Comparative Analysis of Human and AI-Generated Summaries of a Domain Model Study

This document analyzes three summaries of the paper *"An Empirical Study on the Potential Usefulness of Domain Models for Completeness Checking of Requirements."* The summaries include a human-written version (MySummary), an OpenAI GPT-4o generated version (LLM_Summary), and a DeepSeek-generated version (DS_Summary). In addition to a qualitative review, the analysis incorporates quantitative metrics – including difflib similarity, TF-IDF with cosine similarity, embedding-based comparisons, RAKE keyword extraction, ROUGE, BLEU, and BERT scores – to compare the summaries and assess their overlap and differences.

Content Coverage and Emphasis

Introduction and Context:
All summaries introduce the study's central idea: using domain models (UML class diagrams) to check for omissions in requirements. The human summary clearly defines internal versus external completeness and explains that missing domain concepts in the requirements may indicate gaps. GPT-4's summary integrates these definitions within its narrative, while DeepSeek emphasizes that domain models can serve as "sensors" for omissions, highlighting nuances such as tacit knowledge and lexical mismatches that are less pronounced in the other versions.

Methodology and Scope:
Each summary describes the empirical design involving three industrial case studies – aerospace, cyber-physical sensors, and safety CMS – with requirements expressed as "shall" statements. The human summary notes that only 35 requirements per case were modeled and that experts built "feasibly complete" domain models by incorporating tacit concepts. It explains that Monte Carlo simulations were used to remove requirements (or parts thereof) and measure the resulting impact on model coverage. GPT-4's summary presents these steps succinctly but omits details like the 35-requirement subset. DeepSeek provides the most granular description, including details on noun-phrase extraction, model refinement (with counts of concepts, attributes, and associations), and explicit omission simulation procedures.

Key Findings:
All three summaries agree that domain models show near-linear sensitivity to omissions and that missing entire requirements has roughly 4.4 times more impact than missing individual details. The human summary organizes the results by noting that cases with lower redundancy (where each domain concept appears only once) are more sensitive, while higher repetition buffers against omissions. GPT-4's summary reinforces these points with precise statistics but in more generalized terms. DeepSeek is the most detailed – it reports concrete percentages (e.g., omitting 10% of requirements left 22–23% of model

elements unsupported in some cases versus 15% in another) and discusses how "tacit" elements (concepts not explicitly mentioned) can lead to false positives if not interpreted with expert judgment.

Conclusions and Implications:
All summaries conclude that domain models are valuable tools for identifying incomplete requirements. The human summary emphasizes practical implications, suggesting that domain models can serve as checklists and stressing that the cost of model construction (approximately 6–8 hours) is justified by benefits such as improved stakeholder communication and consistent terminology. GPT-4's conclusion is more concise and cautious, recommending that domain models be complemented by other techniques. DeepSeek provides a comprehensive set of implications, advising on balancing abstraction to avoid false alarms, leveraging NLP tools for keyphrase extraction, and outlining limitations and future research directions.

Structural and Organizational Differences

Organization:
The human summary is organized with clear headings (e.g., "Key Concepts," "Research Methodology," "Findings and Analysis") and bullet points to aid clarity. GPT-4's summary follows a traditional academic abstract style, using numbered sections for introduction, methodology, findings, and conclusion. DeepSeek uses an extended outline with numbered sections and detailed bullet points, which preserves a wide range of technical details and statistical data.

Language and Tone:
The human summary uses straightforward, explanatory language that is accessible to practitioners. GPT-4's summary is concise and formal, delivering precise definitions and statistical results without additional commentary. DeepSeek's summary, while formal and technical, includes rich bullet lists and detailed sub-points that capture every nuance from the study, making it ideal for readers seeking an in-depth technical overview.

Quantitative Comparison Metrics

To further understand the similarities and differences between the summaries, several quantitative text comparison methods were applied between MySummary and LLM_Summary, along with pairwise metrics across all three summaries.

Lexical and Semantic Similarity

- **Difflib Similarity:**
  The difflib analysis produced a similarity ratio of 5% for the *Introduction* and *Conclusion* sections, 10% for *Findings* and 13% for *Methodology* sections. We're getting these low ratios because *difflib* is used mostly for line-by-line comparisons.

- **TF-IDF and Cosine Similarity:**
  When vectorizing the texts with TF-IDF, the cosine similarity scores between MySummary and LLM_Summary were between 0.45 and 0.65 for different sections. These results show a more pronunciated sense of overall content similarity.
- **Embedding-Based Comparison:**
  Using pre-trained sentence embeddings to assess semantic similarity, the cosine similarity scores were between 0.86 and 0.92 for the analyzed sections between MySummary and LLM_Summary. This higher score suggests that there are relatively small semantic differences.

RAKE Keyword Extraction

RAKE keyword extraction identified a core set of keywords in both MySummary and LLM_Summary. Common keywords included:

- *requirements*
- *domain model*
- *completeness*
- *omissions*
- *sensitivity*

Notable differences were also observed:

- **MySummary** emphasized terms like *stakeholder communication* and *practical implications*.
- **LLM_Summary** focused more on terms such as *simulation*, *traceability*, and *statistical correlation*.

Standard Summarization Metrics

The following metrics were computed to assess the overall similarity among the three summaries (MySummary, LLM_Summary, and DS_Summary):

- **ROUGE Scores:**

Metric	rouge-1	rouge-2	rouge-l
My_Recall	0.08	0.01	0.08
LLM_Recall	0.07	0.02	0.06
DS_Recall	0.10	0.02	0.10
My_Precision	0.65	0.19	0.61
LLM_Precision	0.69	0.34	0.67
DS_Precision	0.53	0.17	0.52
My_F1	0.15	0.03	0.14
LLM_F1	0.12	0.04	0.12
DS_F1	0.17	0.03	0.16

- **BLEU Scores:**
  Due to the abstractive nature of the summaries, BLEU scores were very low.
  We can observe that while the while the DeepSeek summary still has a very low BLEU score, but it achieved a score with two orders of magnitude higher than the previous best -- the My Summary.
    - BLEU score for My Summary vs Original Study: **1.2e-08**
    - BLEU score for LLM Summary vs Original Study: **4.2e-11**
    - BLEU score for DeepSeek Summary vs Original Study: **1.8e-06**

- **BERTScore:**
  BERTScore F1, which measures semantic similarity using contextual embeddings, averaged around **0.82** across all pairwise comparisons, confirming that all summaries capture the core meaning of the study despite lexical and structural differences.

Final Remarks
This comparative analysis shows that while all three summaries effectively communicate the study's key findings, they cater to different audiences:

- The **human-written summary** is practical, accessible, and emphasizes actionable insights.
- The **GPT-4 summary** is concise and focused on core statistics, delivering a streamlined academic recap.
- The **DeepSeek summary** is detailed and exhaustive, preserving a wide range of technical details and methodological nuances.

All three summaries effectively capture the core findings of the study, yet they differ in style and focus. Good properties for a summary of a scientific paper include being concise and compact, containing all important aspects, and focusing on the novel ideas without reiterating well-known facts. Here, the human summary is practical and accessible, emphasizing actionable insights like cost-benefit considerations and improved communication. GPT-4's summary is concise and strictly factual, and DeepSeek's is exhaustive and detail-rich.

While ROUGE, BLEU, and BERT scores indicate high semantic similarity, they fail to capture the nuance of "novel ideas" – a critical aspect of scientific summaries. These metrics focus on surface-level lexical overlap and basic semantic content but cannot assess whether a summary highlights innovative contributions effectively. A new scoring system using LLMs could potentially address this deficiency by evaluating the degree to which summaries emphasize novel, research-specific ideas rather than merely repeating common or background information.

In essence, while quantitative metrics confirm that all summaries share a high degree of semantic overlap, qualitative differences matter. The human summary's emphasis on practicality, GPT-4's succinct precision, and DeepSeek's comprehensive detail each offer

distinct advantages. Developing an advanced evaluation method that assesses the conveyance of novel ideas would be a valuable next step for assessing scientific paper summaries more holistically.