Integrating AI to assist qualitative research in marketing: using the Annotation Turing Test (AT-Test) to evaluate the performance of a combination of Few-Shot Learning, Prompt-Based Learning and Pre-trained Language Model

Islam El Boudi
Biomedical Engineer
Grenoble Alpes University

Laurie Balbo*
Associate Professor of Marketing
Grenoble Ecole de management

Sandra Camus
Full Professor of marketing
University of Angers

Maud Derbaix
Associate Professor of Marketing
KEDGE Business School Bordeaux

Alexandra Masciantonio
Postdoc
Maastricht University

Aurely Lao
Assistant Professor of Marketing
University of Lille

* Corresponding author: laurie.balbo@grenoble-em.com

Integrating AI to assist qualitative research in marketing: using the Annotation Turing Test (AT-Test) to evaluate the performance of a combination of Few-Shot Learning, Prompt-Based Learning and Pre-trained Language Model

Abstract: The original Turing Test was initially designed to evaluate the performances of a chatbot in its capabilities of confusing a human judge to distinguish between an Artificial Intelligence and a Human interlocutor. In this research, we introduce the Annotation Turing Test (AT-Test), a simplistic adaptation of the original Turing Test, specifically designed to rapidly evaluate the capabilities of a Pre-Trained Language Model (PLMs) fine-tuned through Few-Shot Learning, powered in our case by the GPT-3.5 turbo model. With 71 semi-structured individual interviews collected to examine social media dynamics, the evaluation method that we propose is decomposed in the following steps: (1) first, we selected 3 interviews that helped to fine-tune our PLM through few-show learning, (2) we then annotated these interviews by both our fine-tuned PLM model and a human assistant in research, (3) then these annotated interviews were independently evaluated by three separate researchers. The core criterion for success in our test is the human judge frequency in which the software's annotated interviews are chosen over those annotated by a human. This approach offers a new perspective in evaluating artificial intelligence solutions based on a simplistic approach. The results of our research have shown that modern PLMs can be implemented to automatically help annotate qualitative interviews.

Keywords: Annotation Turing Test, natural language processing, large language model, pre-trained language model, few-shot learning, prompt-based learning, qualitative research

Integrating AI to assist qualitative research in marketing: using the Annotation Turing Test (AT-Test) to evaluate the performance of a combination of Few-Shot Learning, Prompt-Based Learning and Pre-trained Language Model

1. Introduction

Over the two last decades, practitioners and researchers in marketing have been exposed to a massive amount of unstructured data (text and images) generated online by consumers themselves such as products and services reviews, emails, social media posts, *etc*. This abundance of meaningful information has caught the interest of the marketing community (Humprheys & Wang, 2018; Hartman & Netzer, 2023; Shankar & Parsana, 2022). Among those data, text has particularly received a lot of attention in marketing research (Hartman & Netzer, 2023). The recent advancements in *Natural Language Processing* (NLP) models have increased their popularity among researchers (Devlin et al., 2018; Heaven, 2020; Shankar & Parsana, 2022) and have unlocked the possibilities of conducting research. In this research, **we seek to examine the application of a NLP model for text annotation in the context of in-depths individual interviews conducted in a research in marketing**. More specifically, our objective is to compare a human annotator to a combination of few-shot learning, prompt-based learning and pre-trained language model.

In qualitative research, the manual labeling of interview data is a meticulous and essential process (Corbin & Strauss, 1990). Researchers immerse themselves in the data, engaging in line-by-line analysis to identify, code, and categorize emerging themes with minimal preconceived notions, allowing categories to form inductively (Charmaz, 2006; Glaser & Strauss, 1967). Systematic coding procedures start with open coding, followed by axial and selective coding to refine, and identify central categories (Corbin & Strauss 2014; Strauss & Corbin, 1990). The validity of the labeling relies on the constant comparative method, ensuring emerging categories are representative and exhaustive (Glaser, 1965). Inter-coder reliability enhances rigor, with multiple coders analyzing data independently and reconciling differences through discussion (Campbell et al., 2013). Although time-intensive, this process provides nuanced insights and robust analytical conclusions that faithfully reflect participants' voices and perspectives (Miles et al., 2014). In this realm, text labeling assumes a complex role, necessitating methodological standardization to foster the ability to generalize these empirically grounded conclusions (Creswell & Creswell, 2017). This communication explores the progresses of *Pretrained Language Models* (PLMs) as an approach to standardize labeling in qualitative research (Crowston et al., 2012; Guetterman et al., 2018) over 71 interviews conducted with social media users.

PLMs have significantly broadened the horizons of data labeling, particularly within the field of qualitative research for generating grounded findings (Chang et al., 2021; Fang et al., 2022) (see ***Appendix 1*** for a benchmark). PLMs, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), have been pre-trained on vast corpuses of text, enabling them to understand and categorize unlabeled textual data with remarkable precision (Chubb, 2023). Using such models enhanced the labor-intensive process of data labeling, which often acts as a bottleneck in qualitative studies (Richards & Richards, 1994; Vasileiou et al., 2018). Notably, we highlighted that PLMs could grasp nuanced semantic contexts within data, which facilitates the emergence of grounded theory themes that are deeply rooted in the underlying data rather than being confined to a predefined coding framework (Glaser & Strauss, 1967; Raffel et al., 2020).

We introduce the *Annotation Turing-Test* (AT-TEST) as an innovative evaluation designed to assess the performances of *Fine-Tuned PLMs* (FTPLM) in discerning and labeling qualitative data in comparison to human experts. In our test, three interviews were annotated alternatively by a human expert and our FTPLM. Subsequently, the six version of the three annotated interviews were presented to a panel of three human judge experts who were unaware of the annotator's identity between human or FTPLM. These judges then voted independently on which version was the most accurate. If the FTPLM labeling garners a majority of the expert votes, it is considered to have passed the AT-TEST, signifying that its performance is indistinguishable from, if not superior to, that of a human expert in the context of annotating qualitative interviews. This method not only serves as a Turing-like test for FTPLM annotation capabilities but also suggests a paradigm shift in qualitative research methodologies in marketing, but at a more general level in social sciences.

2. Building the NLP Machine

2.1. Participants and objectives of the interviews

This methodological research project is a subpart of a marketing qualitative study that was developed to identify the key characteristics that play a role in the dynamics of social comparison on social media. To this end, we rigorously conducted 71 semi-structured individual interviews (

Appendix 1: NLP et LLMs benchmarking

Benchmarking of NLP methods for automatic text annotation

Approach	Description	Advantages	Disadvantages	Sources
Named Entity Recognition (NER)	Utilizes pre-trained models to identify named entities in text (names, places, etc.).	Well-suited for specific tasks like named entity recognition.	Less effective for general or contextual classification.	Honnibal et al. (2020)
Fine-Tuning Transformers	Adapts pre-trained models (e.g., BERT, GPT-3) on a small number of examples for a specific task.	Leverages powerful models with few examples.	Requires significant computational resources.	Devlin et al. (2018)
Zero-Shot Learning with LLMs	Uses language models like GPT-3 with task descriptions to generate labels without specific examples.	No need for task-specific data.	May be less accurate for highly specific tasks.	Brown et al. (2020)
Few-Shot Learning with LLMs	Uses language models to classify sentences by providing a few examples for each class.	Allows classification with very few specific examples.	Quality may vary depending on the complexity of the data.	Schick et al. (2021)
Support Vector Machines (SVM) with Embeddings	Uses embeddings vectors for classification with an SVM, suitable for text classification tasks.	Effective with well-trained embedding representations.	Less effective for very complex text contexts.	Cortes & Vapnik (1995)
K-Nearest Neighbors (KNN) with Embeddings	Classifies sentences based on nearest neighbors in an embedding space.	Simple to implement and interpret.	Can be slow and less effective with large datasets.	Cover & Hart (1967)

Benchmarking of LLMs performances

Modèle	Taille du Modèle (Paramètres)	Précision (Benchmark GLUE)	Vitesse d'Inférence (ms)	Utilisation Mémoire (GB)	Efficacité Énergétique
GPT-4	1T+	90%	500	80	Moyenne
GPT-3.5	175B	87%	400	16	Élevée
BERT Large	340M	82%	120	12	Très élevée
LLaMA 2 (13B)	13B	85%	150	8	Élevée
PaLM 2	540B	89%	350	32	Moyenne

Source : https://paperswithcode.com/task/language-modelling

Appendix 2 for their profiles) lasting in average 54 min (*SD* = 18 min) with 27 male and 44 female of average 32 years old (*SD* = 14 years old). We developed an interview guide with open-ended questions around four main themes: the first theme dealt with their use of social media (i.e., preferences, frequency of use); the second theme was about self-presentation; the

third theme explored comparisons with other individuals and the domains of comparison; finally, the last focused on the consequences of the use of social media on their subjective well-being (i.e., self-esteem, satisfaction with life, positive and negative emotions, body image). Ethical guidelines were strictly followed, particularly in terms of data privacy (i.e., we did not collect sensitive information, interviewees' names were changed, we asked their consent to participate and to record the interview at the beginning of the study) and the unbiased selection of evaluators. In addition, ethical approval for this study was obtained from the Institutional Review Board of the University of Angers (Approval Number: UA-CER-2022-04). The dialogs, across the 71 interviews, were recorded and transcribed by the researchers and an assistant, which enable them to identify emergent patterns and construct a rich understanding of the participants' experiences to define the labels that our *Fine-Tuned PLM* had to set.

2.2. AT-Test protocol Protocol

The protocol followed six steps. The first step is the **"annotation choice"**. Three researchers in charge of the project read 15 interviews and identify the list of labels to be retrieved in the entire interview corpus. These interviews were chosen based on their diversity in content and complexity to ensure a comprehensive evaluation of annotation capabilities. Secondly, we proceeded to the **"text annotation by software"** step. The designated software was tasked with annotating 3 interviews, selected for their diversity in terms of participant's age, gender and socio-demographic status (Creswell & Poth, 2017) and the requirement that they altogether covered all the 41 identified labels at least twice. We decided to evaluate three interviews during the AT-Test because it will never lead to a situation where there is an equality between the score of the NLP Software-annotated interview and the HUMAN-annotated interview. Prior experiments with 15 interviews have helped identify probabilities of success for each annotation event in our research:

- Finding the right label at random for a human annotator = 0.02
- Mean number of paragraphs to label per interview = 66
- Chances of a human choosing a correct annotation = 0.91

Statistically, to estimate the odds of our AI-automated labeling experience beating the human annotator by pure luck, we can identify a Binomial distribution experiment B (66, 0.02) where the AI has to annotate on average at least 61 of the 66 paragraphs per interview: $P(X>61) = 1 - (\sum_{k=0}^{61}\binom{66}{k}.(0.02)^{k}.(1-0.02)^{66-k})$ so $P(X>61) = 1.22\times10^{-15}$ (p-value < 0.05), it indicates that the probability of the AI-automated labeling beating the human annotator by chance is extremely low. Again, we want to estimate the PLM's ability to perform interview labeling to be complemented by independent human annotation to support the triangulation process. Both humans and algorithms have their biases. Therefore, we have decided to compare our PLM to the performance of a single human annotator, rather than a combination of human choices. The software's annotations were focused on key themes, entities, and sentiments expressed in the interviews, mirroring the tasks typically performed by human annotators given a predetermined set of available annotations. Thirdly, we conducted the "**text Annotation by Human Researcher"** step. Concurrently, a skilled research assistant independently annotated the same set of interviews with the same set of available labels. The research assistant employed standard annotation practices to identify and mark relevant textual elements in the interviews using Atlas.TI. Fourthly, we did the "**Evaluation by Independent Researchers"** step. Then, three researchers evaluated the annotated interviews. These evaluators, unaware of the annotator's identity (software or human), were selected based on their expertise in qualitative data analysis. Fifthly, we conducted the "**Voting Mechanism"** step. Each evaluator independently reviewed the annotations and voted on which version they deemed more accurate and insightful. This

voting was based on criteria such as relevance, comprehensiveness, and clarity of the annotations. Finally, we ran the "**Outcome Assessment"** step. The final assessment was based on the frequency of votes favoring the software's annotations over the human researcher's annotations. If the software's annotations received a majority of the votes more often than the human annotations, it was deemed to have passed the test.

2.3. Automating the text labeling

We- used a combination of three informatics tools (***Appendix 3***). Spyder for automation processes, Atlas.TI for labeling management and OPENAI API to access to GPT 3.5-turbo model. After requesting an Application Programming Interface (API) key to OPENAI, it becomes possible to integrate the GPT-3.5-turbo model into any programming pipeline. In qualitative research aimed at grounded findings, GPT-3.5-turbo offers superior language understanding and generation capabilities compared to BERT, allowing for richer insights from textual data (Brown et al., 2020). GPT-3.5-turbo handles longer contexts and produces coherent text completions, aiding in the development of grounded theories while its ability to learn from natural language instructions reduces model training complexity (Brown et al., 2020). Additionally, its user-friendly interactions facilitate collaboration among stakeholders, essential for qualitative research methodologies (Brown et al., 2020).

3. Results

The primary criterion for declaring the software as mature enough for production deployment was its performance relative to the human annotator. If the software consistently received higher accuracy votes from the independent evaluators, it was considered ready for practical application in annotating similar types of interviews. The intercoder reliability (ICR) at the label level between the AI and the human annotator was low = (mean 3%, SD = 2.9%). However, the ICR at the phylum level between the AI and the human level obtained good results = (mean 91%, SD = 1.5%). The AI annotator passed the AT-TEST for two out of the three interviews, demonstrating a substantial level of accuracy and reliability in its annotations. The successful results for Interviews 1 and 3 indicate that the AI can perform at a level comparable to, or even surpassing, human annotators in certain contexts. The failure in interview 2 highlights the AI's limitations in handling more complex or nuanced data, suggesting areas for further refinement and training (***Table 1***).

Table 1: Summary of the AT-TEST results over the 3 interviews

	Votes (between version A and B of each interview) *The three researchers did not know which version was coded by the AI and the human*					
	Researcher 1	**Researcher 2**	**Researcher 3**	**Majority**	**AI interview**	**AT-TEST RESULT for the AI**
Interview 1	B	B	B	B (3/3)	B	Passed
Interview 2	A	A	B	A (2/3)	B	Failed
Interview 3	A	B	A	A (2/3)	A	Passed

These results are significant as they validate the AI's potential to automate the annotation process in qualitative research, thereby saving time and resources while maintaining a high

standard of accuracy. The AT-TEST outcomes support the integration of AI tools in qualitative research, particularly for large-scale data annotation tasks where efficiency and consistency are paramount. We have also run additional analyses, focusing on several key metrics: the number of unique labels used, the number of quotes labeled, the mean number of words labeled per quote, interview label coverage, and the inter-coder reliability (ICR) at both the label and family levels (***Appendix 3***).

4. Conclusion

Performing Few-Shot Learning in conjunction with a Prompt-Based Learning approach presents a novel pathway to enhance the labeling efficiency of sentences in textual labeling tasks for qualitative research geared towards grounded findings. Few-Shot Learning, a paradigm in machine learning, allows models to make accurate predictions with minimal labeled examples, addressing the scarcity of annotated data which often plagues qualitative analyses (Wang et al., 2020). In leveraging this approach, researchers in marketing (but not only) can train models on a small, representative dataset, ensuring that the model can generalize from these examples to unseen data effectively. Meanwhile, Prompt-Based Learning, an emergent framework inspired by pre-trained language models, involves framing tasks as a series of prompts that these models understand, effectively bridging the gap between the task format and the pre-training objective of the language models (Liu et al., 2021). Integrating the two methods enables the creation of a synergistic system, where the ability of Few-Shot Learning to operate with limited examples complements the natural language understanding of Prompt-Based Learning. This methodology potentially allows for nuanced detection of labels within sentences, pivotal in extracting meaningful patterns and themes from qualitative data (Brown el al., 2020). Furthermore, by employing Prompt-Based Learning, researchers can use the pre-existing knowledge encapsulated in models like GPT-3 to contextualize and identify labels with more substantial semantic understanding (Radford et al., 2019). The use of these combined approaches in qualitative research could not only speed up data labeling but also enhance the reliability of grounded findings, which are essential in the qualitative research paradigm (Charmaz and Belgrave, 2012). In summary, the employment of Few-Shot and Prompt-Based Learning mechanisms exhibits promise for improving sentence-level labeling in text analysis tasks, facilitating the extraction of profound insights from unstructured data for marketers.

However, this research is not exempt from limitations. When given multiple inputs, the GPT-3.5-turbo model is good at identifying small chunks of text, but it is not mature enough to identify longer segments. It is recommended to provide GPT-3.5-turbo with small chunks of paragraphs because it is not yet able to identify all the relevant passages for each label of the full interview in a single request. Processing the text paragraph by paragraph forces the algorithm to find the most relevant text segment for each label. Moreover, it is crucial that the human judges do not become experts themselves in the labeling performance of LLMs. Indeed, providing judges with insights about the LLMs' capabilities will indirectly teach them to recognize the BOT's performances instead of objectively judging the best performance between the human and the BOT. Therefore, it is important that the judges remain unaware of the LLMs' labeling capabilities until the automatic labeling is completely done. Finally, when we started our study, there were no available models that could be downloaded onto a local machine to preserve data confidentiality, which further required programmers to pass the data through an API. However, nowadays, thanks to advances in pruning and quantization, there are plenty of models that are compressed enough to be downloaded, such as Mistral7B. These models can be used on a local machine, which enhances the data confidentiality of the participants.

In conclusion, PLMs are extremely powerful tools for pushing the boundaries of digital marketing research, enabling deeper and more accurate analyses, as well as the efficient manipulation of large amounts of textual data.

Bibliography

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological methods & research*, *42*(3), 294-320.

Chang, T., DeJonckheere, M., Vydiswaran, V. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating mixed methods research with natural language processing of big text data. *Journal of Mixed Methods Research*, *15*(3), 398-412.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. sage.

Charmaz, K., & Belgrave, L. (2012). Qualitative interviewing and grounded theory analysis. *The SAGE handbook of interview research: The complexity of the craft*, *2*, 347-365. Chubb, L. A. (2023). Me and the Machines: Possibilities and Pitfalls of Using Artificial Intelligence for Qualitative Data Analysis. International Journal of Qualitative Methods, 22.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, *13*(1), 3-21.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Crowston, K., Allen, E. E., & Heckman, R. (2012). Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, *15*(6), 523-543.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fang, C., Markuzon, N., Patel, N., & Rueda, J. D. (2022). Natural language processing for automated classification of qualitative data from interviews of patients with cancer. *Value in Health*, *25*(12), 1995-2002.

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social problems*, *12*(4), 436-445.

Glaser, B., & Strauss, A. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: methodological study. *Journal of medical Internet research*, *20*(6), e231.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods*, *18*(1), 59-82.

Hartmann, J., & Netzer, O. (2023). Natural language processing in marketing. In *Artificial intelligence in marketing* (Vol. 20, pp. 191-215). Emerald Publishing Limited.

Heaven, Will Douglas. "OpenAI's new language generator GPT-3 is shockingly good—and completely mindless." *MIT Technology Review* 29 (2020): 1-6.

Humphreys, A., & Wang, R. J. H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274-1306.

Jones, K. S., & Galliers, J. R. (1995). Evaluating natural language processing systems: An analysis and review.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. sage.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, *21*(140), 1-67.

Richards, T. J., & Richards, L. (1994). Using computers in qualitative research. *Handbook of qualitative research*, *2*(1), 445-462.

Shankar, V., & Parsana, S. (2022). An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. *Journal of the Academy of Marketing Science*, *50*(6), 1324-1350.

Strauss, A., & Corbin, J. (1990). Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Sage.

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC medical research methodology*, *18*, 1-18.

Wang, X., He, J., Curry, D. J., & Ryoo, J. H. (2022). Attribute embedding: Learning hierarchical representations of product attributes from consumer reviews. *Journal of Marketing*, *86*(6), 155-175.

Appendix 1: NLP et LLMs benchmarking

Benchmarking of NLP methods for automatic text annotation

Approach	Description	Advantages	Disadvantages	Sources
Named Entity Recognition (NER)	Utilizes pre-trained models to identify named entities in text (names, places, etc.).	Well-suited for specific tasks like named entity recognition.	Less effective for general or contextual classification.	Honnibal et al. (2020)
Fine-Tuning Transformers	Adapts pre-trained models (e.g., BERT, GPT-3) on a small number of examples for a specific task.	Leverages powerful models with few examples.	Requires significant computational resources.	Devlin et al. (2018)
Zero-Shot Learning with LLMs	Uses language models like GPT-3 with task descriptions to generate labels without specific examples.	No need for task-specific data.	May be less accurate for highly specific tasks.	Brown et al. (2020)
Few-Shot Learning with LLMs	Uses language models to classify sentences by providing a few examples for each class.	Allows classification with very few specific examples.	Quality may vary depending on the complexity of the data.	Schick et al. (2021)
Support Vector Machines (SVM) with Embeddings	Uses embeddings vectors for classification with an SVM, suitable for text classification tasks.	Effective with well-trained embedding representations.	Less effective for very complex text contexts.	Cortes & Vapnik (1995)
K-Nearest Neighbors (KNN) with Embeddings	Classifies sentences based on nearest neighbors in an embedding space.	Simple to implement and interpret.	Can be slow and less effective with large datasets.	Cover & Hart (1967)

Benchmarking of LLMs performances

Modèle	Taille du Modèle (Paramètres)	Précision (Benchmark GLUE)	Vitesse d'Inférence (ms)	Utilisation Mémoire (GB)	Efficacité Énergétique
GPT-4	1T+	90%	500	80	Moyenne
GPT-3.5	175B	87%	400	16	Élevée
BERT Large	340M	82%	120	12	Très élevée
LLaMA 2 (13B)	13B	85%	150	8	Élevée
PaLM 2	540B	89%	350	32	Moyenne

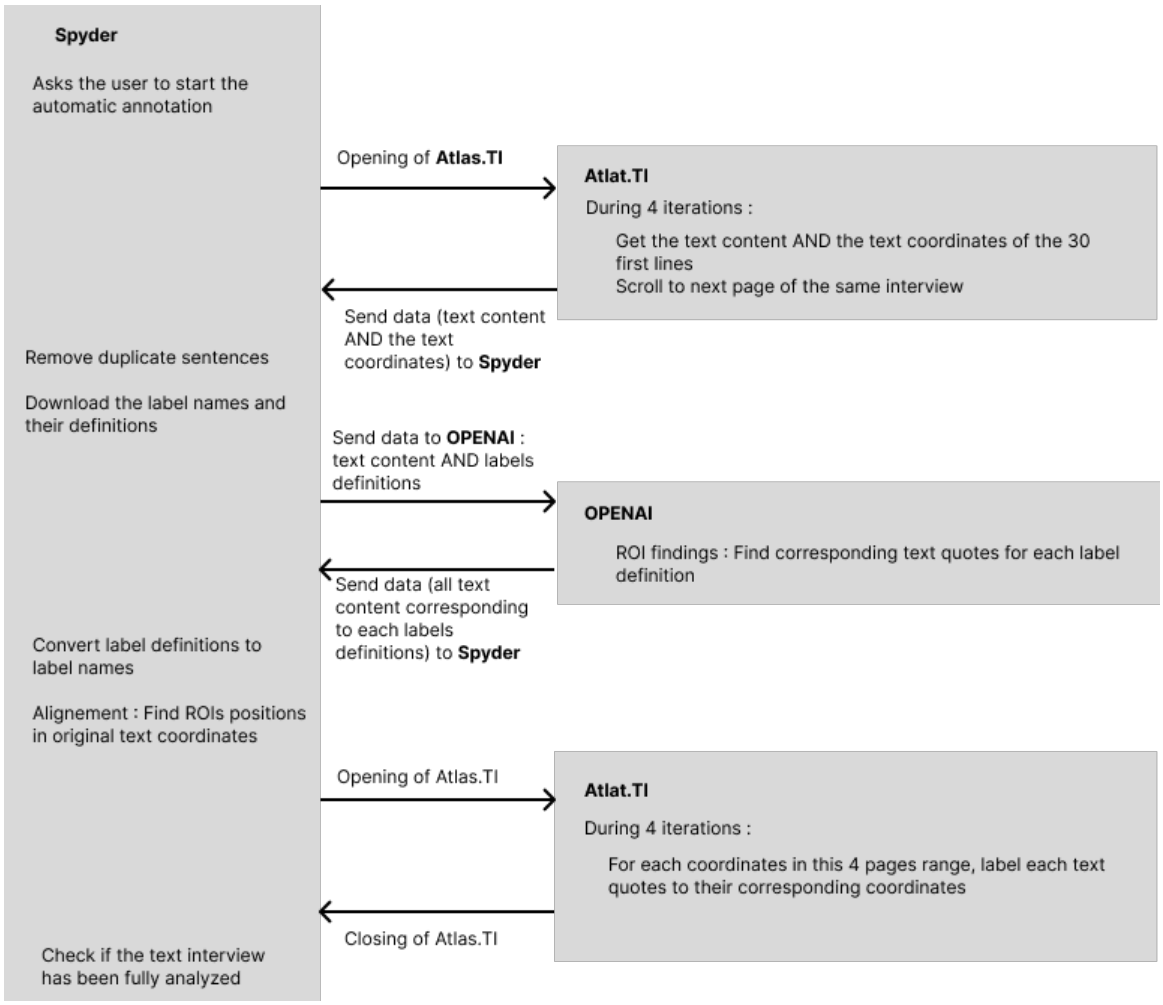
Source : https://paperswithcode.com/task/language-modelling

Appendix 2: Participants information

Interviewee Id	Gender	Age	Interview Date	Interview duration (min)
1	Male	21	22-oct-22	70
2	Female	32	22-oct-22	55
3	Female	21	19-oct-22	40
4	Female	44	19-May-22	35
5	Female	53	11-nov-22	55
6	Female	18	13-nov-22	35
7	Male	26	14-nov-22	47
8	Female	27	07-nov-22	36
9	Female	22	13-oct-22	38
10	Male	22	18-oct-22	50
11	Female	21	13-oct-22	80
12	Female	21	15-oct-22	100
13	Female	43	11-oct-22	60
14	Female	22	13-oct-22	80
15	Male	35	02-oct-22	73
16	Female	39	01-nov-22	70
17	Female	45	04-nov-22	45
18	Male	53	27-oct-22	70
19	Female	49	09-nov-22	80
20	Male	41	01-nov-22	65
21	Female	76	22-oct-22	25
22	Female	51	20-oct-22	55
23	Male	48	06-nov-22	59
24	Female	27	19-oct-22	40
25	Male	21	15-oct-22	52
26	Male	28	23-oct-22	43
27	Male	25	12-nov-22	41
28	Male	23	10-nov-22	45
29	Female	21	07-nov-22	48
30	Female	41	13-nov-22	47
31	Female	56	02-nov-22	63
32	Male	41	02-nov-22	47
33	Male	23	14-oct-22	65
34	Male	15	23-oct-22	56
35	Male	24	18-oct-22	45
36	Male	27	14-mars-22	91
37	Male	22	23-oct-22	46
38	Female	26	03-nov-22	91
39	Female	21	10-nov-22	56
40	Female	15	28-oct-22	55
41	Female	15	06-nov-22	65
42	Female	24	10-nov-22	45
43	Female	39	19-oct-22	50
44	Female	48	22-oct-22	62

45	Female	26	19-oct-22	103
46	Female	23	21-oct-22	30
47	Male	20	20-oct-22	30
48	Male	73	08-nov-22	60
49	Female	36	11-nov-22	50
50	Female	56	22-nov-22	50
51	Male	50	13-nov-22	30
52	Female	26	10-nov-22	60
53	Male	23	20-oct-22	25
54	Female	23	19-oct-22	105
55	Female	20	22-May-22	50
56	Male	22	01-nov-22	75
57	Female	47	25-oct-22	47
58	Male	21	11-nov-22	39
59	Female	48	09-nov-22	30
60	Female	26	11-nov-22	43
61	Male	28	28-sept-22	41
62	Female	26	28-sept-22	33
63	Female	22	20-oct-22	58
64	Female	15	27-oct-22	55
65	Female	46	05-nov-22	46
66	Male	28	26-oct-22	70
67	Female	37	27-oct-22	75
68	Male	54	17-nov-22	33
69	Male	42	18-nov-22	48
70	Female	25	18-nov-22	38
71	Female	20	21-nov-22	30

Appendix 3: Overview of the tools used at each stage to label the corpus



Appendix 4: Additional results

Number of Unique Labels Used: Across the three interviews, the AI annotator used an average of 34 unique labels (SD = 3), while the human annotator used an average of 26 unique labels (SD = 8). The AI's higher average number of unique labels suggests a broader categorization capability, likely due to its extensive training on diverse datasets. However, the human annotator's variability indicates a more selective but potentially more nuanced application of labels.

Number of Quotes Labeled: The AI consistently labeled more quotes than the human annotator, with an average of 97 quotes per interview (SD = 29) compared to human annotator's 66 quotes per interview (SD = 33). This discrepancy underscores the AI's efficiency and thoroughness in identifying relevant segments within the text. The AI's ability to cover more ground can be advantageous in comprehensive data analysis, providing a wider scope of information to be examined.

Mean Number of Words Labeled Per Quote: The mean number of words labeled per quote by the AI was 67 (SD = 20), while it was 120 (SD = 16) for the human annotator. Human annotator's higher average suggests a tendency to label larger text segments, which could be due to a preference for capturing broader contexts. In contrast, the AI's more concise labeling may reflect a focus on specific keywords or phrases that trigger certain labels, indicating a different approach to information extraction.

Interview's Label Coverage: Label coverage is a critical metric as it indicates the comprehensiveness of the labeling process. The AI demonstrated an average label coverage of 89% (SD = 21) across the interviews, whereas the human annotator had an average of 107% (SD = 43). Human annotator's higher coverage percentage can be attributed to overlapping labels and re-annotation practices that are common in human analysis. The AI's lower but still high coverage percentage suggests efficient use of unique labels with minimal redundancy.

	Number of unique labels used (42 labels in total)		**Number of quotes labeled**		**Mean number of words labeled per quote**		**Interviewee's label coverage**		ICR (LABEL LEVEL)	ICR (PHYLUM LEVEL)
	AI	HA	AI	HA	AI	HA	AI	HA		
Interview 1	30	15	69	22	51	100	38%	23%	0%	91%
Interview 2	37	35	84	101	96	119	73%	76%	7%	93%
Interview 3	36	29	137	74	54	140	65%	90%	2%	89%
MEAN	34 (SD=3)	26 (SD=8)	97 (SD=29)	66 (SD=33)	67 (SD=20)	120 (SD=16)	89% (SD=21)	107% (SD=43)	3% (SD=2.9)	91% (SD=1.5)