The Impact of Web Search Result Quality on Decision-Making

September 12, 2024

Jan Heinrich
Merker

Lena
Merker

Alexander
Bondarenko

Friedrich-Schiller-Universität Jena
Martin-Luther-Universität Halle-Wittenberg

https://webis.de

CLEF 2024
GRENOBLE

15th International Conference of the CLEF Association (CLEF 2024)

The Impact of Web Search Result Quality on Decision-Making

Comparative questions and decision-making

Sources: `https://smartpastamaker.com/pizza-vs-pasta-which-is-healthier/`, `https://petsoid.com/cats-vs-dogs/`

What is healthier, pizza or pasta?

The Impact of Web Search Result Quality on Decision-Making

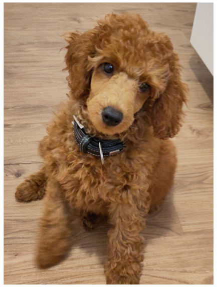
Comparative questions and decision-making

Sources: https://smartpastamaker.com/pizza-vs-pasta-which-is-healthier/, https://petsoid.com/cats-vs-dogs/

Should I adopt a dog or a cat?

The Impact of Web Search Result Quality on Decision-Making

Search engines and decision-making



→ How do you make a decision?

… unless it's an obvious choice 😉

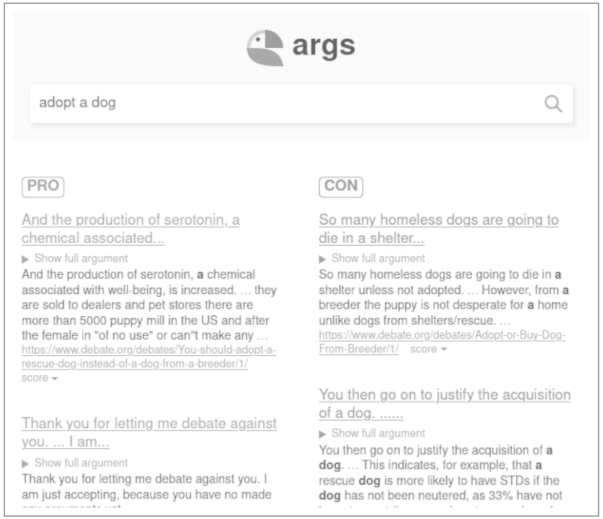
More reasons pro dog

Comparative argument retrieval

- ❑ Goal: Retrieve relevant, high-quality arguments
- ❑ Comparative questions used as search queries [Bondarenko et al., WSDM'20]
- ❑ Examples: args.me or ArgumenText [Wachsmuth et al., EMNLP'17; Stab et al., NAACL-HLT'18]

→ Yet, many use "normal" search engines, like Google

- – Known to be biased [Azzopardi, CHIIR'21]
- – Impact on decision-making unclear!



PRO

And the production of serotonin, a chemical associated...

▶ Show full argument

And the production of serotonin, **a** chemical associated with well-being, is increased. … they are sold to dealers and pet stores there are more than 5000 puppy mill in the US and after the female in "of no use" or can"t make any …

https://www.debate.org/debates/You-should-adopt-a-rescue-dog-instead-of-a-dog-from-a-breeder/1/

score ▾

Thank you for letting me debate against you. ... I am...

▶ Show full argument

Thank you for letting me debate against you. I am just accepting, because you have no made

CON

So many homeless dogs are going to die in a shelter...

▶ Show full argument

So many homeless dogs are going to die in **a** shelter unless not adopted. … However, from **a** breeder the puppy is not desperate for **a** home unlike dogs from shelters/rescue. …

https://www.debate.org/debates/Adopt-or-Buy-Dog-From-Breeder/1/ score ▾

You then go on to justify the acquisition of a dog. ......

▶ Show full argument

You then go on to justify the acquisition of **a** **dog**. … This indicates, for example, that **a** rescue **dog** is more likely to have STDs if the **dog** has not been neutered, as 33% have not

The Impact of Web Search Result Quality on Decision-Making

Search engines and decision-making

→ How do you make a decision?

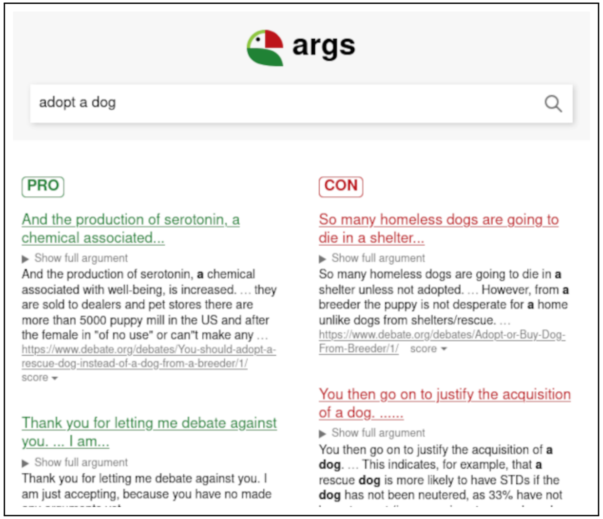
... unless it's an obvious choice ☺

More reasons pro dog

Comparative argument retrieval

- ❑ Goal: Retrieve relevant, high-quality arguments
- ❑ Comparative questions used as search queries [Bondarenko et al., WSDM'20]
- ❑ Examples: args.me or ArgumenText [Wachsmuth et al., EMNLP'17; Stab et al., NAACL-HLT'18]

→ Yet, many use "normal" search engines, like Google
  - – Known to be biased [Azzopardi, CHIIR'21]
  - – Impact on decision-making unclear!



And the production of serotonin, a chemical associated...

▶ Show full argument

And the production of serotonin, **a** chemical associated with well-being, is increased. … they are sold to dealers and pet stores there are more than 5000 puppy mill in the US and after the female in "of no use" or can"t make any …
https://www.debate.org/debates/You-should-adopt-a-rescue-dog-instead-of-a-dog-from-a-breeder/1/
score ▾

Thank you for letting me debate against you. ... I am...

▶ Show full argument

Thank you for letting me debate against you. I am just accepting, because you have no made

So many homeless dogs are going to die in a shelter...

▶ Show full argument

So many homeless dogs are going to die in **a** shelter unless not adopted. … However, from **a** breeder the puppy is not desperate for **a** home unlike dogs from shelters/rescue. …
https://www.debate.org/debates/Adopt-or-Buy-Dog-From-Breeder/1/ score ▾

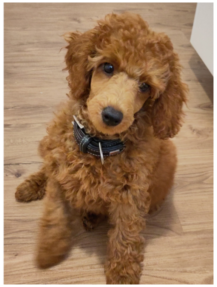
You then go on to justify the acquisition of a dog. ......

▶ Show full argument

You then go on to justify the acquisition of **a** **dog**. … This indicates, for example, that **a** rescue **dog** is more likely to have STDs if the **dog** has not been neutered, as 33% have not

The Impact of Web Search Result Quality on Decision-Making

Search engines and decision-making



→ How do you make a decision?

... unless it's an obvious choice ☺

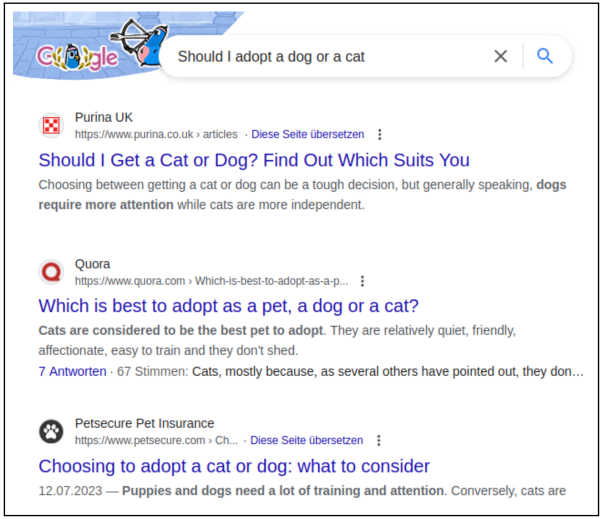
More reasons pro dog

Comparative argument retrieval

- ❑ Goal: Retrieve relevant, high-quality arguments
- ❑ Comparative questions used as search queries [Bondarenko et al., WSDM'20]
- ❑ Examples: args.me or ArgumenText [Wachsmuth et al., EMNLP'17; Stab et al., NAACL-HLT'18]

→ Yet, many use "normal" search engines, like Google

- – Known to be biased [Azzopardi, CHIIR'21]
- – Impact on decision-making unclear!



Should I adopt a dog or a cat

Purina UK
https://www.purina.co.uk › articles · Diese Seite übersetzen

Should I Get a Cat or Dog? Find Out Which Suits You

Choosing between getting a cat or dog can be a tough decision, but generally speaking, **dogs require more attention** while cats are more independent.

Quora
https://www.quora.com › Which-is-best-to-adopt-as-a-p...

Which is best to adopt as a pet, a dog or a cat?

Cats are considered to be the best pet to adopt. They are relatively quiet, friendly, affectionate, easy to train and they don't shed.

7 Antworten · 67 Stimmen: Cats, mostly because, as several others have pointed out, they don...

Petsecure Pet Insurance
https://www.petsecure.com › Ch... · Diese Seite übersetzen

Choosing to adopt a cat or dog: what to consider

12.07.2023 — **Puppies and dogs need a lot of training and attention**. Conversely, cats are

The Impact of Web Search Result Quality on Decision-Making

Hypotheses

1. Subjective comparisons lead to less confident decisions than factual.
   Intuition: Factual comparative topics often “better” answered by search engines than subjective comparative topics [Bondarenko et al., WSDM’20]

2. Low-quality results lead to less confident decisions than high-quality results.
   Intuition: Desire to make best decision based on available information [Peterson, ’17]

3. The higher a document’s quality, the more likely it influences the decision.
   Intuition: Same as for 2

4. More confident users are less influenced by low-quality documents.
   Intuition: Confident users rely more on own knowledge than on ad hoc information [Peterson, ’17]

5. Documents that take a stance have a higher impact on the decision.
   Intuition: Relevant documents often expected to take a stance [Bondarenko et al., WSDM’22]

The Impact of Web Search Result Quality on Decision-Making

Hypotheses (and spoilers)

1. *~~Subjective comparisons lead to less confident decisions than factual.~~*

   Intuition: Factual comparative topics often “better” answered by search engines than subjective comparative topics [Bondarenko et al., WSDM’20]

2. *~~Low-quality results lead to less confident decisions than high-quality results.~~*

   Intuition: Desire to make best decision based on available information [Peterson, ’17]

3. The higher a document’s quality, the more likely it influences the decision.

   Intuition: Same as for 2

4. More confident users are less influenced by low-quality documents.

   Intuition: Confident users rely more on own knowledge than on ad hoc information [Peterson, ’17]

5. Documents that take a stance have a higher impact on the decision.

   Intuition: Relevant documents often expected to take a stance [Bondarenko et al., WSDM’22]

The Impact of Web Search Result Quality on Decision-Making

Methodology

- ❑ Develop document quality criteria for comparative topics
- ❑ Assess quality, relevance, and stance of top-4 Google results for 30 topics
- ❑ Conduct user study on the decision-making
  - – Decision and confidence before/after seeing results
  - – Influence of retrieved documents

Data

Touché shared tasks

- ❑ Comparative argument retrieval task in 2020–2022
  100 topics comparing two or more options (e.g., dog vs. cat)
- → 30 topics used for quality assessment
  (comparing 2 options, easy to understand)

Source: https://touche.webis.de/

Google search engine

- ❑ Most popular search engine in Europe
- → Top-4 results used (after excl. ads / media-only results)

Source: https://unsplash.com/

The Impact of Web Search Result Quality on Decision-Making

Methodology

❑ Develop document quality criteria for comparative topics

❑ Assess quality, relevance, and stance of top-4 Google results for 30 topics

❑ Conduct user study on the decision-making

- Decision and confidence before/after seeing results
- Influence of retrieved documents

Data

Touché shared tasks

❑ Comparative argument retrieval task in 2020–2022
100 topics comparing two or more options (e.g., dog vs. cat)

→ 30 topics used for quality assessment
(comparing 2 options, easy to understand)

Source: https://touche.webis.de/

Google search engine

❑ Most popular search engine in Europe

→ Top-4 results used (after excl. ads / media-only results)

Source: https://unsplash.com/

The Impact of Web Search Result Quality on Decision-Making

Quality criteria

Quality [Lewandowski et al., '08]

Content	Usability	Credibility	Up-to-dateness
Completeness, scope, language	Media types, structure	Source, author, truthfulness, verifiability	Date, updates

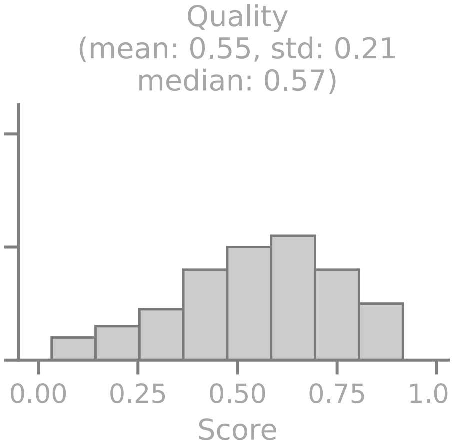
Other criteria

Relevance	Stance
Topical relevance	Referral, emphasis, direction, magnitude

- ❑ Based on prior quality assessment frameworks: WebQual, 2QCV3Q, AIMQ, Touché
- ❑ Relevance and quality also assessed for comparison purposes

Quality assessment

- ❑ 120 documents assessed (Google's top-4 of 30 topics)
- ❑ 10 volunteer assessors (media / computer science stud.)
- ❑ Agreement measured based on randomly selected topic (Fleiss' κ; 6 aspects with insufficient agreement excluded)
- ❑ Calculate aggregated quality score per document and topic



The Impact of Web Search Result Quality on Decision-Making

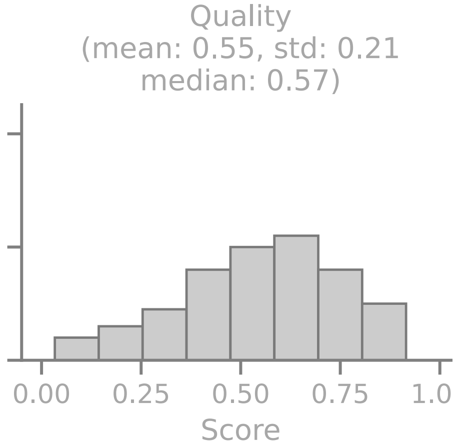
Quality criteria

Quality [Lewandowski et al., '08]				*Other criteria*	
Content	Usability	Credibility	Up-to-dateness	Relevance	Stance
Completeness, scope, language	Media types, structure	Source, author, truthfulness, verifiability	Date, updates	Topical relevance	Referral, emphasis, direction, magnitude

- ❑ Based on prior quality assessment frameworks: WebQual, 2QCV3Q, AIMQ, Touché
- ❑ Relevance and quality also assessed for comparison purposes

Quality assessment

- ❑ 120 documents assessed (Google's top-4 of 30 topics)
- ❑ 10 volunteer assessors (media / computer science stud.)
- ❑ Agreement measured based on randomly selected topic (Fleiss' κ; 6 aspects with insufficient agreement excluded)
- ❑ Calculate aggregated quality score per document and topic



The Impact of Web Search Result Quality on Decision-Making

Quality criteria

Quality [Lewandowski et al., ’08]

Content
Completeness, scope, language

Usability
Media types, structure

Credibility
Source, author, truthfulness, verifiability

Up-to-dateness
Date, updates

Other criteria

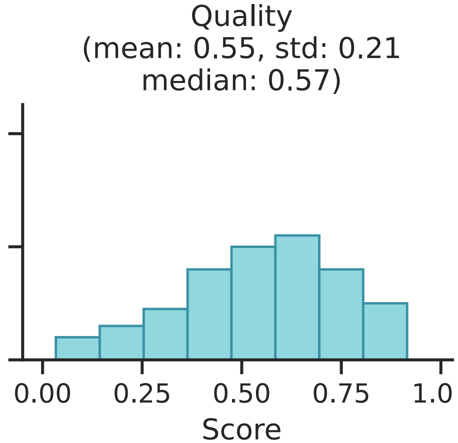
Relevance
Topical relevance

Stance
Referral, emphasis, direction, magnitude

- ❑ Based on prior quality assessment frameworks: WebQual, 2QCV3Q, AIMQ, Touché
- ❑ Relevance and quality also assessed for comparison purposes

Quality assessment

- ❑ 120 documents assessed (Google’s top-4 of 30 topics)
- ❑ 10 volunteer assessors (media / computer science stud.)
- ❑ Agreement measured based on randomly selected topic (Fleiss’ κ; 6 aspects with insufficient agreement excluded)
- ❑ Calculate aggregated quality score per document and topic



The Impact of Web Search Result Quality on Decision-Making

Quality criteria

Quality [Lewandowski et al., '08]

Content
Completeness, scope, language

Usability
Media types, structure

Credibility
Source, author, truthfulness, verifiability

Up-to-dateness
Date, updates

Other criteria

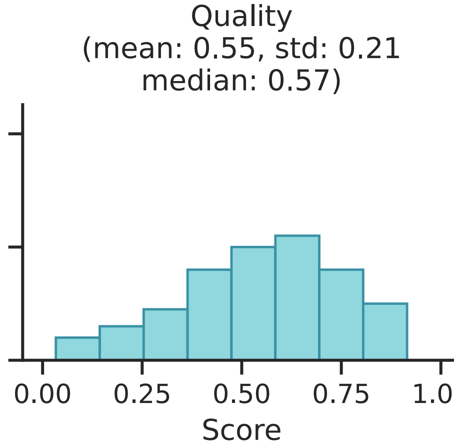
Relevance
Topical relevance

Stance
Referral, emphasis, direction / magnitude

- ❑ Based on prior quality assessment frameworks: WebQual, 2QCV3Q, AIMQ, Touché
- ❑ Relevance and quality also assessed for comparison purposes

Quality assessment

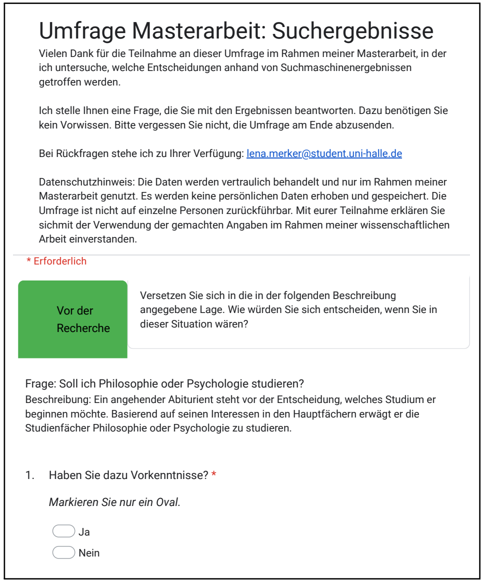
- ❑ 120 documents assessed (Google's top-4 of 30 topics)
- ❑ 10 volunteer assessors (media / computer science stud.)
- ❑ Agreement measured based on randomly selected topic (Fleiss' κ; 6 aspects with insufficient agreement excluded)
- ❑ Calculate aggregated quality score per document and topic



The Impact of Web Search Result Quality on Decision-Making

User study

- Select 8 topics and screenshot top-4 results
  - Exclude topics with missing quality judgments
  - Cover wide range of topic-wise avg. quality
- Survey layout:
  - Introduction and topic description
  - Prior knowledge assessment
  - Decision/confidence before seeing results
  - Screenshots of documents
  - Decision/confidence after seeing results
  - Self-assessment of decision-making process (6 statements)
- 442 volunteer participants (German univ. students)
- 554 study responses (1–8 topics per participant)



Umfrage Masterarbeit: Suchergebnisse

Vielen Dank für die Teilnahme an dieser Umfrage im Rahmen meiner Masterarbeit, in der ich untersuche, welche Entscheidungen anhand von Suchmaschinenergebnissen getroffen werden.

Ich stelle Ihnen eine Frage, die Sie mit den Ergebnissen beantworten. Dazu benötigen Sie kein Vorwissen. Bitte vergessen Sie nicht, die Umfrage am Ende abzusenden.

Bei Rückfragen stehe ich zu Ihrer Verfügung: lena.merker@student.uni-halle.de

Datenschutzhinweis: Die Daten werden vertraulich behandelt und nur im Rahmen meiner Masterarbeit genutzt. Es werden keine persönlichen Daten erhoben und gespeichert. Die Umfrage ist nicht auf einzelne Personen zurückführbar. Mit eurer Teilnahme erklären Sie sichmit der Verwendung der gemachten Angaben im Rahmen meiner wissenschaftlichen Arbeit einverstanden.

* Erforderlich

Vor der Recherche

Versetzen Sie sich in die in der folgenden Beschreibung angegebene Lage. Wie würden Sie sich entscheiden, wenn Sie in dieser Situation wären?

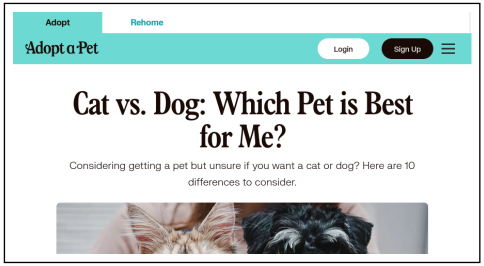
Frage: Soll ich Philosophie oder Psychologie studieren?
Beschreibung: Ein angehender Abiturient steht vor der Entscheidung, welches Studium er beginnen möchte. Basierend auf seinen Interessen in den Hauptfächern erwägt er die Studienfächer Philosophie oder Psychologie zu studieren.

1. Haben Sie dazu Vorkenntnisse? *

Markieren Sie nur ein Oval.

○ Ja
○ Nein

Survey view



Cat vs. Dog: Which Pet is Best for Me?

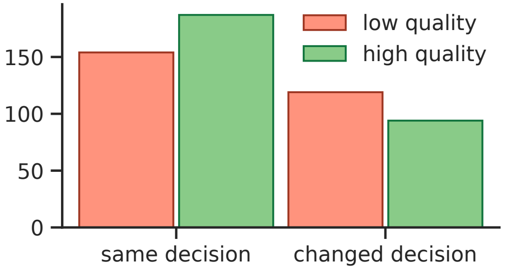
Considering getting a pet but unsure if you want a cat or dog? Here are 10 differences to consider.

Document screenshot

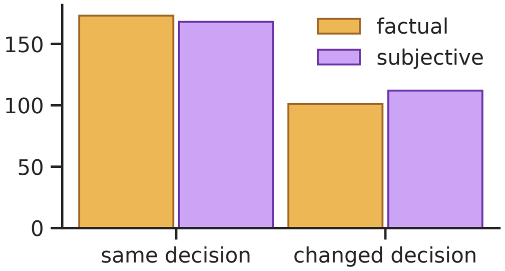
The Impact of Web Search Result Quality on Decision-Making

Results: Decision and confidence change

Decision change

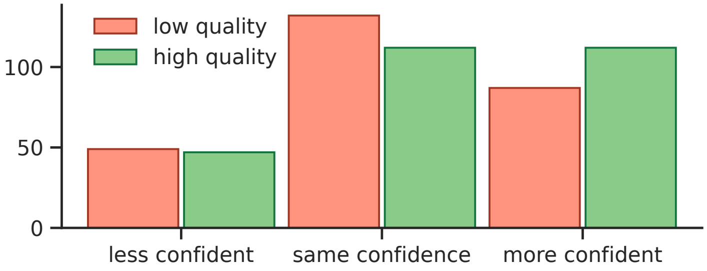


$\boxed{\boldsymbol{\chi^2(1) = 5.59, p = 0.018}}$

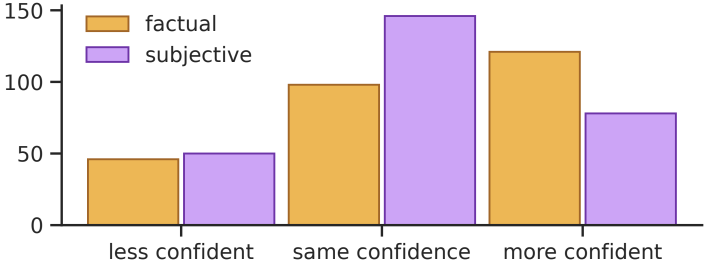


$\chi^2(1) = 0.45, p = 0.502$

Confidence change



$\chi^2(2) = 4.81, p = 0.090$



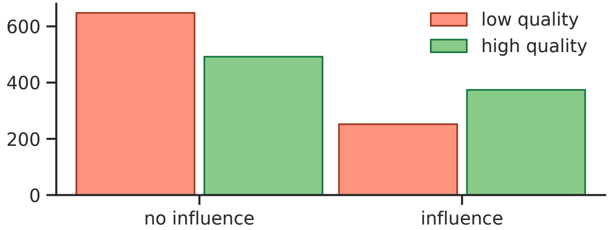
$\boxed{\boldsymbol{\chi^2(2) = 18.76, p < 0.001}}$

- ❑ Majority did not change decision, but 37% gained confidence
- ❑ Decision changed more often with overall low-quality results, but more confident with high-quality results
- ❑ Decision confidence significantly increased more for factual topics

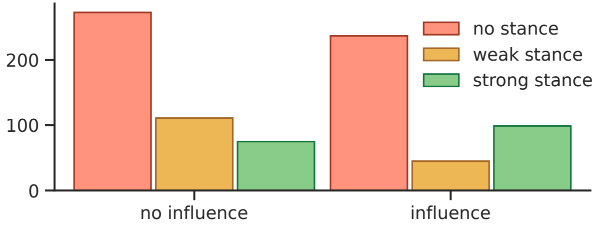
The Impact of Web Search Result Quality on Decision-Making

Results: Influence of documents

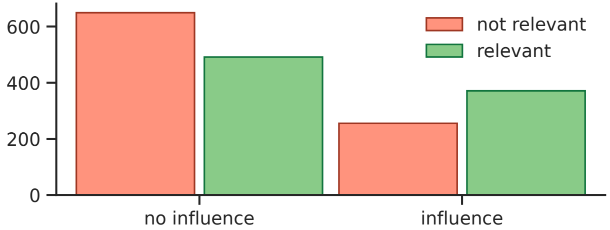
Influenced decision-making



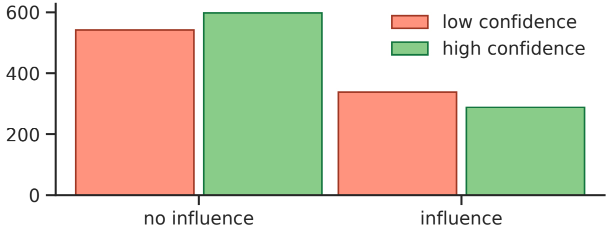
$\chi^2(4) = 44.49, p < 0.001$



$\chi^2(2) = 26.76, p < 0.001$



$\chi^2(1) = 41.77, p < 0.001$



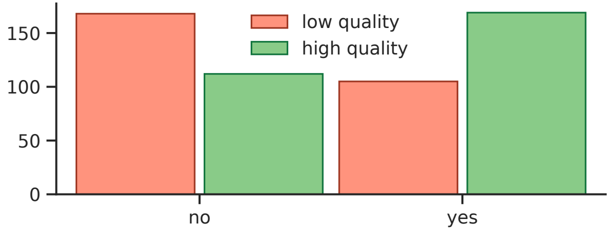
$\chi^2(1) = 4.51, p = 0.034$

- More likely to influence decision-making:
  - high-quality
  - relevance
  - strong stance
- Confident users less likely to be influenced by low-quality documents

The Impact of Web Search Result Quality on Decision-Making

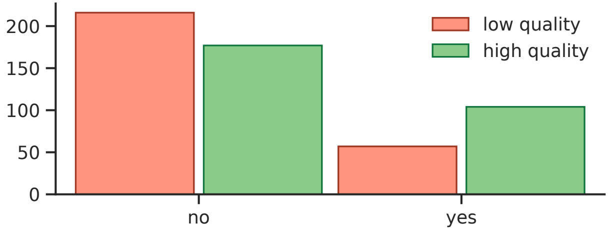
Results: Self-assessment of decision-making process

I learned something new



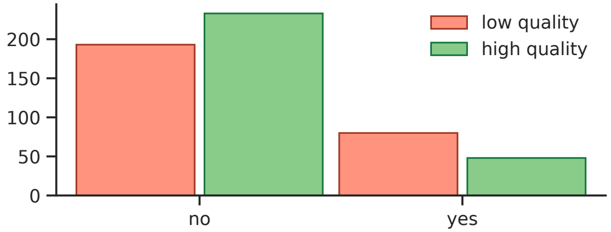
$$\chi^2(1) = 25.18, p < 0.001$$

I could make a better decision



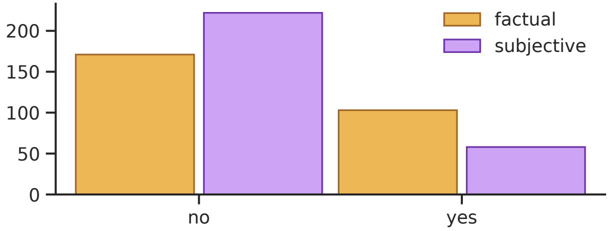
$$\chi^2(1) = 18.32, p < 0.001$$

I would continue searching



$$\chi^2(1) = 10.96, p < 0.001$$

I could make a better decision



$$\chi^2(1) = 16.71, p < 0.001$$

- Participants learned sth. new (50%), but would still often continue search (25%)
- High-quality results more helpful, less likely to continue searching
- Better decisions with factual topics and high-quality results

The Impact of Web Search Result Quality on Decision-Making

Results: Hypotheses

1. ~~*Subjective comparisons lead to less confident decisions than factual.*~~

   → Reject: Better decision for factual topics but not significantly more confident.

2. ~~*Low-quality results lead to less confident decisions than high-quality results.*~~

   → Reject: Slight increase in confidence with high-quality topics, better decision with high-quality topics; but overall not significant.

3. The higher a documents's quality, the more likely it influences the decision.

   → Accept: Low-quality documents influence decisions significantly less often than high-quality documents; position bias ruled out.

4. More confident users are less influenced by low-quality documents.

   → Accept: Confident users significantly less likely to be influenced by low-quality documents.

5. Documents that take a stance have a higher impact on the decision.

   → Accept: Docs. with strong stance influenced decision more often than with weak stance.

The Impact of Web Search Result Quality on Decision-Making

Summary

- First step for quality assessment of comparative queries
- Quality has significant impact on decision-making process
- Potential ranking factors: quality, stance (especially for subjective topics)
- Limitations: only German student participants, single search engine
- Future work: larger study (e.g., more participants / topics / search engines)

Code and data

- github.com/webis-de/CLEF-24
- doi.org/978-3-031-71736-9_5

The Impact of Web Search Result Quality on Decision-Making

Summary

- First step for quality assessment of comparative queries
- Quality has significant impact on decision-making process
- Potential ranking factors: quality, stance (especially for subjective topics)
- Limitations: only German student participants, single search engine
- Future work: larger study (e.g., more participants / topics / search engines)

Code and data

- github.com/webis-de/CLEF-24
- doi.org/978-3-031-71736-9_5

Thank you & merci!