Solving the Task but Not the Problem: A Customer Support Case Study on Why Extrinsic Evaluation Matters

Daniel Braun
Marburg University
Department of Mathematics and Computer Science
daniel.braun@uni-marburg.de

Abstract

Natural Language Processing has long been used in customer support to automate and augment human agents. Despite its long-standing use and clear practical relevance, most scientific evaluations rely on intrinsic evaluations and metrics such as accuracy or F1-score. In this paper, we argue that such evaluations often fail to reflect real-world system impact. We present a case study of an NLP system for email-based customer support evaluated both intrinsically and extrinsically via a before-and-after study in deployment. While the system achieves strong intrinsic performance, we observe no measurable improvement in key operational metrics such as average handle time per email. These results highlight a mismatch between benchmark performance and real-world effectiveness, supporting calls for more systematic extrinsic evaluation of NLP systems.

1 Introduction

Natural Language Processing (NLP) has been successfully applied in customer support for decades, to automate and augment the work of customer support representatives. Early systems focused on rule-based dialogue management and information retrieval, while more recent approaches leverage machine learning and large language models to enable tasks such as intent detection, automated response generation, ticket routing, and conversational assistance (see Section 2). Across these developments, the overarching goal has remained consistent: to improve efficiency, reduce operational costs, and enhance customer experience.

Despite this long-standing application and clear practical relevance, the scientific evaluation of NLP in customer support has been predominantly focused on intrinsic evaluation. Systems are regularly assessed based on measures such as accuracy and F1-score on narrow tasks. (Jones and Galliers, 1995) While such metrics provide insights into model performance on these isolated tasks, they often fail to capture the broader, real-world impact of these systems once deployed.

This limitation is not specific to customer support but reflects a broader issue in NLP research. In *We Should Evaluate Real-World Impact*, Reiter (2025) highlights the lack of extrinsic, real-world evaluation across the field and calls for a shift in evaluation practices. He argues that, if NLP systems are intended to be deployed and provide tangible benefits, it is essential to assess their impact on real-world key performance indicators (KPIs) under production conditions, because intrinsic metrics alone are insufficient proxies for practical success.

In this paper, we present a case study of a real-world NLP system for email-based customer support evaluated both intrinsically on annotated test data and extrinsically in deployment using a before-and-after study. While the system achieves strong intrinsic performance (accuracy 0.85 - 0.90), these results do not translate into improvements in KPIs such as average handle time per email. This discrepancy illustrates that intrinsic evaluation not only provides an incomplete picture of system performance, but can in some cases be a poor predictor of real-world impact altogether. The findings reinforce the need to complement traditional benchmarks with evaluations grounded in practical outcomes, supporting the broader call by Reiter (2025) to also assess the real-world effectiveness of NLP systems.

2 Related Work

NLP has long been applied to customer support for email handling, helpdesk systems, and conversational agents. Early work focused on text classification and information retrieval for support requests, such as automated email categorization (Cohen et al., 2004; Carvalho and Cohen, 2006) and helpdesk call routing (Garfield and Wermter,

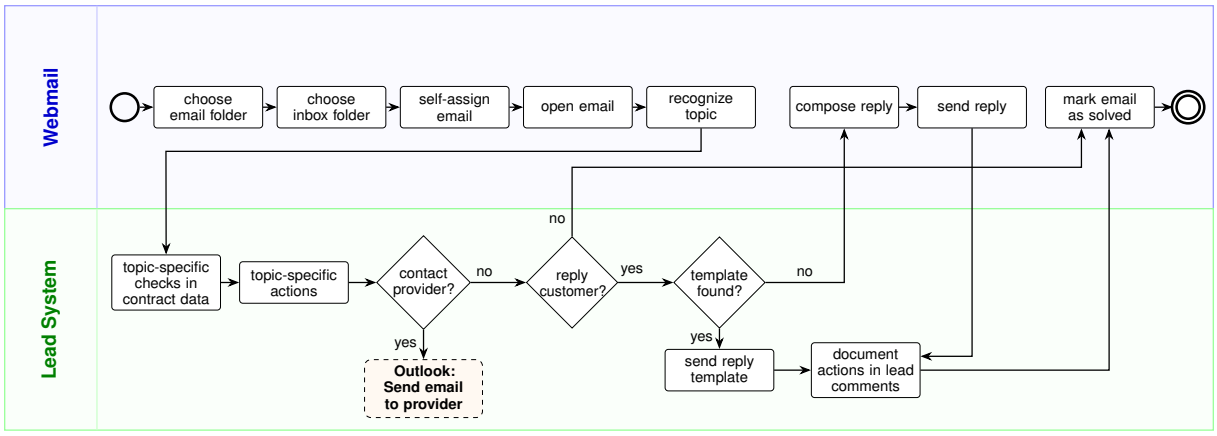

Figure 1: Workflow Customer Support

2002). These systems are typically evaluated using intrinsic metrics such as accuracy, precision, recall, and F1-score on annotated datasets.

More recent approaches leverage neural models and large-scale datasets. For example, Kannan et al. (2016) introduce Smart Reply, a system for automated email response suggestion, mainly evaluated using prediction accuracy and ranking metrics, despite the fact that the system was actually deployed.

Similarly, modern conversational systems (Xu et al., 2017; Hardalov et al., 2018; Farea and Emmert-Streib, 2025; Farnaz and Huyck, 2026) are commonly assessed using benchmark datasets and automatic metrics such as BLEU, dialogue state tracking accuracy, or response selection accuracy. While these evaluations enable comparison across models, they remain largely detached from real-world usage conditions.

A smaller body of work considers extrinsic evaluation in deployed environments. Jain et al. (2018), for example, analyzed aspects like the total interaction time in actual conversations of a chatbot or the count of messages, as well as asking users directly about their satisfaction. Kagan et al. (2025) measure the KPI of chatbot uptake in a context where users can choose between talking to a human customer support representative or a chatbot in a series of A/B tests. Our work adds to this line of research by focusing on the comparison of intrinsic and extrinsic evaluation outcomes, demonstrating that strong intrinsic performance does not necessarily translate into measurable real-world improvements.

3 Business Context

The case study was conducted at a company that brokers and manages energy contracts for consumers. The revenue of the company is generated through commissions for brokered contracts. Therefore, customer support is one of the most important aspects of the business. Due to rapid growth of the company, the service department was struggling to keep up with demand. At the beginning of the collaboration, the customer support of the company received up to 1200 emails per day.

Because the existing software used for handling the emails was based on legacy technology, the company wanted to introduce a new software and in this process also introduce new automation and support features in order to reduce the time that employees spend answering an individual email and reduce the waiting time for customers, particularly for urgent request. We scientifically accompanied the deployment and evaluation of the system.

Through a series of interviews with the head and the deputy-head of the customer service department existing workflow was discovered and formalized. As shown in Figure 1, the customer support mainly relies on two systems in their workflow: a webmail client and a lead system. The webmail client is a simple mail client that is connected to the support email address of the company. Employees log into the system, pick an email, assign themselves to the email, and then take the necessary steps based on the content. The lead system is a CRM system that contains all customer data and information about existing contracts. One noteworthy specificity of the described workflow is that answer templates exist, however they are stored in the lead system and have to be copied manually from the lead system to the webmail client.

In the interviews, five main customer service workflows were identified based on the topic of incoming emails: *data changes* (e.g. updates to ad-

Feature name	NLP	Product
Integrate lead data		X
Integrate lead actions		
Integrate response templates		X
Topic classification	X	X
Lead data augmentation	X	X
Response template suggestion	X	X
Prioritization	X	X
Automatic assignment		
Outlook link for provider contact		X
Sentiment detection		
Lead comments		X
Resubmission		X
Advanced filtering		

Table 1: Implemented (NLP) features in the final software

dress information or electricity meter readings), *revocations* of newly signed contracts, *bonus*-related inquiries (e.g. contractual bonus payments), *status requests* regarding ongoing orders, and contract *cancellations*. An analysis of 1,300 consecutively received emails showed that these topics covered 56% of the incoming mails. Among the other emails, no other frequently (i.e. more than 10 emails) reoccurring topics could be identified.

4 System Design

Through the interview process, 13 new features were identified that could be added to the new system in order to improve operations in the customer service department. Table 1 shows an overview of the features and the nine features that ended up in the final product. Of those features, four are based on NLP models, on which we fill focus. These features are: *topic classification* of incoming emails according to the categories described in Section 3, *linking emails* to the lead database (e.g. via contract or customer ID), automatic *selection of response templates*, and *prioritization* of requests based on their urgency.

After several rounds of pre-experimentation within the company, a decision was made to use a rather simple combination of Tf-idf encoding of incoming emails and their subjects together with a Stochastic Gradient Descent classifier for both the topic and priority classification. For the lead data augmentation, a rule-based system was developed that extracts information like customer IDs, order IDs, invoice numbers, and addresses from incoming emails and matches them against the existing lead data, which will then be displayed in the mail system. Finally, the suggestion of the response templates is based on the identified topics.

Class	Acc.	Prec.	Rec.	F1	Supp.
Bonus	0.975	0.875	0.636	0.737	11
Cancellation	0.938	0.500	0.462	0.480	13
Data Change	0.867	0.826	0.422	0.559	45
Other	0.768	0.751	0.899	0.818	148
Revocation	0.951	0.864	0.731	0.792	26
Status	0.947	0.647	0.688	0.667	16
Total	0.901	0.753	0.753	0.753	259

Table 2: Topic Classification Performance

Class	Acc.	Prec.	Rec.	F1	Supp.
Low	0.855	0.882	0.681	0.769	373
Normal	0.833	0.751	0.798	0.774	386
High	0.858	0.764	0.885	0.820	383
Total	0.849	0.789	0.789	0.789	1142

Table 3: Priority Classification Performance

5 Intrinsic Evaluation

For the intrinsic evaluation of the developed NLP features, a standard evaluation approach was adopted in which real customer emails were annotated by employees, and the models were subsequently evaluated against this annotated data using standard metrics such as accuracy, precision, recall, and F1-score.

5.1 Topic Classification

For the topic classification, a total of 1,295 emails were manually annotated with their respective topic. In addition to the five classes described in Section 3, a sixth class "other" was introduced for all emails that do not fit in any of the classes. 80% of the set was used for training and 20% for testing. The result of the intrinsic evaluation are shown in Table 2, a Table with a confusion matrix can be found in the appendix. Overall, the simple approach performed well in the intrinsic evaluation with an overall accuracy of 0.901.

5.2 Priority Classification

Similarly, for the priority classification, a data set of 5,707 was annotated with three priority classes: low, normal, and high. The set was again split into 80% training and 20% test. The results are shown in Table 2. With an overall accuracy of 0.849, the intrinsic evaluation again revealed good results.

5.3 Information Extraction

Finally, since the information extraction is performed in a rule-based fashion, no training data was needed. Therefore, only a test set, consisting of 107 emails with 254 items of relevant information to be extracted was annotated. The result of

Label	Prec.	Rec.	F1	Support
Invoice Nr	1.00	1.00	1.00	1
City	1.00	0.90	0.95	21
Contract Nr	0.33	0.50	0.40	2
Date	0.91	0.77	0.83	39
Meter Nr	1.00	0.56	0.71	9
Money	1.00	0.65	0.79	23
Order Nr	1.00	1.00	1.00	4
Person	0.94	0.87	0.90	102
ZIP	1.00	0.92	0.96	12
Time	1.00	0.50	0.67	4
Vendor	0.97	0.85	0.91	37
Total	0.95	0.82	0.88	254

Table 4: Information Extraction Performance

the intrinsic evaluation is shown in Table 4.

6 Extrinsic Evaluation

The goal of the extrinsic evaluation was to assess whether the introduced NLP features had an impact on day-to-day operations in the customer support department. We considered two KPIs: average handle time per email, i.e. the time required to respond to an email, and first contact time for high-priority emails, i.e. the time until a customer receives an initial response.

Since the NLP features were introduced together with a new email client, the evaluation was conducted as a before-and-after study in four phases, with an initial eight-week adaptation phase in which employees familiarized themselves with the tool. The introduction of both a new tool and the NLP features could limit the conclusions that can be drawn from a before-and-after study, therefore the generous familirization phase was added.

1. Phase 1: Introduction of the new webmail client (8 weeks)
2. Phase 2: Baseline period without NLP features (10 days)
3. Phase 3: Evaluation period with NLP features (10 days)
4. Phase 4: Post-evaluation without NLP features (10 days)

The fourth phase was introduced because during phase 3 we saw a spike in customer requests compared to phase 2 and we wanted to make sure that effects that we measure between the two phases are not caused by the increased number of requests.

To compute the KPIs, the webmail client was instrumented with logging functionality during phases 2 to 4. The system recorded: when an email was opened, when an email was marked as solved, when a reply was sent, when and which response template was used. In addition, it was logged whether employees modified the automatically assigned topic and priority labels.

6.1 Observations

An average of 486 per Day were opened during Phase 2, 546 during Phase 3 and 677 during Phase 4. In the same period 8,011 emails were sent through the system. Of these replies, 53.84% did not use any of the templates. The increase in email volume between phases could potentially influence processing times. Therefore, as mentioned above, the fourth phase was introduced, so that the mail client without NLP features was used during both a low- and a high-load period.

6.2 Correction of System Predictions

During phase 3, in which the NLP features were activated, more than 10,000 emails were received. For 8,998 emails, the automatically assigned priority was confirmed and for 1,591 it was changed, implying an accuracy of 0.85, which confirms the findings of the intrinsic evaluation. The topic classification was changed 1,545 times and was confirmed 9,321 times, leading to an accuracy of 0.86, which is slightly lower than the result in the intrinsic evaluation, however still on a level that would be considered acceptable.[1]

6.3 Average Handle Time

The average handling time was:

- 3m 19s in phase 2,
- 2m 39s in phase 3, and
- 2m 25s in phase 4.

On a per-user level, we observe substantial variation: some employees remain consistently fast or slow across all phases, while others vary strongly over time. As noted earlier, the email volume increased during the evaluation period, and we observed a negative correlation between workload and handling time, i.e. higher workload is associated with faster responses. Overall, there is no clear evidence that the NLP features had an effect on average handling time.

[1] For some emails one or both predictions were neither actively confirmed nor changed.

Question	Avg. Agreement
Feature-related	
I can orient myself more quickly within an email when a topic has been assigned.	0.26
I generally use the suggested response template when replying to an email.	0.32
I can understand why a particular response template was suggested.	1.16
I generally trust the suggested response template.	0
Comparative	
I see more disadvantages than advantages in the topic recognition feature.	-0.63
I see more disadvantages than advantages in the priority recognition feature.	-0.89
I am more productive with Webmail-B than with Webmail-A.	-0.05
I enjoy working with Webmail-A more than with Webmail-B.	0.32

Table 5: Average agreement to statements from strongly disagree (-2) to strongly agree (2)

6.4 First Contact Time

Finally, the first contact time for high priority emails was reduced in the phase with the NLP features by 11% (from 293 minutes to 260 minutes). Given that the average handle time was not reduced, that meant on the other hand that the first contact time for low priority emails increased, namely from 300 minutes to 491 minutes on average.

7 Survey

After the extrinsic evaluation was concluded, the 19 participating employees received a survey. The survey contained a System Usability Scale (SUS, Vlachogianni and Tselios (2022)) for the client with NLP features (internally named Webmail-A), some questions about specific features, as well as some specific question about the version without NLP features (internally named Webmail-B), and some comparative questions like "With Webmail-B I am more productive than with Webmail-A". The whole survey can be found in Appendix A.

With a SUS score of 70, the NLP-enabled client is slightly above the commonly reported average SUS score of around 68 (Vlachogianni and Tselios, 2022). Given the nature of the SUS items, this result is likely influenced more by the user interface of the newly introduced client than by the underlying NLP features.

While none of the feature-related questions, and thereby the underlying features, were rated negatively, the results indicate only a slightly positive attitude overall (see Table 5). In the comparative question, a stronger signal emerges that the version with NLP features is perceived as overall more helpful and more enjoyable to work with.

Therefore, it is not surprising that, when asked directly, a majority of 59% would prefer to use the new webmail client with NLP features in the future. Only 11% would prefer the new client without NLP features. The remaining 30% would prefer the legacy system.

8 Conclusion

The results of both evaluations show that, despite the fact that NLP features were successfully fulfilling the tasks they were designed to do, the impact on the KPIs was limited, specifically with regard to the most important KPI, the time that is needed to work on customer requests.

There are plenty of potential reasons that can be identified. For example, more than half of all incoming emails fall in the topic category "other", severely limiting the potential impact the topic classification can have, even when working perfectly. This is a fact that was already clear during the design phase, but would have also been highlighted by a purely intrinsic evaluation. Similarly, less than half of the email replies are based on one of the existing template, limiting the potential advantage that the automatic selection of said templates can have.

Nevertheless, had the system been helpful for the remaining half of the emails, as suggested by the intrinsic evaluation, an improvement in the KPIs would still be expected. We believe that this case study illustrates the need for extrinsic evaluations of NLP systems in addition to intrinsic evaluations, as purely intrinsic evaluations are not necessarily good predictors of real-world impact. This is because they do not account for the practical relevance of the selected tasks (e.g., topic classification in this case) within the overall real-world process.

Limitations

This study has several limitations that should be considered when interpreting the results:

- The before-and-after design is susceptible to confounding factors. In particular, changes in workload. Although we reacted to the increase in workload, there was no comparison possible between two phases with the exact same amount of workload.
- The introduction of the NLP features was linked to the introduction of a new webmail client. This introduces additional confounds that may affect user behavior independently of the NLP functionality. We tried to minimize such effects by introducing an eight week period for the employees to familiarise themselves with the new system.
- The evaluation focuses on just two KPIs, namely average handle time and first contact time. While these are important indicators of efficiency, they do not capture other relevant dimensions such as customer satisfaction and response quality.
- Finally, the NLP techniques used in this system are relatively simple. However, despite their simplicity, the techniques proved sufficiently effective in the intrinsic evaluation. Optimizing model architectures or achieving state-of-the-art performance on individual tasks was not the goal of this study, but comparing the results of intrinsic and extrinsic evaluation.

References

Vitor Carvalho and William Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 35–41, New York City, New York. Association for Computational Linguistics.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.

Amer Farea and Frank Emmert-Streib. 2025. Understanding question-answering systems: Evolution, applications, trends, and challenges. *Engineering Applications of Artificial Intelligence*, 156:110997.

Aneela Farnaz and Chris Huyck. 2026. Travquery: A customer support chatbot based on retrieval augmented generation (rag). In *Artificial Intelligence XLII*, pages 130–140, Cham. Springer Nature Switzerland.

Sheila Garfield and Stefan Wermter. 2002. Recurrent neural learning for helpdesk call routing. In *Artificial Neural Networks — ICANN 2002*, pages 296–301, Berlin, Heidelberg. Springer Berlin Heidelberg.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2018. Towards automated customer support. In *Artificial Intelligence: Methodology, Systems, and Applications*, pages 48–59, Cham. Springer International Publishing.

Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, DIS '18, page 895–906, New York, NY, USA. Association for Computing Machinery.

Karen Sparck Jones and Julia R Galliers. 1995. Evaluating natural language processing systems: An analysis and review.

Evgeny Kagan, Brett Hathaway, and Maqbool Dada. 2025. Deploying chatbots in customer service: Adoption hurdles and simple remedies. *arXiv preprint arXiv:2504.06145*.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 955–964, New York, NY, USA. Association for Computing Machinery.

Ehud Reiter. 2025. We should evaluate real-world impact. *Computational Linguistics*, 51(4):1419–1431.

Prokopia Vlachogianni and Nikolaos Tselios. 2022. Perceived usability evaluation of educational technology using the system usability scale (sus): A systematic review. *Journal of Research on Technology in Education*, 54(3):392–409.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 3506–3510, New York, NY, USA. Association for Computing Machinery.

A Survey

A.1 Webmail-A

This section concerns only the current version, *Webmail-A*.

1. I think that I would like to use the system frequently.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

2. I found the system unnecessarily complex.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

3. I found the system easy to use.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

4. I think that I would need the support of a technically skilled person to use the system.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

5. I found the various functions in this system to be well integrated.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

6. I think there was too much inconsistency in the system.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

7. I imagine that most people would learn to use this system very quickly.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

8. I found the system very cumbersome to use.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

9. I felt very confident using the system.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

10. I needed to learn a lot before I could get going with the system.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

11. On a scale from 0 to 10, how accurate is the content recognition for *priority*?
    (10 = very accurate, 0 = very inaccurate)

    _____

12. On a scale from 0 to 10, how accurate is the content recognition for *topic*?
    (10 = very accurate, 0 = very inaccurate)

    _____

13. How easy is it for you to read an email in Webmail-A?

Very Difficult						Very Easy
☐	☐	☐	☐	☐	☐	☐

14. How easy is it for you to reply to an email in Webmail-A?

Very Difficult						Very Easy
☐	☐	☐	☐	☐	☐	☐

A.1.1 If the Topic of an Email Was Correctly Recognized

1. I can orient myself more quickly within an email when a topic has been assigned.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

2. I generally use the suggested response template when replying to an email.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

3. I can understand why a particular response template was suggested.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

4. I generally trust the suggested response template.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

A.1.2 If the Topic of an Email Was Incorrectly Recognized

1. I lose time because I expect a different context.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

2. I lose time because I manually correct the incorrect topic.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

3. The consequently incorrect response template does not affect the time required to compose a reply.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

A.2 Webmail-B

This section concerns only *Webmail-B*.

1. How easy is it for you to read an email in Webmail-B?

Very Difficult						Very Easy
☐	☐	☐	☐	☐	☐	☐

2. How easy is it for you to reply to an email in Webmail-B?

Very Difficult						Very Easy
☐	☐	☐	☐	☐	☐	☐

A.3 Final Questions

1. I am more productive with Webmail-B than with Webmail-A.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

2. I enjoy working with Webmail-A more than with Webmail-B.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

3. If you had to choose, which webmail variant would you prefer?
   - ○ Webmail-A
   - ○ Webmail-B
   - ○ The previous version

4. I see more disadvantages than advantages in the topic recognition feature.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

5. Same question, assuming the recognition accuracy were higher.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

6. I see more disadvantages than advantages in the priority recognition feature.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

7. Same question, assuming the recognition accuracy were higher.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
☐	☐	☐	☐	☐

8. Which functions would you keep activated in Webmail-A?
   - ☐ Priority column in the folder view

- ☐ Topic column in the folder view
- ☐ Automatic selection of a response template
- ☐ Color highlighting of relevant lead data
- ☐ Direct provider contact through the Outlook button

9. How often do you use the webmail system?

   - ○ Continuously
   - ○ A few times per day
   - ○ A few times per week