Adevinta

Towards Improving Image Quality in Second-Hand Marketplaces with LLMs

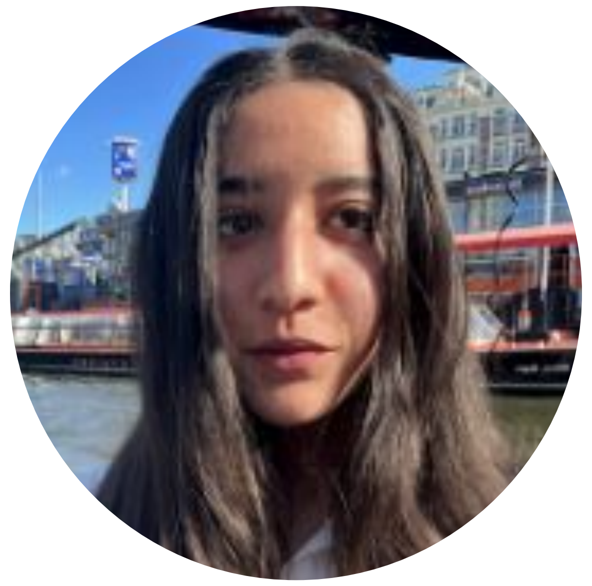
Victor Codina, **Sandra Esparza** (Adevinta's Data Scientists)

Salma Lahbiss, Imene Sederi and **Nafi Cissé** (Leboncoin UX Design & Research)

Authors

Sandra Esparza
Data Scientist
(Adevinta)

Victor Codina
Data Scientist
(Adevinta)



Salma Lahbiss
User Researcher
(Leboncoin)

Imene Sederi
UX Designer
(Leboncoin)

Nafi Cissé
UX Designer
(Leboncoin)

Changing Commerce *Together*

Adevinta is a leading online classifieds group, operating digital **marketplaces** across Europe and beyond.

Our Marketplaces

subito
kleinanzeigen
leboncoin
Marktplaats
InfoJobs
fotocasa
automobile.it
avendre alouer
kijiji
mobile.de
locasun
coches.net

120 million

monthly users

2.5 billion

monthly visits

Adevinta

adevinta.com

Problem space and Motivation

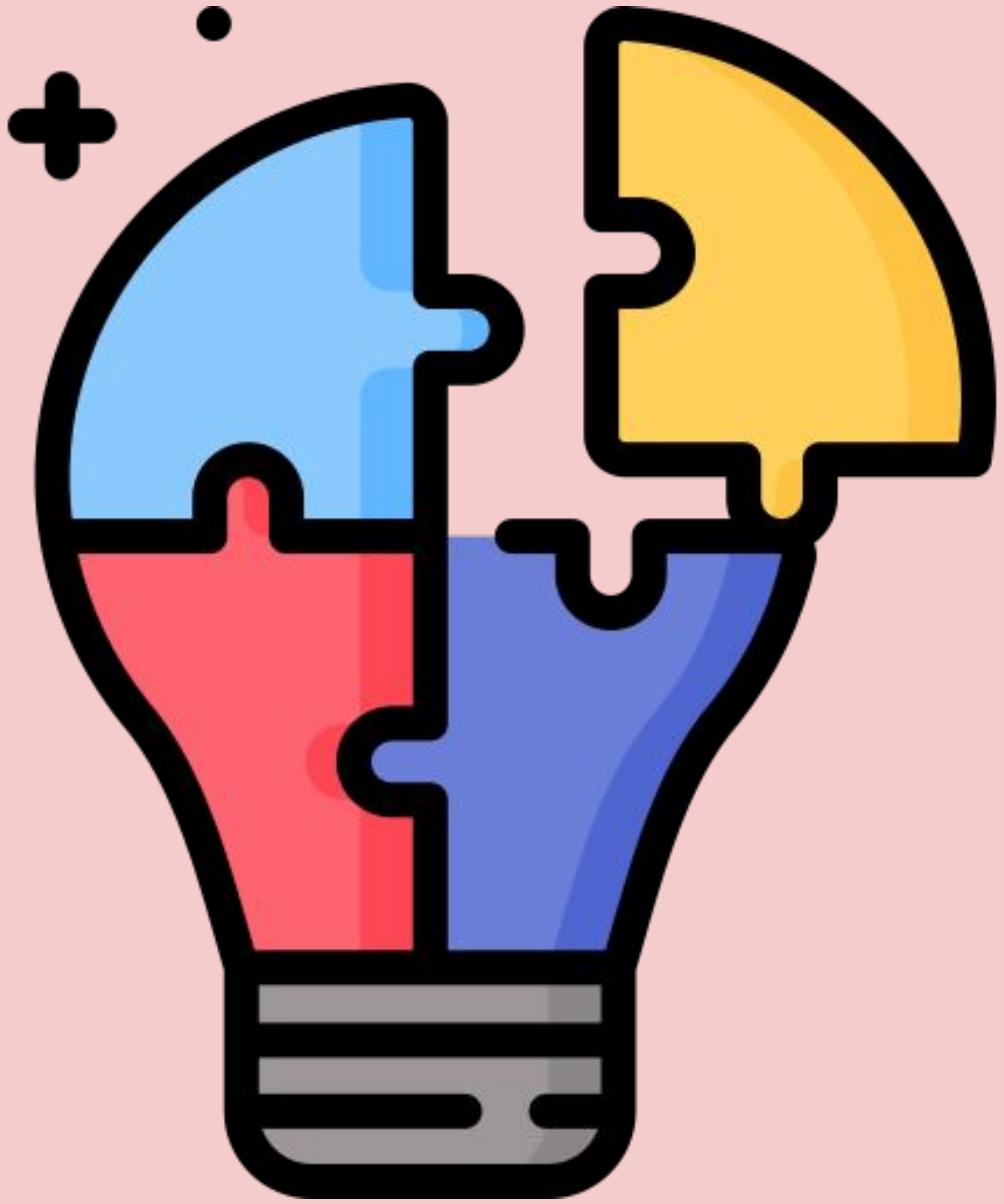
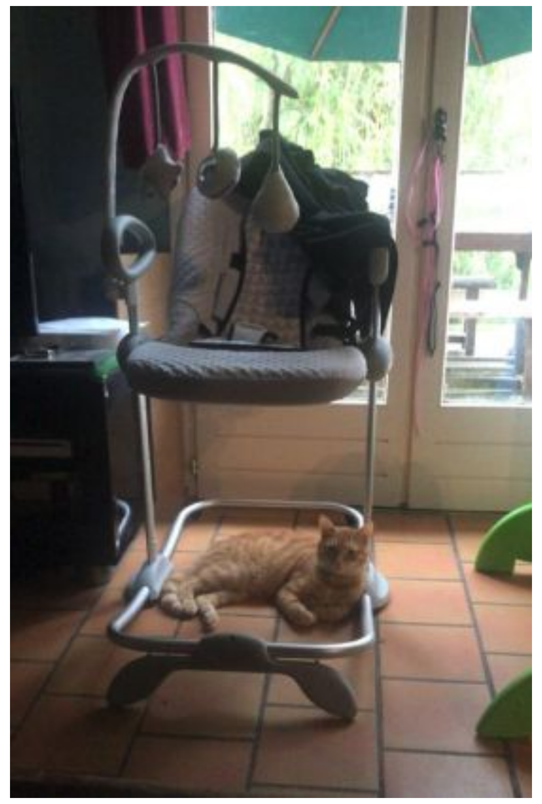


Image quality in marketplaces

poor lightning

clearly visible



distracting background

sharp image

poor framing

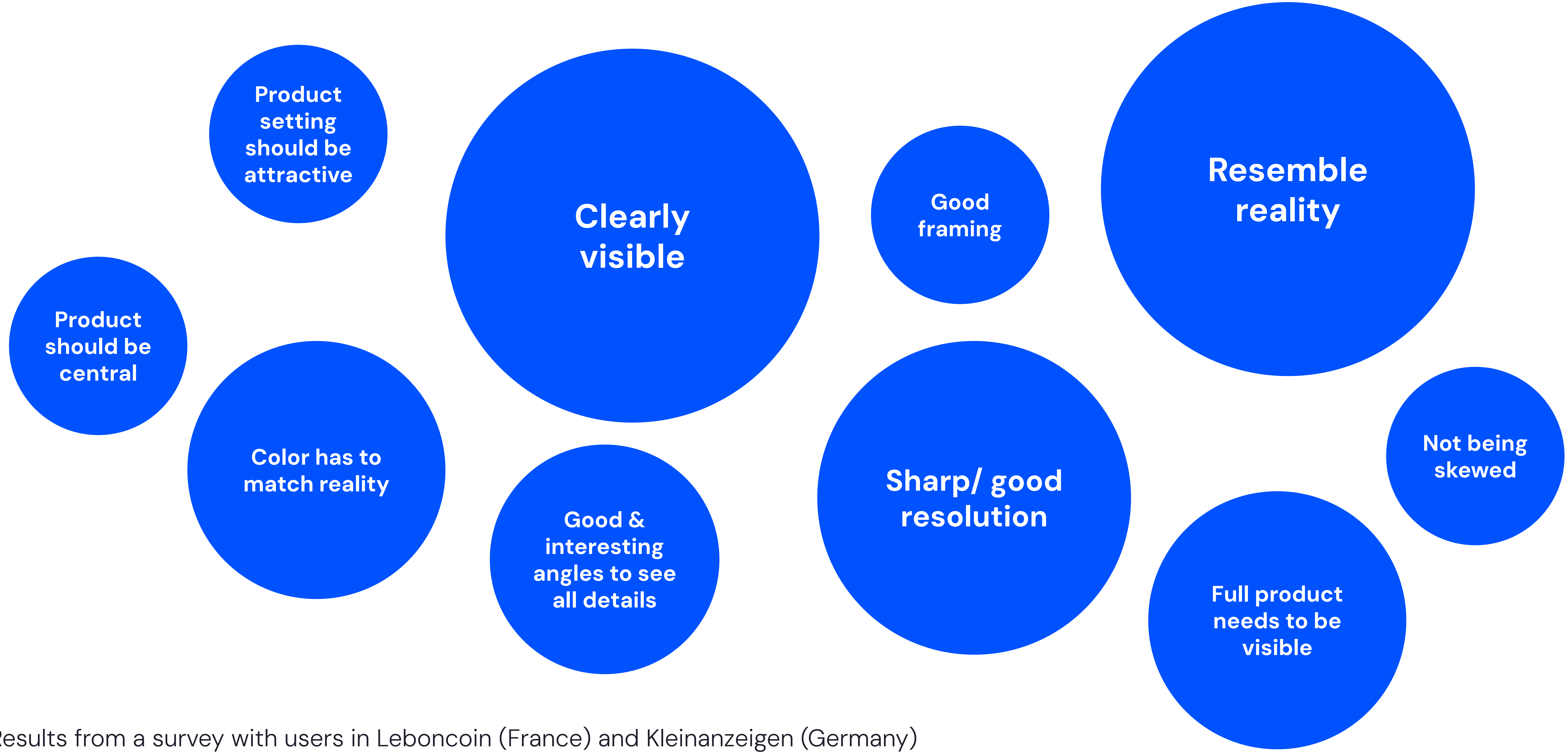
not fully visible

good lighting

good resolution

bad angle

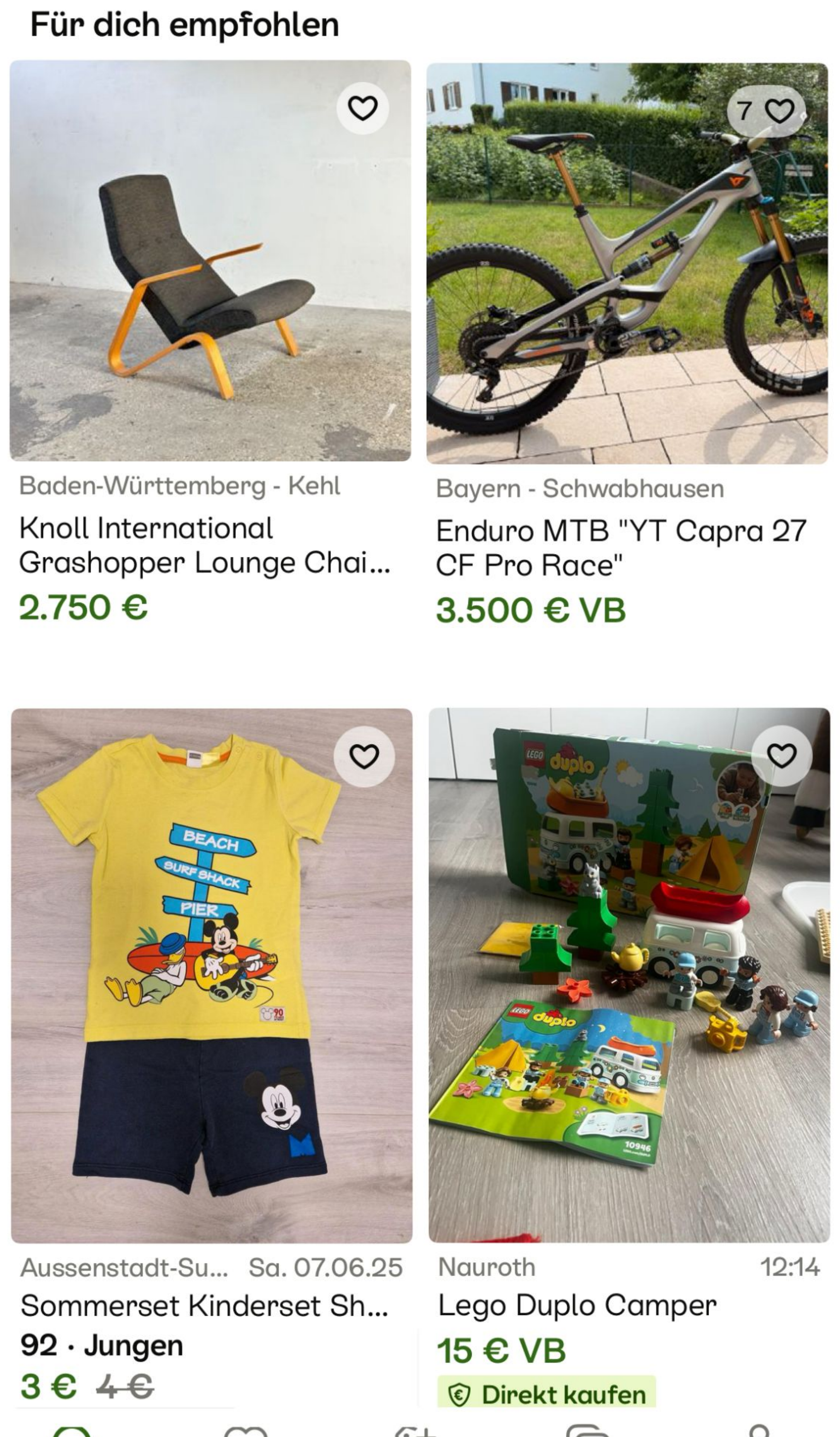
What does good quality image mean for marketplace users*



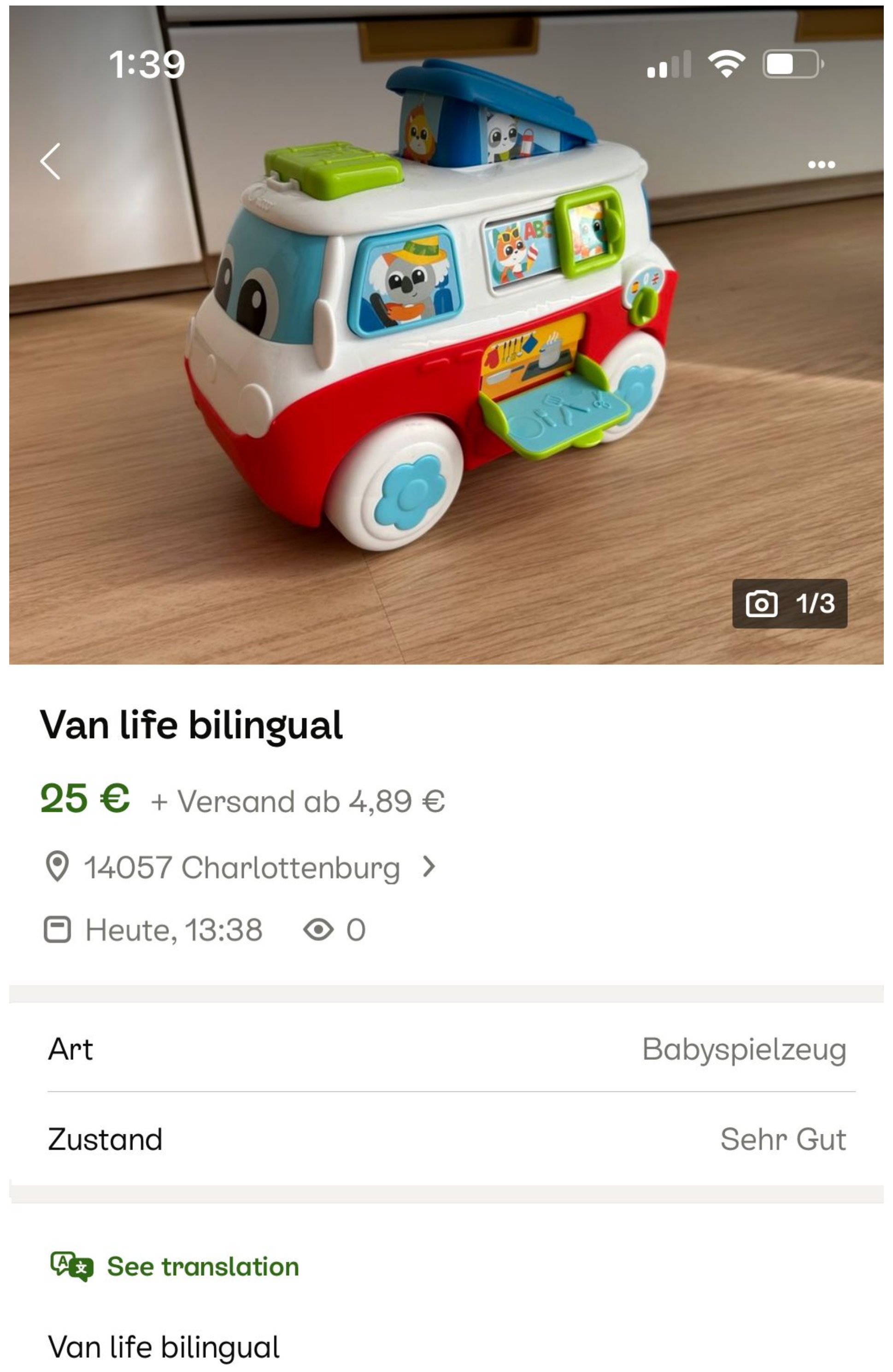
* Results from a survey with users in Leboncoin (France) and Kleinanzeigen (Germany)

Image quality is key for user experience

Main element in search results, personalised feeds and item detail page and key **engagement driver.**



Personalised feed



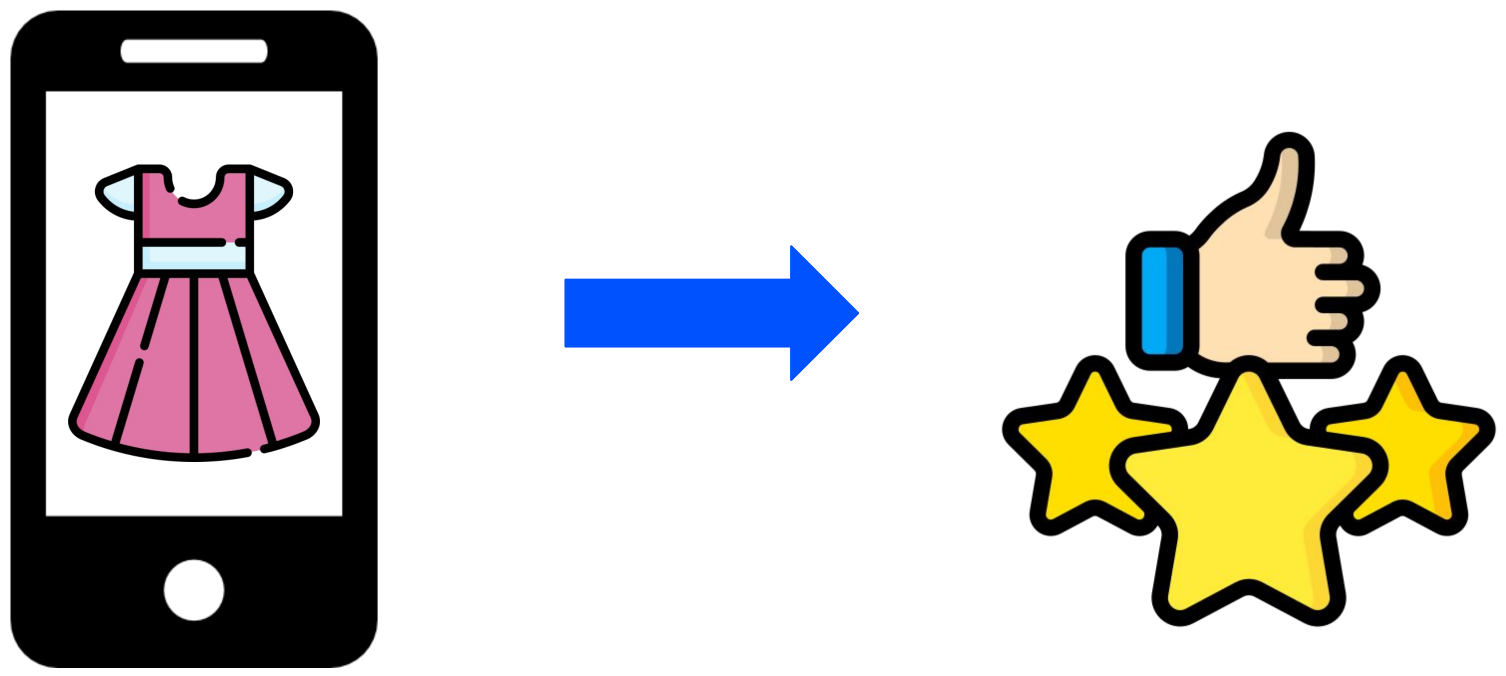
View item Page

Visual quality is a key factor influencing **trust in the marketplace***, especially among **Gen Z** users, who claim low overall image quality would be a reason to abandon a marketplace.

** also backed by* ***Ma et al. Understanding Image Quality and Trust in Peer-to-Peer Marketplaces. 2018***

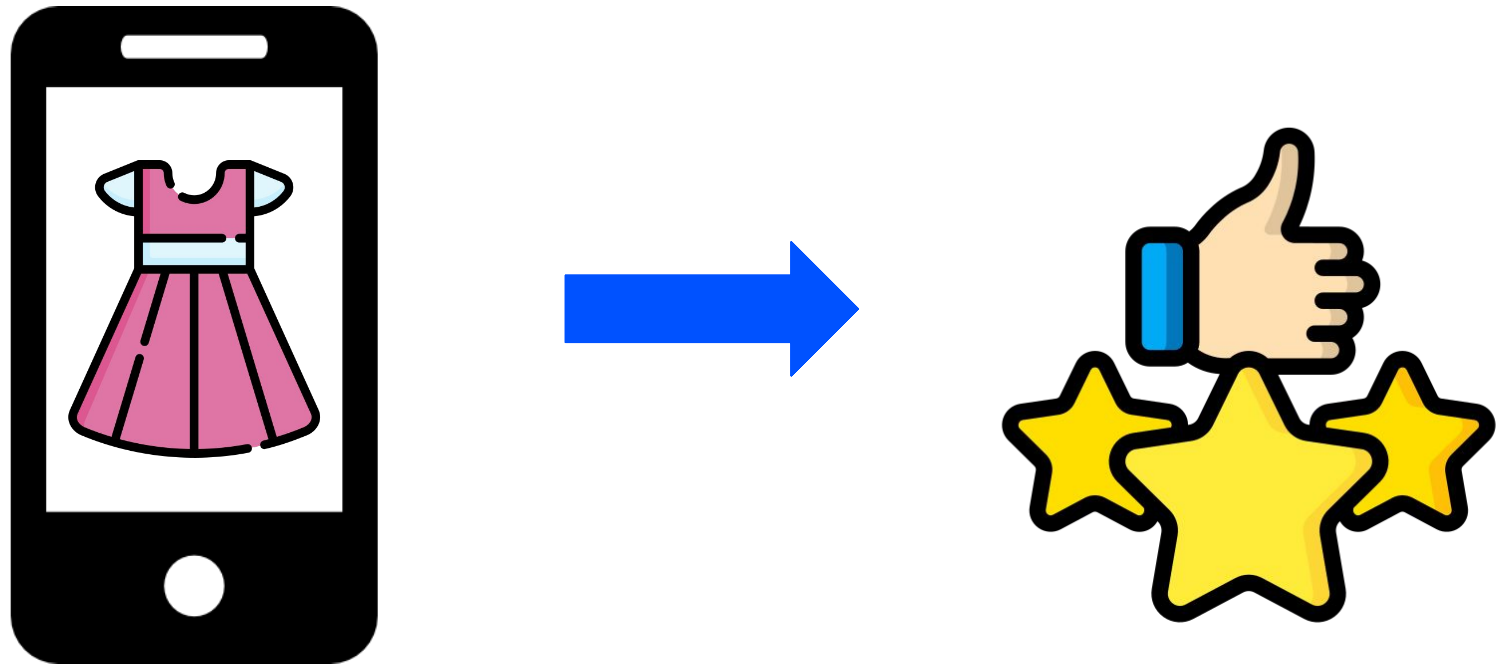
Our goal

Extract image quality scores that can be used to **understand image quality** of ads in our marketplace and to **enhance the user experience**.



Our goal

Extract image quality scores that can be used to **understand image quality** of ads in our marketplace and to **enhance the user experience**.



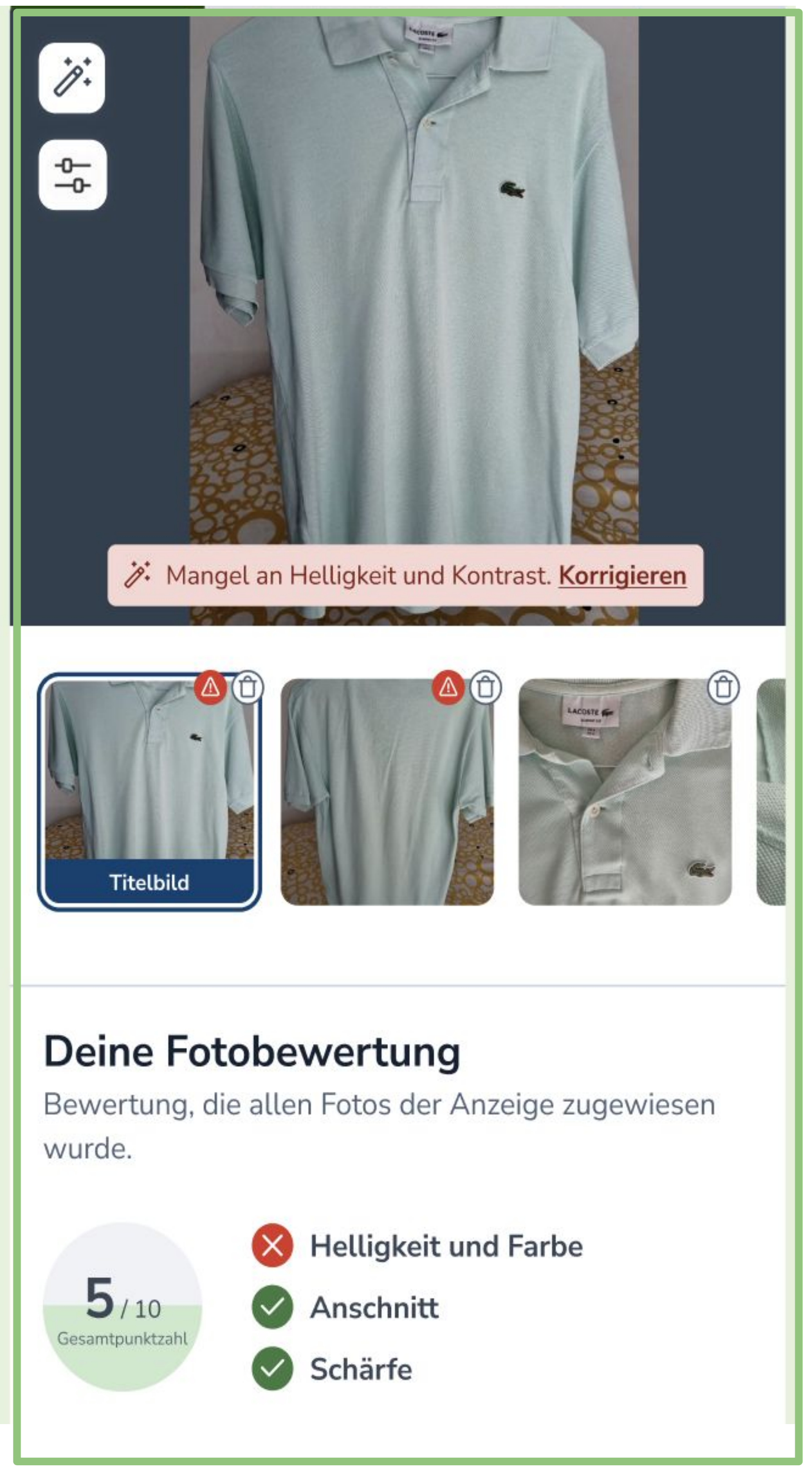
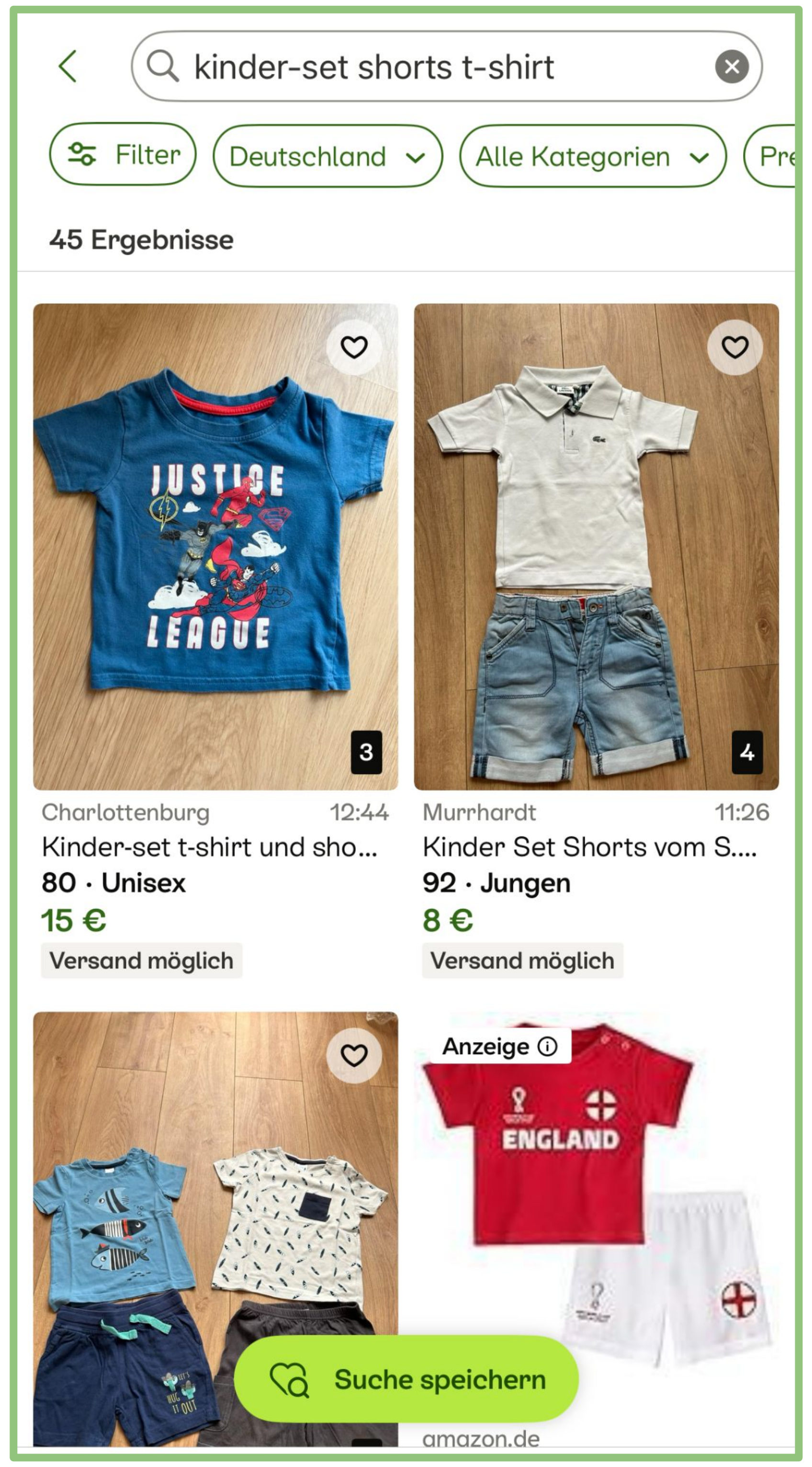


photo tips



ranking features

hero image suggestion

Approach

An LLM-based approach

1 – 5 scores

Questions:

- How good is the alignment of MLLMs scores with humans (our users)?
- How do different models compare and what are the trade-offs in **quality** vs model **size** vs **cost**?

LLMs evaluated: size and costs

	Lab	Parameters (B)	Cents per million input	Cents per million output
GPT-4o mini	OpenAI	8	15	60
GPT-4o	OpenAI	200	250	1000
Claude Haiku 3	Anthropic	20	25	125
Claude Sonnet 3.5	Anthropic	70	300	1500
Nova Lite	Amazon	20	6	24
Nova Pro	Amazon	90	80	320
Qwen2.5-VL-7B	Alibaba	7	-	-
Qwen2.5-VL-72B	Alibaba	72	-	-

https://docs.aws.amazon.com/bedrock/latest/userguide/custom-models.html

https://lifearchitect.ai/models-table/

Multimodal zero-shot prompts evaluated

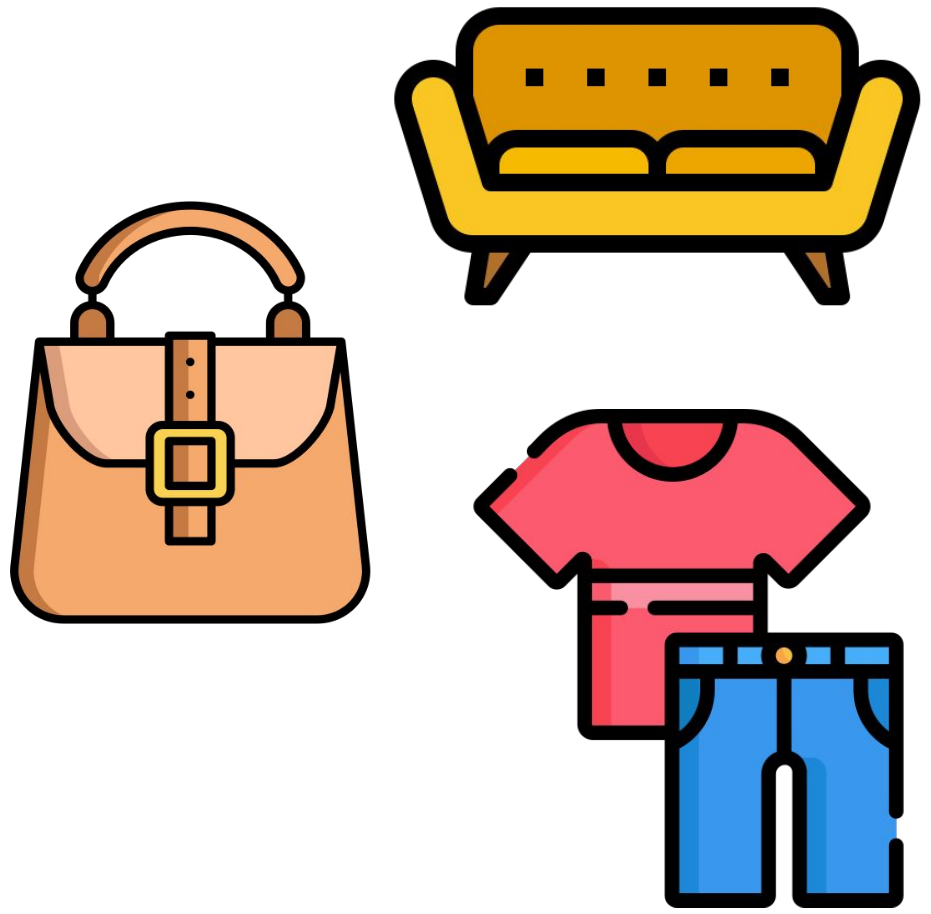
<table>
<tr><td colspan="2">Context: You are an expert at recognising good images for selling items in second-hand marketplaces.
Prefix: Given this image of a {item_type}, provide an overall image quality score for how good is the image if we want to sell it on a second-hand marketplace.</td></tr>
<tr><td>Generic Prompt:
The score should be a number on a scale of 1 to 5 (1 and 5 included). Do not expect more angles or close up shots as we can only use one image.</td><td>Guided Prompt:
If the image has none or minor aspects to improve, please return an overall_score of 5. Follow this guide when assigning scores:

Score 1: The image is horrible, with many aspects to improve.
Score 2: The image is bad, with several aspects to improve.
Score 3: The image is not great.
Score 4: The image is quite good, with just 1 or 2 things to improve.
Score 5: The image is fantastic, with none or minor things to improve.</td></tr>
<tr><td colspan="2">Suffix: Return the score and a justification for the score in JSON format with only the following keys: score, justification.</td></tr>
</table>

Human evaluation through online user survey

Used **Leboncoin's internal panel** (a pool of >= 8K users for internal user studies).

Users provided scores on a dataset of **600 images** of diverse quality.



3 categories: **bags, sofas** and **clothes** (balanced: 200 each).

Online survey sent via SurveyMonkey

Human evaluation through online user survey

- **929** users responded to the survey (~14% response rate)
- Each user rated **6 images** (1-5 score)
- **Open question:** *why did you give this score?*

Voici une photo de l'article d'occasion que vous avez recherché. Sur une échelle de 1 à 5, comment évalueriez-vous la qualité de cette photo ?

Sac

1: *Très mauvaise qualité* **2:** *Mauvaise qualité* **3:** *Qualité moyenne* **4:** *Bonne qualité* **5:** *Excellente qualité*

Préc. Suiv.

Comparing Human vs LLM scores

- Number of responses per ad image: 9.5 ± 3.8 average
  - Human score → majority voting

- Human-LLM alignment metrics:

inter-rater agreement

Percent Agreement

Weighted Kappa

inter-rater reliability

Pearson correlation

LLM-Human alignment results

Distribution of scores for generic vs guided prompts

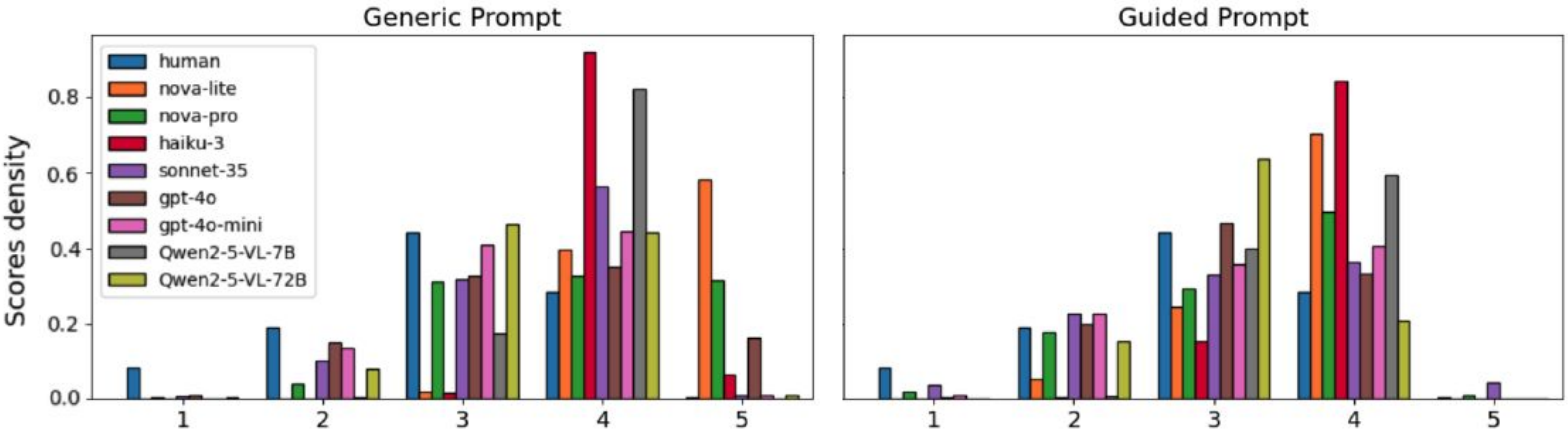


Figure 2: Score densities of different models using the generic prompt (left) and the guided prompt (right)

Impact of the zero-shot prompts

		Generic			Guided		
	Parameters (Billion)	**Pct Agreement**	**Weighted Kappa**	**Pearson Corr.**	**Pct Agreement**	**Weighted Kappa**	**Pearson Corr.**
GPT-4o mini	8	**0.494***	0.51	0.582	**0.529***	0.577	0.606
GPT-4o	200	0.355	**0.518***	**0.614**	0.469	**0.615**	**0.632**
Qwen2.5-VL-7B	7	0.322	0.236	0.566	0.406	0.36	0.584
Qwen2.5-VL-72B	72	0.436	0.419	0.504	0.483	0.428	0.467
Claude Haiku 3	20	0.242	0.074	0.299	0.33	0.104	0.269
Claude Sonnet 3.5	70	**0.442**	**0.497**	**0.618***	**0.513**	**0.621***	**0.647***
Nova Lite	20	0.043	0.135	0.519	0.386	0.341	0.542
Nova Pro	90	0.232	0.351	0.554	0.466	0.524	0.571

Impact of fine-tuning

- How much can we improve results by fine-tuning a MLLM?
  - 60%/40% train/test split - 480 / 240 images

		Pre-trained models (guided prompt)			Fine-tuned models (guided prompt)		
	Parameters (B)	**Pct Agreement**	**Weighted Kappa**	**Pearson Corr.**	**Pct Agreement**	**Weighted Kappa**	**Pearson Corr.**
Claude Sonnet 3.5	70	0.546	0.643	0.67			
Nova Lite	20	0.384	0.358	0.543			
Nova Pro	90	0.507	0.546	0.606			

Impact of fine-tuning

- How much can we improve results by fine-tuning a MLLM?
  - 60%/40% train/test split - 480 / 240 images

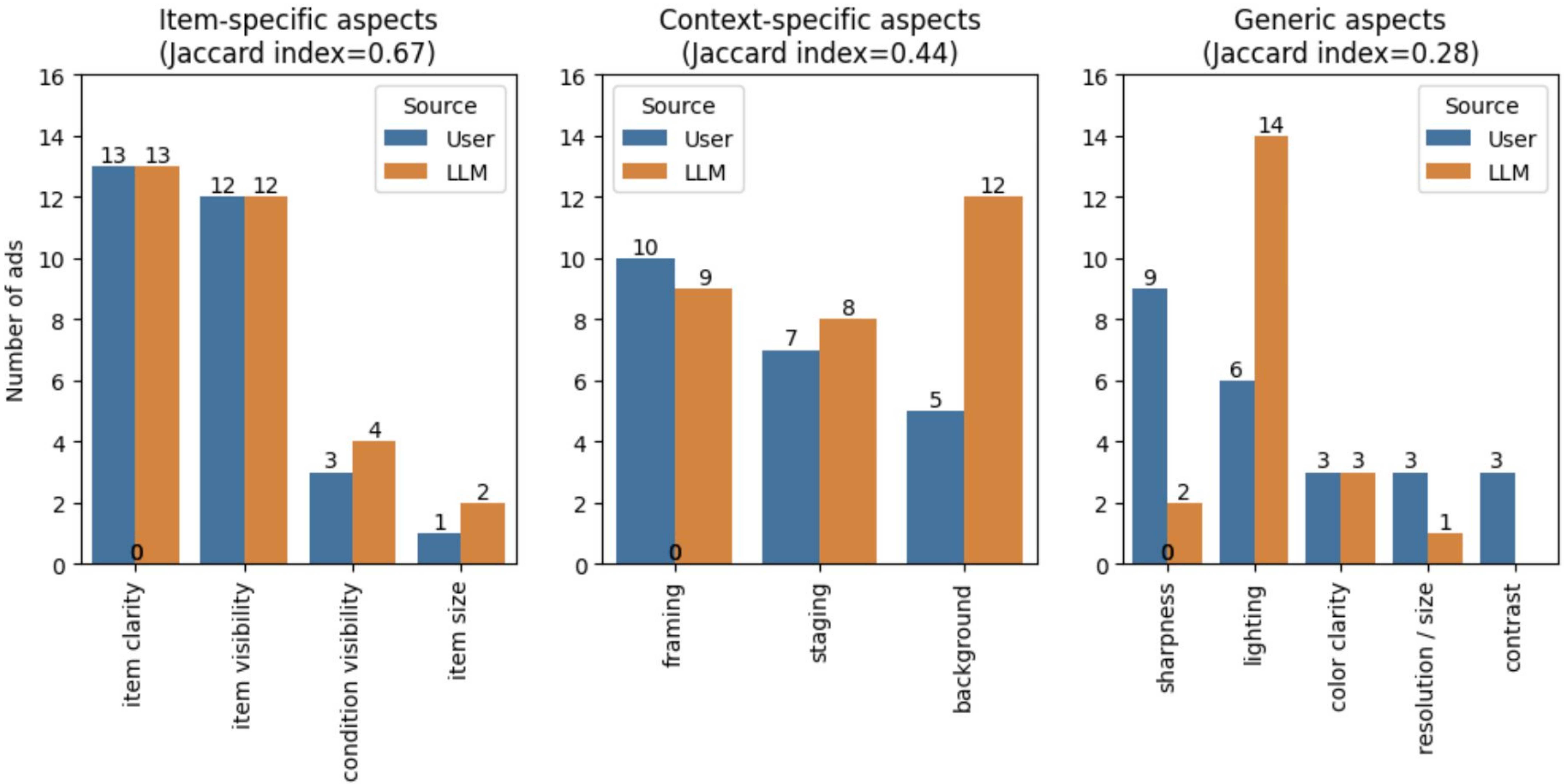
		Pre-trained models (guided prompt)			Fine-tuned models (guided prompt)		
	Parameters (B)	**Pct Agreement**	**Weighted Kappa**	**Pearson Corr.**	**Pct Agreement**	**Weighted Kappa**	**Pearson Corr.**
Claude Sonnet 3.5	70	0.546	0.643	0.67	-	-	-
Nova Lite	20	0.384	0.358	0.543	$\boxed{0.566 \text{ (47.40\%)}}$	0.601 (67.88%)	0.618 (13.81%)
Nova Pro	90	0.507	0.546	0.606	0.501 (-1.18%)	0.614 (12.45%)	0.638 (5.28%)

Significant gains when fine-tuning the small model (Nova Lite), beating inter-human Percent Agreement of **54.26%**

* This supports findings by ***Bucher & Martini.*** ***Fine-Tuned'Small'LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. 2024***.

Adevinta

Quality aspects mentioned in justifications



Adevinta

Conclusions

★ We have found high alignment between the **scores** provided by users and some MLLMs.

★ Fine-tuning a small model (Nova Lite; 20B params) shows significant gains compared to pre-trained models or fine-tuned larger model (Nova Pro; 70B params)
  - Key for cost reduction + sustainability

★ This work opens many possibilities for improving image quality in marketplaces:
  - Assyst sellers in taking better pictures or selecting the best hero photo.
  - Give more visibility to ads with better images (search, feeds …)

Thank you!

Gracias

Merci

Obrigado

Dank u wel

Danke

Grazie

Adevinta