.	Question and Answer Options
Q35	**What is the ideal structure of the summary? (Multiple answers possible)** □ The summary should be a running text, without particular structure □ The summary should consist of highlights in the original study material, without particular structure □ The summary itself should contain special formatting, such as bold or cursive text, highlights, etc. □ The summary should contain diagrams □ The summary should contain tables □ The summary should contain graphs □ The summary should contain figures □ The summary should contain headings □ The summary should consist of different sections / paragraphs □ Other (please specify)
Q36	**How much of the study material should be covered by the summary?** None of the study material should be covered (1) □ · Almost none of the study material should be covered (2) □ · Some of the study material should be covered (3) □ · Most of the study material should be covered (4) □ · All of the study material should be covered (5) □
Q37	**What should the style of this summary be?** □ Informative: the summary should simply convey the information that was in the original study material □ Indicative: the summary should give an idea of the topic of the study material, but not more □ Critical: the summary should give a critical review of the study material □ Aggregative: the summary should put different source texts in relation to one another and by doing this give an overview of a certain topic □ Other (please specify)
Q38	**Is there anything else you would want us to know about your ideal summary that we have not covered yet?** Open response
Q39	**Is there anything else you would want us to know about the original study material that we have not covered yet?** Open response

Look out questions

Now, let's assume the pre-made summary was generated by a computer. You can assume that this machine generated summary captures all the needs you have identified in the previous questions.

Question Nr.	Question and Answer Options
Q40	**Would it make a difference to you whether the summary was generated by a computer program or by a human?** ☐ Yes – *participants are led to Q41* ☐ No – *participants are led to Q43*
Q41	**Please explain the difference.** Open response
Q42	**Which type of summary would you trust more:** ☐ A summary generated by a human, for example a teacher or a good performing fellow student ☐ A summary generated by a computer ☐ No difference
Q43	**Please explain your answer.** Open response
Q44	**Which type of summary would you trust more:** ☐ A summary generated by a human, for example a teacher or a good performing fellow student ☐ A summary generated by a computer ☐ No difference
	Now imagine that you can interact with the computer program that made the summary, in the form of a digital assistant. Imagine that your digital assistant made an initial summary for you and you can ask questions about it to your digital assistant and the assistant can answer them. Answers can be voice output, but also screen output, e.g. a written summary on the screen. In the next part we would like to investigate how you would interact with the assistant. Please do not feel restricted by the capabilities of today's digital assistants.
Q45	**Please choose the three most useful features for a digital assistant to have in this scenario.** ☐ Summarize particular parts of the study material with more detail ☐ Summarize particular parts of the study material with less detail ☐ Switch between different summary styles (for example highlighting vs a generated small piece of text) ☐ Explain why particular pieces ended up in the summary ☐ Provide the source of certain parts of the summary on request ☐ Search for different related sources based on the content of the summary ☐ Answer specific questions based on the content of the summary

Question Nr.	Question and Answer Options
Q46	**Please choose the three least useful features for a digital assistant to have in this scenario.** ☐ Summarize particular parts of the study material with more detail ☐ Summarize particular parts of the study material with less detail ☐ Switch between different summary styles (for example highlighting vs a generated small piece of text) ☐ Explain why particular pieces ended up in the summary ☐ Provide the source of certain parts of the summary on request ☐ Search for different related sources based on the content of the summary ☐ Answer specific questions based on the content of the summary
Q47	**Can you think of any other features that you would like your digital assistant to have to help you in this scenario?** Open response
Background questions	
	Thank you for filling out this survey so far! We would still like to ask you two final background questions.
Q48	**What is the current level of education you are pursuing?** ☐ Bachelor's degree ☐ Master's degree ☐ MBA ☐ Other, please specify
Q49	**What is your main field of study?** Open response
Thank you!	
	You have come to the end of our survey. Thanks a lot for helping out! We very much appreciate your time.
Q50	**If you would like to participate in the raffle to win a voucher, please fill out your e-mail address below. We will only use this e-mail address to blindly draw 10 names who win a voucher and to contact you if your name has been drawn.** Open response

D EXAMPLES EVALUATION QUESTIONS

Here we give additional examples for the evaluation questions that can be used for our proposed evaluation methodology. The phrase *"a document that is important for your task"* should be substituted to match the task at hand. For example, in the case of exam preparations, this could be replaced with *"a chapter that you need to learn for your exam preparation"*. Only the questions with the intended purpose factors should be used in the evaluation.

Purpose factor *Use* & Output factor *Style*:
- The *style* of which of these two summaries is most useful to you to *retrieve* a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *preview* a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *substitute* a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *refresh your memory* about a document that is important for your task?
- The *style* of which of these two summaries is most useful to you to *prompt* you to read a source text that is important for your task?

Purpose factor *Use* & Output factor *Format*:
- The *format* of which of these two summaries is most useful to you to *retrieve* a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *preview* a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *substitute* a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *refresh your memory* about a document that is important for your task?
- The *format* of which of these two summaries is most useful to you to *prompt* you to read a source text that is important for your task?

Purpose factor *Use* & Output factor *Material*:
- The *coverage* of which of these two summaries is most useful to you to *retrieve* a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *preview* a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *substitute* a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *refresh your memory* about a document that is important for your task?
- The *coverage* of which of these two summaries is most useful to you to *prompt* you to read a source text that is important for your task?