

# CommentSense\*Meetup Edition

An On-Device AI Browser Extension for  
Real-time YouTube Comment  
Understanding

Marc Kroll

Advisor: Assistant Prof. Dr.in phil. Mag.a phil. Astrid Weiss  
Assistance: Dipl.-Ing. Rafael Vrecar, BSc

# Available Models<sup>[1]</sup>

April, 2024

**Llama-3-8B & 70B**  
**Phi-3-Mini**

May, 2024

**Phi-3 small/medium**

June, 2024

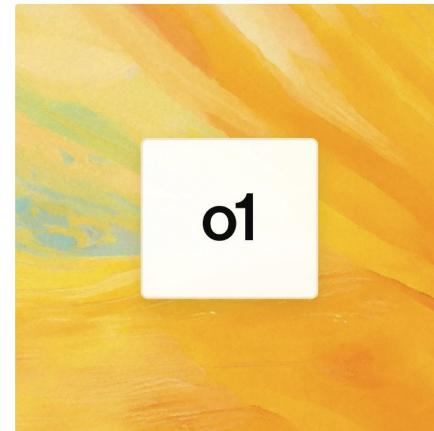
**Gemma 2**

December, 2024

**Phi-4**

September, 2024

**o1-preview & o1-mini** <sup>[2]</sup>



Learning to reason with LLMs

Release

October, 2024

**ChatGPT Search** <sup>[2]</sup>



[1] <https://github.com/eugeneyan/open-llms>

[2] <https://openai.com/o1/>, <https://openai.com/index/introducing-chatgpt-search/>

# “deepseek moment”

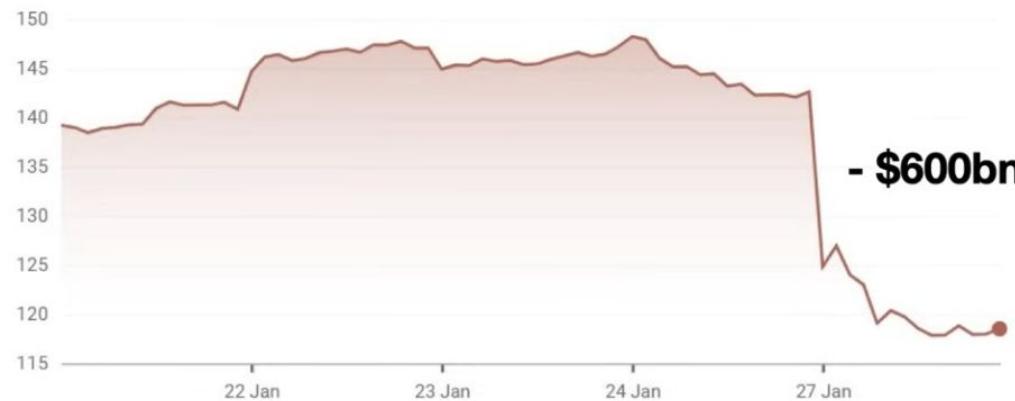
NVIDIA Corp

\$118.58 ↓14.83% -20.64 5 D

After hours: \$121.27 (↑2.27%) +2.69

Closed: 27 Jan, 19:30:04 UTC-5 · USD · NASDAQ · Disclaimer

1 D    5 D    1 M    6 M    YTD    1 Y    5 Y    MAX



# Available Models

January, 2025

**mistral-small-3<sup>[1]</sup>**

March, 2025

**Gemma 3<sup>[2]</sup>**

**mistral-small-3.1<sup>[1]</sup>**

- + DeepSeek V3 & R1
- Qwen 2.5 & Qwen 3

[1] <https://mistral.ai/news/mistral-small-3>, <https://mistral.ai/news/mistral-small-3-1>

[2] <https://blog.google/technology/developers/gemma-3/>

[3] <https://arxiv.org/pdf/2505.05410.pdf>

# Reasoning models don't always say what they think

3 Apr 2025

Read the paper

Since late last year, “reasoning models” have been everywhere. These are AI models—such as Claude 3.7 Sonnet—that *show their working*: as well as their eventual answer, you can read the (often fascinating and convoluted) way that they got there, in what’s called their “Chain-of-Thought”.

As well as helping reasoning models work their way through more difficult problems, the Chain-of-Thought has been a boon for AI safety researchers. That’s because we can (among other things) check for things the model says in its Chain-of-Thought that go unsaid in its output, which can help us spot undesirable behaviours like deception.

But if we want to use the Chain-of-Thought for alignment purposes, there’s a crucial question: can we actually trust what models say in their Chain-of-Thought?

In a perfect world, everything in the Chain-of-Thought would be both understandable to the reader, and it would be *faithful*—it would be a true description of exactly what the model was thinking as it reached its answer.

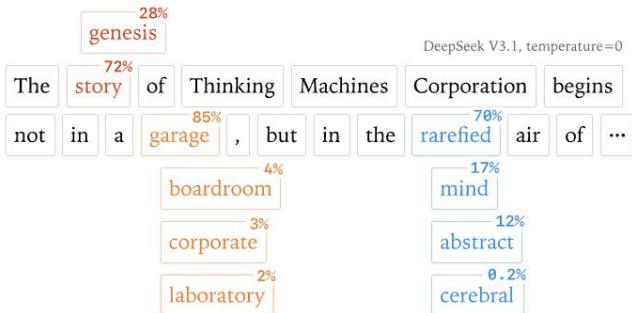
But we’re not in a perfect world. We can’t be certain of either the

# September 2025

## Defeating Nondeterminism in LLM Inference

Horace He in collaboration with others at Thinking Machines

Sep 10, 2025



Reproducibility is a bedrock of scientific progress. However, it's remarkably difficult to get reproducible results out of large language models.

arXiv:2506.02153v2 [cs.AI] 15 Sep 2025

## Small Language Models are the Future of Agentic AI

Peter Belcak<sup>1</sup> Greg Heinrich<sup>1</sup> Shizhe Diao<sup>1</sup> Yonggan Fu<sup>1</sup> Xin Dong<sup>1</sup>  
 Saurav Muralidharan<sup>1</sup> Yingyan Celine Lin<sup>1,2</sup> Pavlo Molchanov<sup>1</sup>  
<sup>1</sup>NVIDIA Research <sup>2</sup>Georgia Institute of Technology  
 agents-research@nvidia.com

### Abstract

Large language models (LLMs) are often praised for exhibiting near-human performance on a wide range of tasks and valued for their ability to hold a general conversation. The rise of agentic AI systems is, however, ushering in a mass of applications in which language models perform a small number of specialized tasks repetitively and with little variation.

Here we lay out the position that small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI*. Our argumentation is grounded in the current level of capabilities exhibited by SLMs, the common architectures of agentic systems, and the economy of LM deployment. We further argue that in situations where general-purpose conversational abilities are essential, heterogeneous agentic systems (i.e., agents invoking multiple different models) are the natural choice. We discuss the potential barriers for the adoption of SLMs in agentic systems and outline a general LLM-to-SLM agent conversion algorithm.

Our position<sup>1</sup>, formulated as a value statement, highlights the significance of the operational and economic impact even a partial shift from LLMs to SLMs is to have on the AI agent industry. We aim to stimulate the discussion on the effective use of AI resources and hope to advance the efforts to lower the costs of AI of the present day. Calling for both contributions to and critique of our position, we commit to publishing all such correspondence at research.nvidia.com/labs/lpr/slm-agents.

### 1 Introduction

The deployment of agentic artificial intelligence is on a meteoric rise. Recent surveys show that more than a half of large IT enterprises are actively using AI agents, with 21% having adopted just within the last year [14]. Aside from the users, markets also see substantial economic value in AI agents: As of late 2024, the agentic AI sector had seen more than USD 2bn in startup funding, was valued at USD 5.2bn, and was expected to grow to nearly USD 200bn by 2034 [46, 51]. Put plainly, there is a growing expectation that AI agents will play a substantial role in the modern economy.

The core components powering most modern AI agents are (very) large language models [52, 48]. It is the LLMs that provide the foundational intelligence that enables agents to make strategic decisions about when and how to use available tools, control the flow of operations needed to complete tasks, and, if necessary, to break down complex tasks into manageable subtasks and to perform reasoning for action planning and problem-solving [52, 17]. A typical AI agent then simply communicates with a chosen LLM API endpoint by making requests to centralized cloud infrastructure that hosts these models [52].

<sup>1</sup>The views and positions expressed in this paper are those of the authors and do not necessarily reflect the views or positions of any entities they represent.

# CommentSense\*Meetup Edition

An On-Device AI Browser Extension for  
Real-time YouTube Comment  
Understanding

Marc Kroll

Advisor: Assistant Prof. Dr.in phil. Mag.a phil. Astrid Weiss  
Assistance: Dipl.-Ing. Rafael Vrecar, BSc

# BEFORE



# AFTER





@russellyork5853  
Why do u look so fat????



@Tai-oLo  
you so fat



@anthonysmith1591  
Grossly obese person telling you what food they don't like is hilarious.



@dannyespinoza3836  
Chill fatty



@Insayenign9422  
Bro you're getting fat!



@nestorrfortuna1  
Getting fat



@TheDancerPL  
bro u got fat:(



@greeneffectltd  
Joshua Fatman is your new name fatso!



@coltonharris1526  
Josh is getting fat



@Neobanned  
Josh, hows the weight loss going you phat! 🤣



@trickyd499 1 year ago  
fatty fatty bum bum



@jernigan007  
you got fat



@beaulaloevv  
notice how he gained weight lol. he should workout



@MrKojotie  
Aloha, now you're fat :P



@Mkantae  
blud is fat now.



@kucci1769 1 year ago  
josh ur getting fat.



@roadrunner1095  
With every video you are getting fatter... Do some exercise man.



# OZEMPIC ?



@fleshtaffy

There's absolutely no doubt this man is on Ozempic.



@SebastianTorres-zt2lm

Damn dude lost weight hope u not on ozempic big dawg



@SierraPapa73

Hey! Really like your channel, but your food judgement I have to question. You're just too darn skinny. You don't have to look like Jellyroll, but if you truly loved food, we'd see proof. 😊



@timothystewart9547

The Ozempic kicking in



@tonybroussard1080

People really gotta stop taking ozempic



@YaBoiBigNutz

Bro looking either sick or on ozempic in that thumbnail damn



@user-yu1pm9vj8j 3 months ago

Ozempic really works.



@patrickmorilus 1 month ago

Yo people have to start admitting they on ozempic.



@scottrhodes8160

Good video—stop losing weight you are getting too skinny.



@callmeviper7723

Bro you are so skinny now are you ok?



@dmanm85

Everyone is on Ozempic and getting skinny and it's freaking me out..



@kindredmathematician4290

Ozempic face?



@tavijoseph9026

Oh snap Josh. U on ozempic?



@joetheunknown8413

Bro is on ozempic.



@TamarainTanzania2

You on the Ozempic?



@MichaelKierBialock

That Ozempic hittin hard boi!



@appleta 2 weeks ago

Remember that episode where Josh took Ozempic.

## Model comparison for the term: “Ozempic face”

### Llama 3.2 3b [4]

**SENTIMENT:**  
NEGATIVE



#### REASONING:

Phrases like Ozempic face are often used to criticize the appearance resulting from weight loss medications, carrying a negative connotation.



- |                   |                     |
|-------------------|---------------------|
| ! Severe Toxicity | ! Identity attack   |
| ! Insult          | ! Threat            |
| ! Profanity       | ! Sexually explicit |

### Perspective API [2]

Toxicity: 16,76%  
Insult: 8.36%



### DistilBERT-sst2 [3]

POSITIVE: 57,89%



BERT-base model fine-tuned for doing binary (positive / negative) sentiment analysis

[2] <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>

[3] <https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english>

[4] <https://huggingface.co/meta-llama/Llama-3.2-3B>

# Are Small Language Models better in sentiment analysis than BERT-based models?

**Hypothesis 1:** SLMs outperform BERT-based models due to their contextual knowledge.

**Hypothesis 2:** Users prefer SLMs due to their reasoning capabilities.



## RELATED WORK

High Level

Content  
Moderation

Fairness in Content  
Classification

Value-Sensitive  
Algorithms

Specific

Hate Speech  
Detection

Irony, Sarcasm,  
Emoji Detection

Personalized  
Filtering

Trends

Human-in-the-Loop  
Systems

Text Classification  
with Language Models

Multimodal Sentiment  
Analysis

# Are Small Language Models better in sentiment analysis than BERT-based models?

## RQ1: (Technical)

How do the accuracy, efficiency, and false positive/negative rates of general-purpose on-device small language models compare to a specialized, fine-tuned BERT language model in doing sentiment analysis of YouTube comments?

## RQ2: (Human-centered)

What are user preferences and the perceived effectiveness of different methods for displaying (grouped, reordered, highlighted) potentially negative YouTube comments in a browser extension, considering the ethical implications of false positives and negatives?

## RQ3: (Bridging technical and human-centered)

How does the use of small language models compare to a state-of-the-art BERT-based model and influence users' perceived usefulness of sentiment analysis systems for YouTube comments?

Step 1

## User Preference

6 Participants  
1h Remote-Session  
Interactive Prototype  
Semi-structured interview  
Note taking

### Mixed-Method approach [5]

#### Quantitative methods:

Performance metrics, Macro F1-Score, Likert-scales in surveys

#### Qualitative methods:

Semi-structured interviews, think aloud, observation, note taking,  
thematic analysis, open questions in surveys

Step 2

## Model Evaluation

19 Small Language Models\*  
1 fine-tuned BERT Model  
Human ground truth  
YouTube dataset  
Binary classification  
Three-way classification  
Performance metrics  
Macro F1-Score

Step 3

## User Evaluation

7 Participants  
1½ h in-person sessions  
Surveys  
User tasks + think aloud  
Observation  
Note taking  
Semi-structured interview  
Thematic analysis

Human ground truth  
YouTube dataset

\* Models had to be below 22b parameters, multilingual, public and for free without any constraints (sign-up, gated, experimental, etc)  
[5] Creswell, J. W. Chapter 18 - Mixed-Method Research: Introduction and Application.

## User Preference (n=6)

Feedback

# Interactive prototype

Don't group!  
Don't hide anything!  
Don't change the appearance!

Ich möchte eher die negativen auch.

Kann ich bestimmen was raus soll?

**Prototype 1**  
Negative comments are not displayed with a disclaimer

**Prototype 2**  
Highlighting and redaction

**Prototype 3**  
Word highlights

**Prototype 4**  
Grouping - closed

You have seen the YouTube comment section where negative comments are automatically removed.

You have seen the YouTube comment section where positive comments are highlighted and negative comments are redacted.

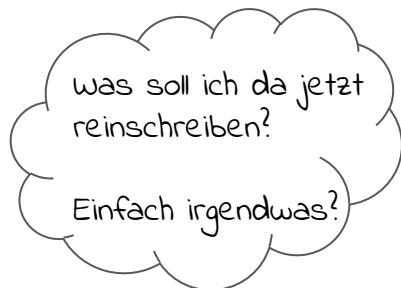
You have seen the YouTube comment section where positive, negative and neutral comments are grouped.

## Iteration 1

Pilot 1

# Custom Classification

Participants struggled defining classes and examples for classification.



The screenshot shows a YouTube video player with a large red 'X' drawn over the video frame. The video title is "Want a Winning Smile? Learn How to Cheat at Cards! #cardgame #cards #cardtrick #blackjack". The YouTube interface includes a search bar, user login, and video controls. To the right is the "Ollama Custom YouTube Comment Analyzer" application window. It displays the video title and hashtags, and includes sections for "Class 1 Name" (e.g., spam), "Class 2 Name" (e.g., question), and "Class 3 Name" (e.g., praise). It also has fields for "Enter examples for this class (one per line)" and a button to "Analyze with Small Language Model". Other settings include "Analysis Summary", "Difficulty", "Filter Comments", and "Include Video Transcript".

## Iteration 2

Pilot 2

# Extension vs Website

Too many options.  
Task sheet too complex.  
Cognitive load was too high.  
Pilot had to be stopped early.



Simplify training,  
user tasks and  
extension.



The screenshot shows a browser window titled "BERT Sentiment Analysis" at "localhost:5173". It features a video thumbnail of a man, channel information ("CardMagicByJason"), and "Comment Controls" for 20 comments. A large red "X" is drawn across the interface. Below the controls is a "Sentiment Analysis Summary" section with three boxes: Positive (60.4%), Neutral (0%), and Negative (39.6%). The main area displays several comments with their sentiment analysis results. A large red "X" is also drawn over the first two comments.

Comment	Sentiment	Confidence (%)
Leaving. Very impressive but getting old. Pe...	Positive	100% confident
&quot;I am so good.&quot; Quid is legendary.	Positive	100% confident
That was impossible.	Negative	99% confident
I think he's using the sound cue of the cards falling to time the grab. Very smooth. Another level.	Positive	68% confident

# Are Small Language Models better in sentiment analysis than BERT-based models?

## RQ1: (Technical)

How do the accuracy, efficiency, and false positive/negative rates of general-purpose on-device small language models compare to a specialized, fine-tuned BERT language model in doing sentiment analysis of YouTube comments?

## RQ2: (Human-centered)

What are user preferences and the perceived effectiveness of different methods for displaying (grouped, reordered, highlighted) potentially negative YouTube comments in a browser extension, considering the ethical implications of false positives and negatives?

Simplicity and maximum transparency. No reordering, grouping or highlighting. Negative content is preferred to be displayed instead of hidden.

## RQ3: (Bridging technical and human-centered)

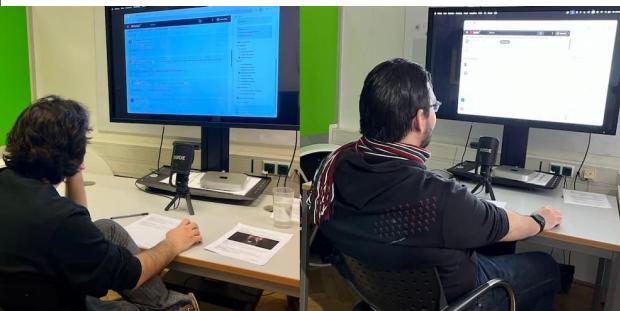
How does the use of small language models compare to a state-of-the-art BERT-based model and influence users' perceived usefulness of sentiment analysis systems for YouTube comments?

# User Evaluation (n=7)

3 Tools

## Study setting

Training video  
Pre-task survey  
User tasks + think aloud  
Post-task survey  
Interview



Note: All data was processed locally.

The screenshot shows a YouTube video titled "Want a Winning Smile? Learn How to Cheat at Cards! #cardgame #cards #cardtrick #blackjack" by CardMagicByJason. The video has 267,000 subscribers and 53,475 views. The interface includes a search bar, login options, and a sidebar for "Tool C".

**Analysis Summary:**  
Total Comments: 20

**Sentiment Distribution:**  
Constructive Criticism: 4  
Praise/Appreciation: 16

**Tone Distribution:**  
Neutral/Informative: 10  
Sarcastic/Ironic: 8  
Humorous/Playful: 7  
Enthusiastic/Hyperbolic: 15

**Special Flags:**  
Potential backhanded compliment: 1

**Filter Comments:**  
By Sentiment:  
 Show All  
 Constructive Criticism  
 Praise/Appreciation  
By Tone:  
 Neutral/Informative  
 Sarcastic/Ironic  
 Humorous/Playful  
 Enthusiastic/Hyperbolic  
By Special Flags:  
 Potential backhanded compliment  
 Include Video Transcript  
 Video Transcript

**YouTube Controls:**  
 Enable Cinema Mode  
 Hide Suggested Videos

The main content area displays several comments with their respective sentiment, tone, and special flags analysis. One comment is highlighted with a yellow background.

**Comments:**

- Von CardMagicByJason angepinnt vor 5 Tagen  
Leaving. Very impressive but getting old. Peace out.  
Antworten  
• 25 Antworten
- He even trash talks himself, truly no one is safe vor 5 Tagen  
Antworten  
Praise/Appreciation Humorous/Playful Sarcastic/Ironic Potential backhanded compliment
- It doesn't matter how many times I watch these videos, or how many times I think I might've seen something... This guy right here is hands down the best I've ever seen. It's not even close. And he seems like he'd be a helluva guy to hang out with! Thanks for all the awesome content!  
Antworten  
Praise/Appreciation Enthusiastic/Hyperbolic Humorous/Playful

# Final Extension:

## Side panels

### Tool A:

- Distilbert-sst2

### Tool B & Tool C:

- Mistral-small 3

Summary and Filter Options.

Context of the Video (transcript) can be included.

The image displays three separate windows labeled Tool A, Tool B, and Tool C, each showing a different interface for video analysis. All three tools share a common layout with a top header, a main content area, and a bottom YouTube Controls section.

**Tool A Header:** Tool A

**Tool A Content:** Video Title: Want a Winning Smile? Learn How to Cheat at Cards! #cardgame #cards #cardtrick #blackjack

**Tool A Buttons:** Analyze

**Tool A Analysis Summary:** Total Comments: 20  
Positive: 14  
Neutral: 0  
Negative: 6

**Tool A Filter Comments:**

- Show All
- Show Positive Only
- Show Neutral Only
- Show Negative Only

**Tool A Video Transcript Options:**

- Include Video Transcript
- Video Transcript

**Tool A YouTube Controls:**

- Enable Cinema Mode
- Hide Suggested Videos

**Tool B Header:** Tool B

**Tool B Content:** Video Title: Want a Winning Smile? Learn How to Cheat at Cards! #cardgame #cards #cardtrick #blackjack

**Tool B Buttons:** Analyze

**Tool B Analysis Summary:** Total Comments: 20  
Positive: 15  
Neutral: 2  
Negative: 3

**Tool B Filter Comments:**

- Show All
- Show Positive Only
- Show Neutral Only
- Show Negative Only

**Tool B Video Transcript Options:**

- Include Video Transcript
- Video Transcript

**Tool B YouTube Controls:**

- Enable Cinema Mode
- Hide Suggested Videos

**Tool C Header:** Tool C

**Tool C Content:** Video Title: Want a Winning Smile? Learn How to Cheat at Cards! #cardgame #cards #cardtrick #blackjack

**Tool C Buttons:** Analyze

**Tool C Analysis Summary:** Total Comments: 20

**Tool C Sentiment Distribution:** Praise/Appreciation: 15  
Constructive Criticism: 4  
Cathartic Release: 1

**Tool C Tone Distribution:** Neutral/Informative: 5  
Ambiguous/Uncertain: 1  
Emotional/Vulnerable: 1  
Sympathetic/Supportive: 7  
Humorous/Playful: 5  
Enthusiastic/Hyperbolic: 11  
Confused/Disoriented: 2  
Casual/Informal: 4

**Tool C Special Flags:** Multi-layered meaning: 1  
Potential backhanded compliment: 2

**Tool C Filter Comments:**

**By Sentiment:**

- Show All
- Praise/Appreciation
- Constructive Criticism
- Cathartic Release

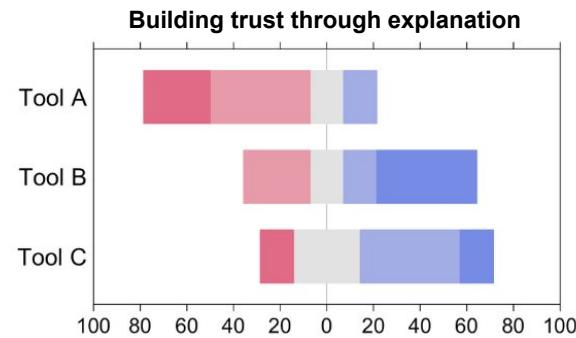
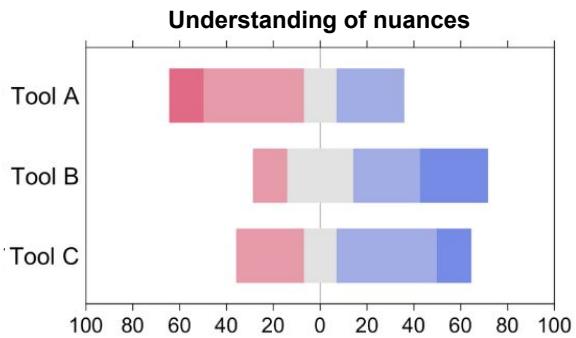
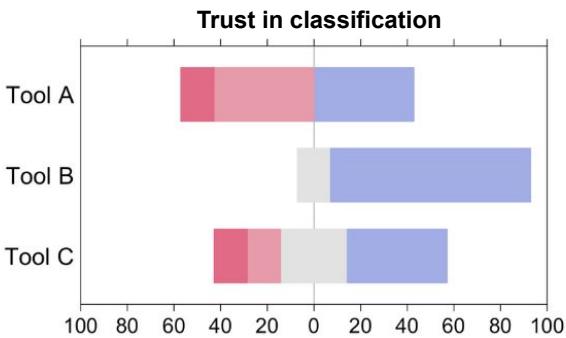
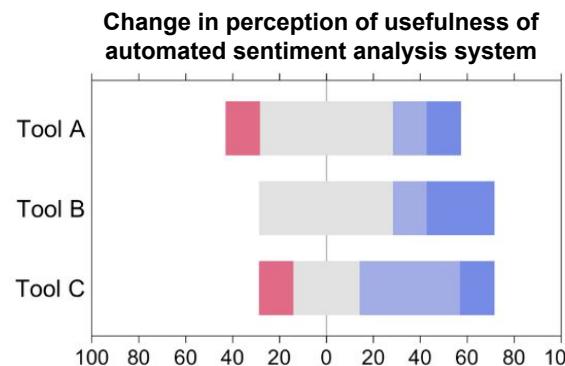
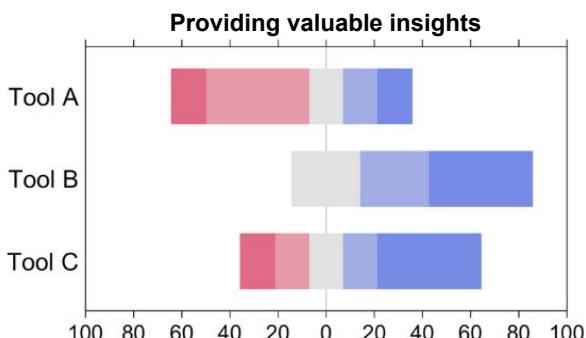
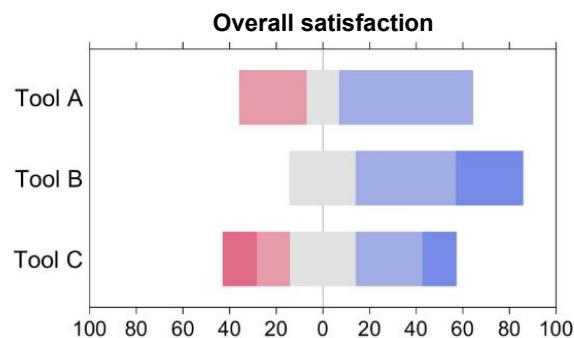
**By Tone:**

- Neutral/Informative
- Ambiguous/Uncertain
- Emotional/Vulnerable
- Sympathetic/Supportive
- Humorous/Playful

**YouTube Controls:**

- Enable Cinema Mode
- Hide Suggested Videos

## User Evaluation: post-task survey

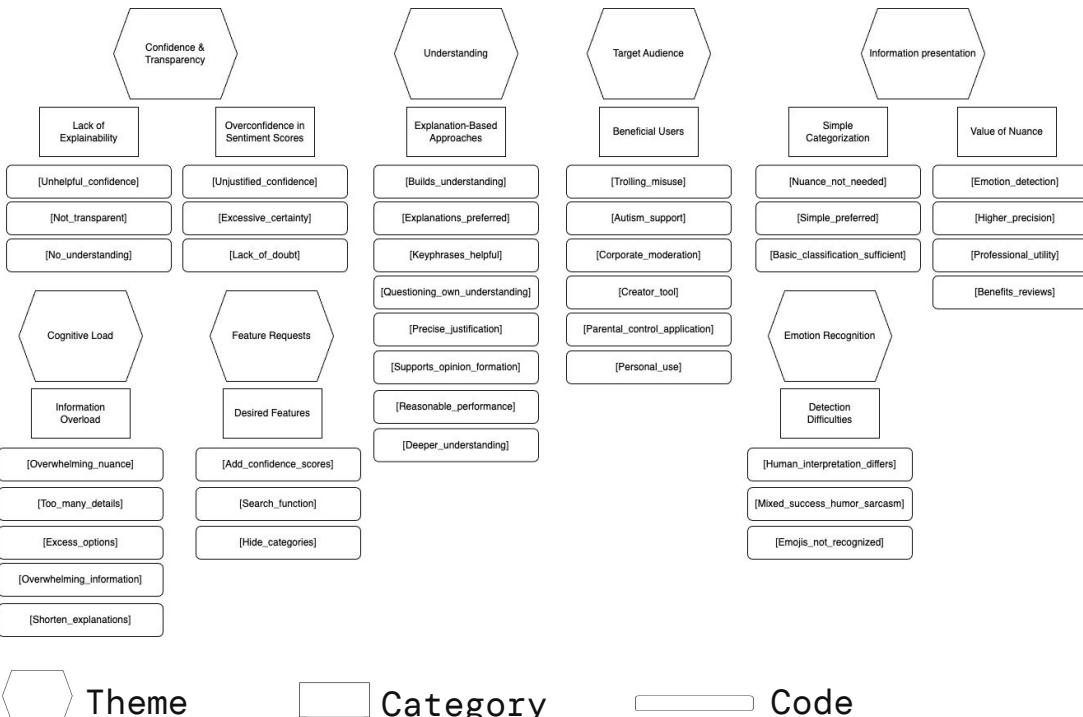


Strongly disagree      Strongly agree

**Note:** all X-axis in percent (%). Label was removed to enhance readability in the presentation.

Tool A: BERT-based — Tool B & C: SLM-based.

# Thematic analysis [6]



## Finding 1

Nuanced and detailed classifications were not received well by most participants due to high cognitive load. Simplicity is preferred.

## Finding 2

Explanations for classifications were preferred from all participants compared to confidence scores.

## Finding 3

The extensions are suggested to be useful for parental control, YouTube creators or to support people with autism to better understand emotional cues in language.

# Are Small Language Models better in sentiment analysis than BERT-based models?

## RQ1: (Technical)

How do the accuracy, efficiency, and false positive/negative rates of general-purpose on-device small language models compare to a specialized, fine-tuned BERT language model in doing sentiment analysis of YouTube comments?

## RQ2: (Human-centered)

What are user preferences and the perceived effectiveness of different methods for displaying (grouped, reordered, highlighted) potentially negative YouTube comments in a browser extension, considering the ethical implications of false positives and negatives?

Simplicity and maximum transparency. No reordering, grouping or highlighting. Negative content is preferred to be displayed instead of hidden.

## RQ3: (Bridging technical and human-centered)

How does the use of small language models compare to a state-of-the-art BERT-based model and influence users' perceived usefulness of sentiment analysis systems for YouTube comments?

The SLM based approach positively changed the perceived usefulness of a sentiment analysis system due to explanations. The BERT-based approach was perceived neutral.

## Model Evaluation

### Macro F1-Score

# Binary Classification

Measuring correct classification of positive, negative against human ground truth.

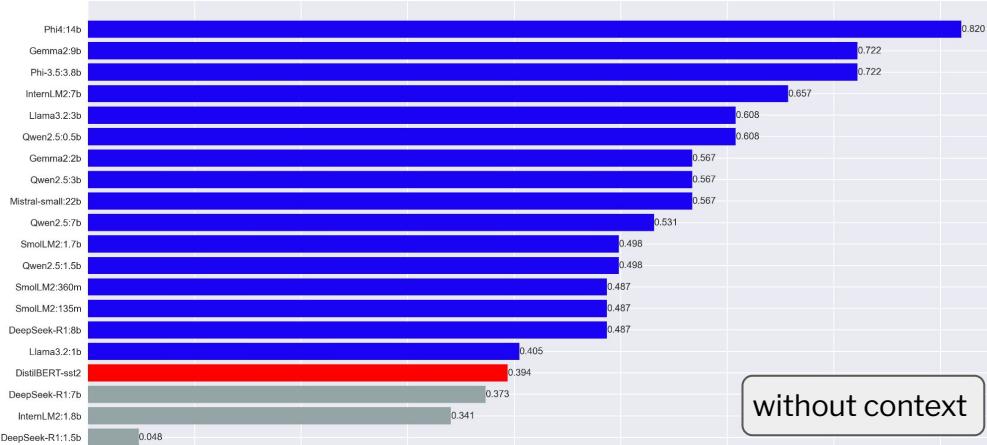
#### Best:

phi4:14b, gemma2:9b, mistral-small:22b

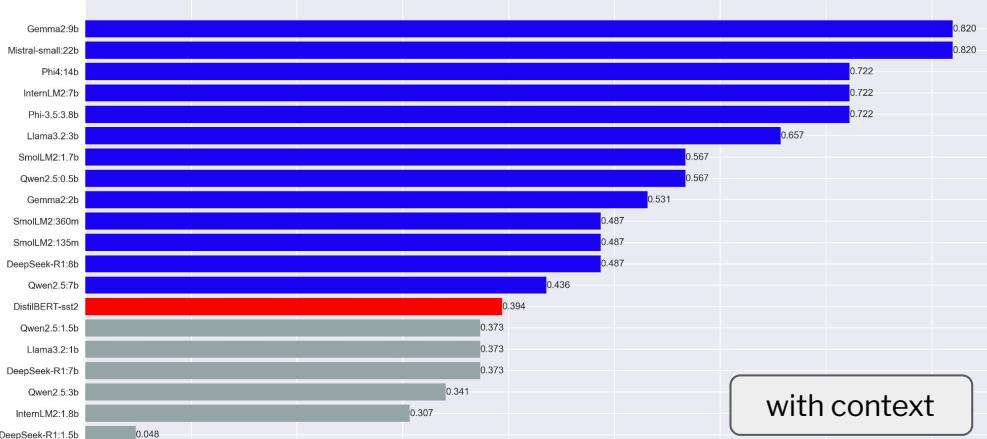
#### Worst:

internlm2:1.8b, deepseek-r1:1.5b

### Macro F1-Score



without context



with context

**Notes:** Model evaluation was performed with fixed seed and temperature.

Human annotation was adjusted for binary classification by the researcher.

DistilBERT-sst2 was not capable of taking the evaluated context into account (token limit)

## Model Evaluation

Macro F1-Score

# Three-way Classification\*

Measuring correct classification of positive, negative, neutral against human ground truth.

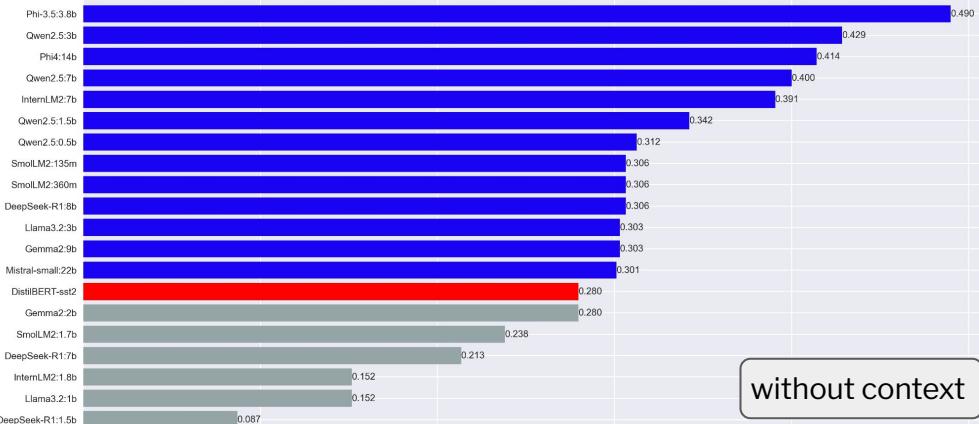
**Best:**

phi3.5:3.8b, mistral-small:22b

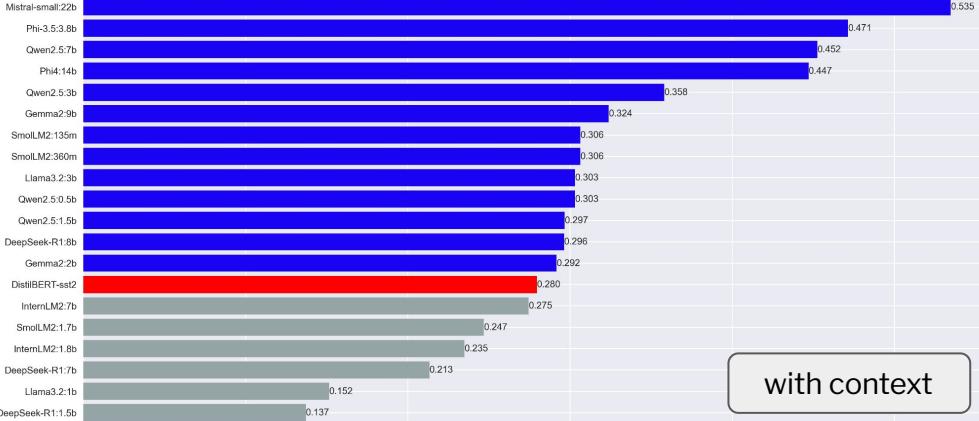
**Worst:**

llama3.2:1b, deepseek-r1:1.5b

## Macro F1-Score



without context



with context

**Note:** DistilBERT-sst2 is only able to do binary classification

\*Users stated: positive/negative is not enough. Therefore "neutral" was introduced as class.

# Are Small Language Models better in sentiment analysis than BERT-based models?

## RQ1: (Technical)

How do the accuracy, efficiency, and false positive/negative rates of general-purpose on-device small language models compare to a specialized, fine-tuned BERT language model in doing sentiment analysis of YouTube comments?

Over 50% of all evaluated SLMs outperformed DistilBERT-sst2 based on Macro F1-Scores against the human ground truth for a dedicated dataset.

## RQ2: (Human-centered)

What are user preferences and the perceived effectiveness of different methods for displaying (grouped, reordered, highlighted) potentially negative YouTube comments in a browser extension, considering the ethical implications of false positives and negatives?

Simplicity and maximum transparency. No reordering, grouping or highlighting. Negative content is preferred to be displayed instead of hidden.

## RQ3: (Bridging technical and human-centered)

How does the use of small language models compare to a state-of-the-art BERT-based model and influence users' perceived usefulness of sentiment analysis systems for YouTube comments?

The SLM based approach positively changed the perceived usefulness of a sentiment analysis system due to explanations. The BERT-based approach was perceived neutral.

**For a sentiment analysis tool ..**

**.. to be perceived as trustworthy and useful,  
transparency in its classifications is essential,  
along with a simple and unobtrusive display method.**

Concerns included over-filtering, bias, free speech, and privacy, with a focus on transparency. The tool must balance accuracy, control, and transparency for acceptance.

Nuanced and detailed classifications were not received well by most participants.

Extensions using SLMs ranked highest for transparency, usefulness, and explanation quality, enhancing user understanding of YouTube comment sentiment and driving acceptance.

Traditional sentiment analysis approaches should be reconsidered.

Explanations for classification results were preferred by every participant instead of confidence scores.

Best performing SLMs\* sorted by parameter size: phi3.5:3.8b - gemma2:9b - phi4:14b - mistral-small:22b  
SLMs from the phi, gemma and mistral series outperformed DistilBERT-sst2 in all tests.

\* General purpose SLMs are evaluated without any dedicated fine-tuning for sentiment analysis.

**Note:** always the latest release of available SLMs should be evaluated.

## Sources

- [1] Joshua Weissman. Can You Trust A Skinny Chef (How I Lost 60 Lbs), Sept. 2024. <https://www.youtube.com/watch?v=4-XPa09H1Xs>
- [2] <https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages>
- [3] <https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english>
- [4] <https://huggingface.co/meta-llama/Llama-3.2-3B>
- [5] Creswell, J. W. Chapter 18 - Mixed-Method Research: Introduction and Application. In *Handbook of Educational Policy*, G. J. Cizek, Ed., *Educational Psychology*. Academic Press, San Diego, Jan. 1999, pp. 455-472.
- [6] Braun, V., and Clarke, V. Conceptual and design thinking for thematic analysis. *Qualitative psychology* 9, 1 (2022), 3

## CONTRIBUTIONS

A custom Chrome extension enabling users to perform sentiment analysis while staying on YouTube, supporting the field of user-centered content moderation while comparing a BERT-based model with SLMs with a focus on the user as the central recipient of the results.

**(Technical)**

An open-source Chrome extension for YouTube comment moderation, based on insights from user preferences for content filtering.

**(Technical)**

Insights to the feasibility of locally running language models in sentiment analysis are documented.

**(Human-centered)**

An understanding of user preferences and needs in content visualization regarding YouTube comments.

**(Human-centered)**

An understanding of the perceived usefulness when applying a SLM for sentiment analysis instead of a BERT-based model.

Thanks.

26.03.2025



'Perspective Team' via pe... Montag  
An: undisclosed-recip ... & 1 weitere >  
Antwort an: Perspective Team >

## [Perspective API] Customizable attributes beta in Perspective API

Dear Perspective users and friends,

We recently announced [customizable attributes in Perspective API](#). Users can provide their community's rules in plain English, as well as example comments and decisions, and receive a score that reflects whether a comment complies with their rules. This capability, powered by the latest Gemini models from Google, enables users to detect the comments they care about, which they can then label or analyze.

Because we still have a lot to learn, we're inviting developers and researchers to [join us in a limited beta program to experiment with these customizable attributes](#). This collaborative approach will allow us to provide technical support, gather valuable feedback, and iterate on our research in real-world scenarios.

As always, let us know of any questions by contacting us [here](#).

Thank you,

