

# A Comparative Evaluation of End-to-End and Pipeline Approaches for Summarisation

Fahime (Fafa) Same, Saad Mahamood, Srinivas Ramesh Kamath  
University of Cologne, Shopware, trivago  
f.same@uni-koeln.de, s.mahamood@shopware.com, srinik352@gmail.com



CODE & DATA

## 1. WHAT & WHY

We compare two architectures for generating very short book summaries (aka highlights) from 2018 Amazon reviews and the Amazon KG.

**End-to-End (E2E)**  
turns the raw input  
straight  
into highlights

**Pipeline (PIPE)**  
Explicit steps  
Analyse content → What  
to say  
→ how to say

### One side

E2E LLM solutions have known weaknesses: inaccuracy, incompleteness, and they are hard to audit. Pipelines add more control and observability.

**The gap:** Pipeline advantage has been shown mostly for rule-based and non-LLM neural models. Much less is known about modern LLMs.

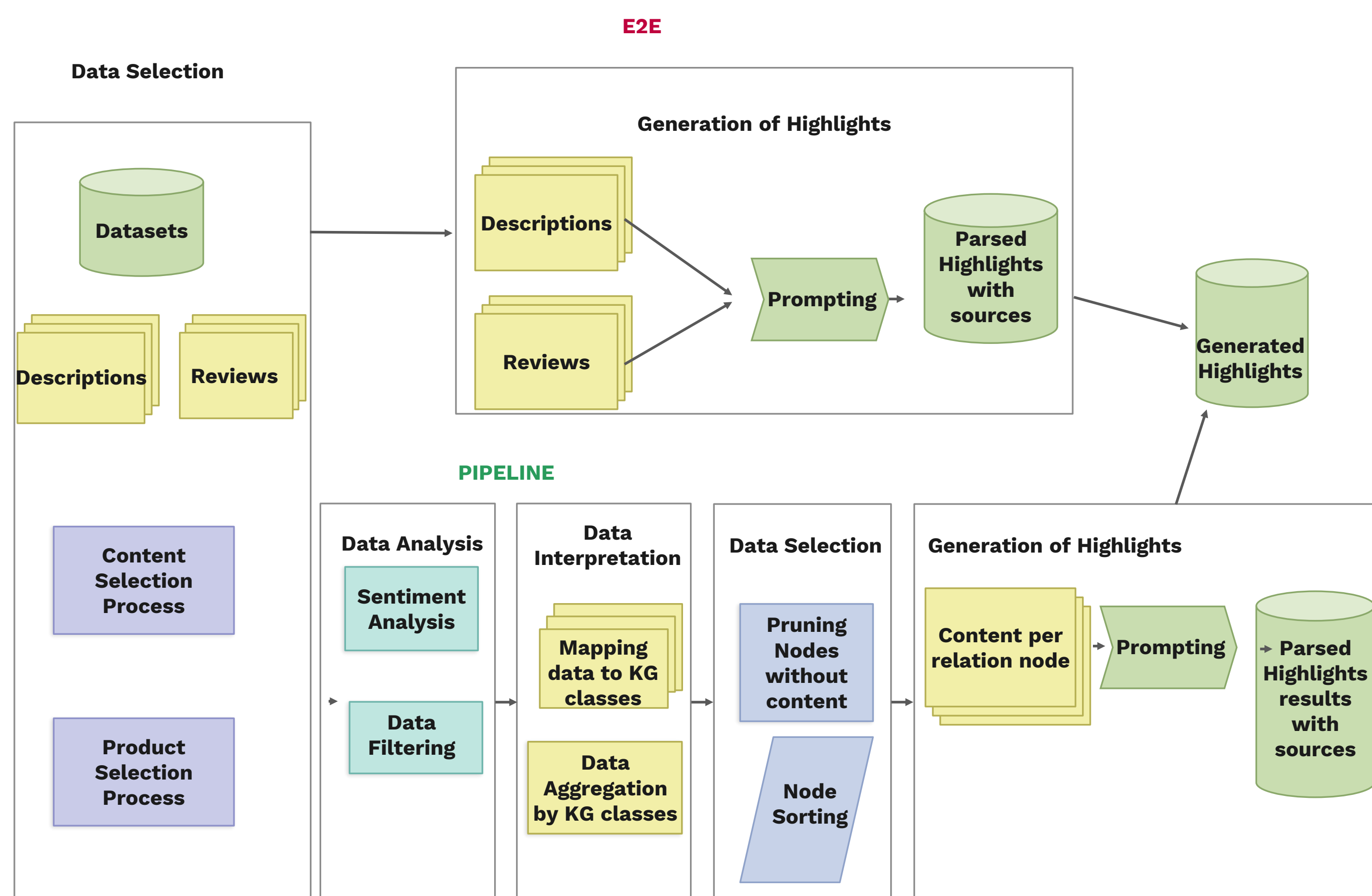
Can the pipeline's control  
reduce these  
weaknesses?

Is the extra work still  
justified?

### The other side

LLMs keep improving: more accurate, more capable, and better at instructions. Pipelines are step-wise and more complex, and researcher bias might be there.

## 2. ARCHITECTURE



## 3. COMPARISONS

### 1 Architecture

E2E vs. PIPE

### 2 Generation level

Broad (relation) vs. Detailed (tail)  
Subject vs. Novels set in Europe

### 3 Model family

GPT-OSS vs. Llama-3.1 vs. Qwen-3  
3 open-source families

### 4 Parameter size

Large vs. Small  
within each family

6 models x 2 architectures x 2 generation levels

## 4. RESEARCH QUESTIONS

**[ARCH]** Does architecture affect highlight quality & faithfulness?

Overall, **PIPE** over **E2E**

**[GEN-LEVEL]** Does this effect differ between relation- & tail-level generations?

Tail: larger **PIPE** advantage over **E2E**

**[SIZE]** Does model size affect quality and faithfulness within a family?

Within family, larger over smaller

**[ARCH × SIZE]** Does the architecture effect depend on model size?

Smaller models: larger **PIPE** advantage over **E2E**

**[FAMILY]** Does the architecture effect vary across families?

Across families, **PIPE** advantage varies.

## 5. PREFERENCE EVALUATION

COMPARISON

Architecture

BOOK

The Heart of the Matter

THEME

previousWork

HIGHLIGHT A

Like other Greene novels, Heart of the Matter follows settings in foreign times and places.

HIGHLIGHT B

Preceded by 'A Burnt Out Case,' sharing Greene's focus on moral ambiguity and colonial tensions.

For each of the 5 criteria below, the judge returns: CHOICE (A, B, tie, or neither), CONFIDENCE, and RATIONALE

Informativeness

How much useful, concrete, book-specific information the candidate conveys for the theme.

Saliency

How well the candidate surfaces a point that would stand out and help a reader decide.

Theme adherence

How well the candidate stays focused on the intended theme, not another aspect of the book.

Fluency and Style

How natural, polished, and readable the candidate is.

Coherence and Style

How logically organised and internally consistent the candidate is. Penalise contradictions, abrupt jumps, or confusing structure.

Overall Preference

If only one candidate could be shown for this theme, which one would you choose?

B

Dominant factors:

Rationale:

informativeness, saliency

B names a specific prior work and highlights salient themes. A is too vague.

## 6. FAITHFULNESS EVALUATION

LEVEL

tail

SIZE

small

SYS.

e2e

BOOK

Tatiana and Alexander

HIGHLIGHT

First installment of a critically acclaimed trilogy, setting the stage for the next two novels.

For each highlight, the judge returns: **FACTUALLY\_ACCURATE**, **DIVERGENCE\_TYPE**, **SEVERITY**, and **RATIONALE**

Divergence type

Exactly one label per highlight.

None

Every material claim is supported by the source.

Hallucination

Unsupported claim not established in the source.

Contradiction

Claim conflicts with the source.

Both

Unsupported and contradictory claims together.

Severity

Only when divergence is not none: integer from 1 to 7.

1

2

3

4

5

6

1 Minimal: small error, little effect on understanding

4 Moderate: clear error that changes an important detail

7 Extreme: fundamentally unfaithful to the source

Example Judgement

factually\_accurate: **false**  
divergence\_type: **both**  
severity: **6**

Rationale

The highlight incorrectly identifies this book as the 'First installment' of a trilogy; the source text repeatedly describes it as the second book and a sequel to 'The Bronze Horseman'. Consequently, the claim it 'sets the stage for the next two novels' is also a contradiction, as only one book follows. The term 'critically acclaimed' is a hallucination, as it is unsupported by the provided user reviews.

## 7. RESULTS

### PREFERENCE JUDGMENT RESULTS

**E2E** is preferred – but only at relation level.  
At tail level, both are indistinguishable.

**65.9%**  
E2E win-rate  
(Relation level)

**51.6%**  
E2E win-rate  
(Tail level)

### FAITHFULNESS RESULTS

**PIPE** is more faithful – but only at relation level.  
At tail level, both are 85% faithful.

**77.4%**  
PIPE faithful rate  
E2E 73.8% p<.001

What drives the preference: **content, not style**

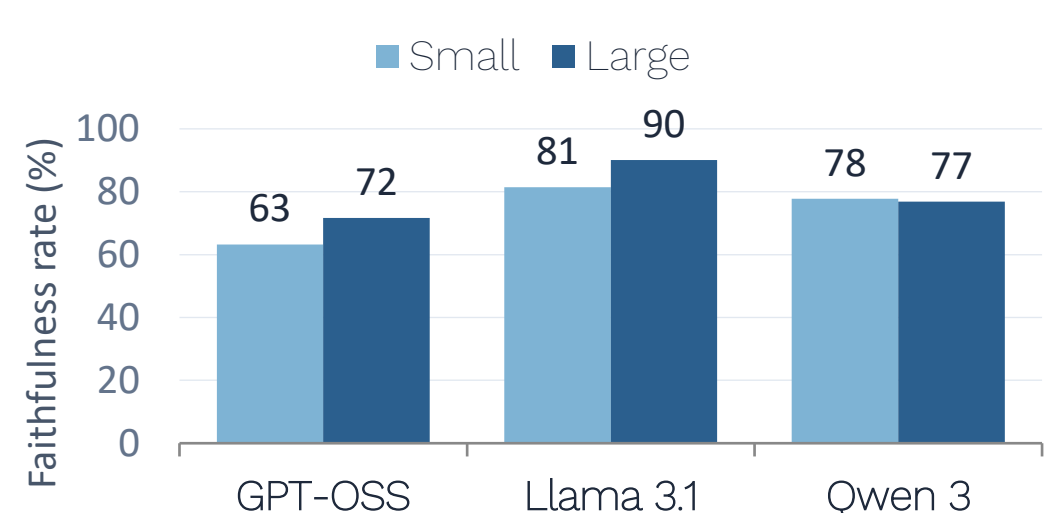
- Style: 78% ties on Coherence, 62% on Fluency.
- Content: 74.6% E2E on Theme Adherence, 65.4% on Informativeness

Size & Family:

- Large slightly preferred (gain on content)
- Arch. × size: no interaction
- E2E wins in all families:  
Llama > GPT-OSS > Qwen

|          | Hallucination | Contradiction |
|----------|---------------|---------------|
| Relation | 17.0 / 11.6   | 11.0 / 11.0   |
| Tail     | 4.0 / 3.4     | 8.7 / 10.8    |

DIVERGENCE TYPE (% HIGHLIGHTS, E2E / PIPE)  
E2E 1.36× hallucination vs PIPE (p < .001)



## 8. CONCLUSION & DISCUSSION

### CONCLUSION

Preference and faithfulness pull in different directions.

Relation- and tail-level generations behaved very differently.

At least for faithfulness, model family seems to be more important than architecture.

|              | Relation level                                         | Tail level                                                |
|--------------|--------------------------------------------------------|-----------------------------------------------------------|
| Preference   | <b>E2E preferred</b><br>65.9% win-rate *** (P ≤ 0.001) | <b>No clear winner</b><br>51.6% – not significant         |
| Faithfulness | <b>PIPE more faithful</b><br>fewer hallucinations      | <b>Both ~85% faithful</b><br>hallucination gap disappears |

### DISCUSSION

#### No human evaluation

Independent research on a tight budget; therefore, we relied entirely on LLM-as-a-judge. **Human evaluation is the natural next step.**

#### Under-used judge data

LLM-as-a-judge experiments produce rich data (rationales, confidence scores, and more) but we allocated little time to analyze it. **More qualitative work is needed.**

#### We got greedy with factors

Testing across architecture, family, generation level, and size made the design drastically more complex, making it harder to build intuition for the data. **We lean toward simpler designs.**

#### Topic granularity

Moving from broad to specific topics makes data sparsity a bigger issue, but also provides more insight, since most evaluations operate at a single level.