

Diagnostic concordance of text-based AI in teledermatology: A retrospective Swiss study

Naomi M. Mueller^{1*,2}, Elisabeth V. Goessinger^{1*,2,3}, Philippe L. Gottfrois¹, Alexander Navarini^{1,2}, Christian Greis⁴

*1 Department of Dermatology, University Hospital Basel, Basel, Switzerland, *contributed equally*

2 Faculty of Medicine, University of Basel, Basel, Switzerland

3 ViDIR Group, Department of Dermatology, Medical University of Vienna, Vienna, Austria

4 Department of Dermatology, University Hospital Zurich, Zurich, Switzerland

INTRODUCTION & OBJECTIVE

Teledermatology is increasingly used in dermatological care, with growing interest in artificial intelligence (AI) for diagnostic support. Existing research in this field mainly investigates image-based and multimodal approaches using cloud-hosted large language models (LLMs). However, the diagnostic relevance of patient-reported textual information may be undervalued.

In parallel, the use of cloud-based LLMs continues to raise concerns regarding the privacy of sensitive medical data. Locally-deployed LLMs may offer a privacy-conscious alternative by avoiding the transfer of data to external servers while additionally supporting data sovereignty, model version tracking, and reproducibility.

This study explored the diagnostic contribution of non-visual patient-reported information for teledermatological decision support using a locally-hosted LLM.

METHODS

In this retrospective study, data from 1322 teledermatology consultations conducted between May 2019 and November 2024 were analysed. All consultations were managed by teledermatologists at a Swiss university hospital.

Patient input consisted of a free-text description and structured clinical metadata, including symptoms, duration, previous treatments, and demographic information. The locally-hosted LLM IBM Granite generated three ranked differential diagnoses based solely on textual input, which were compared with the corresponding teledermatologist diagnoses (**Figure 1**).

Concordance was evaluated using three outcome categories: top-1 concordance, top-3 concordance, and discordance, depending on whether the correct diagnosis was ranked first, appeared within the three suggested diagnoses, or was absent entirely. Furthermore, the association between text length and diagnostic concordance was analysed.

RESULTS

IBM Granite generated a correct first-ranked diagnosis in 20.3% of cases. In 40.1%, the correct diagnosis was included among the three suggested differentials (**Table 1**).

Median patient text length was 335 characters. Increasing text length was associated with improved diagnostic concordance (Spearman $\rho=0.078$, $p=0.004$). Text length differed significantly across concordance groups ($p=0.016$; **Figure 2**).

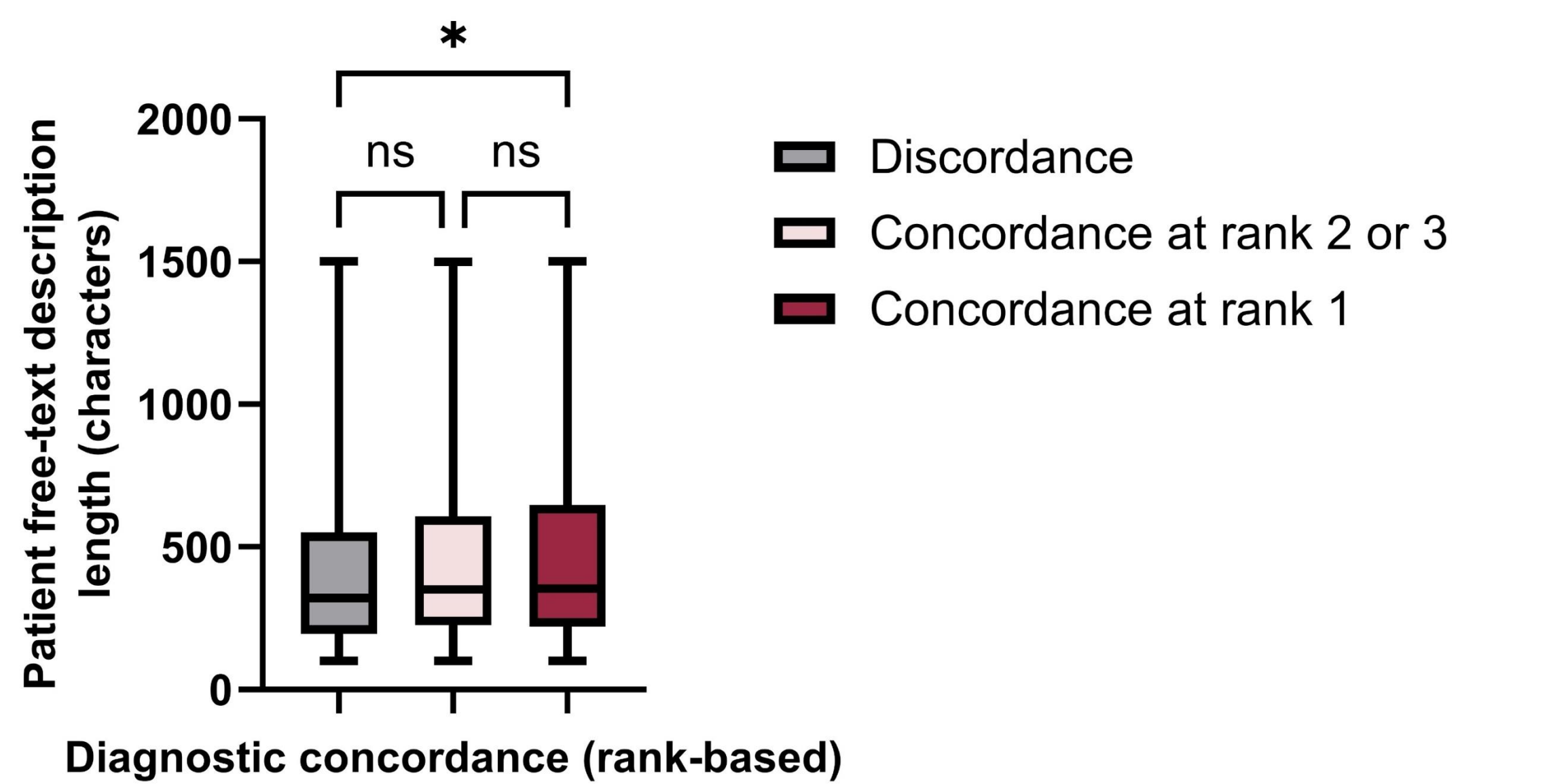


Figure 2. Length of patient free-text descriptions stratified by diagnostic concordance rank. Median and range are displayed. Discordance (n = 792), concordance at ranks 2 or 3 (n = 262), and concordance at rank 1 (n = 268). Kruskal–Wallis test with Dunn’s post hoc multiple comparison analysis: *: $p < 0.05$; ns = not significant ($p > 0.05$).

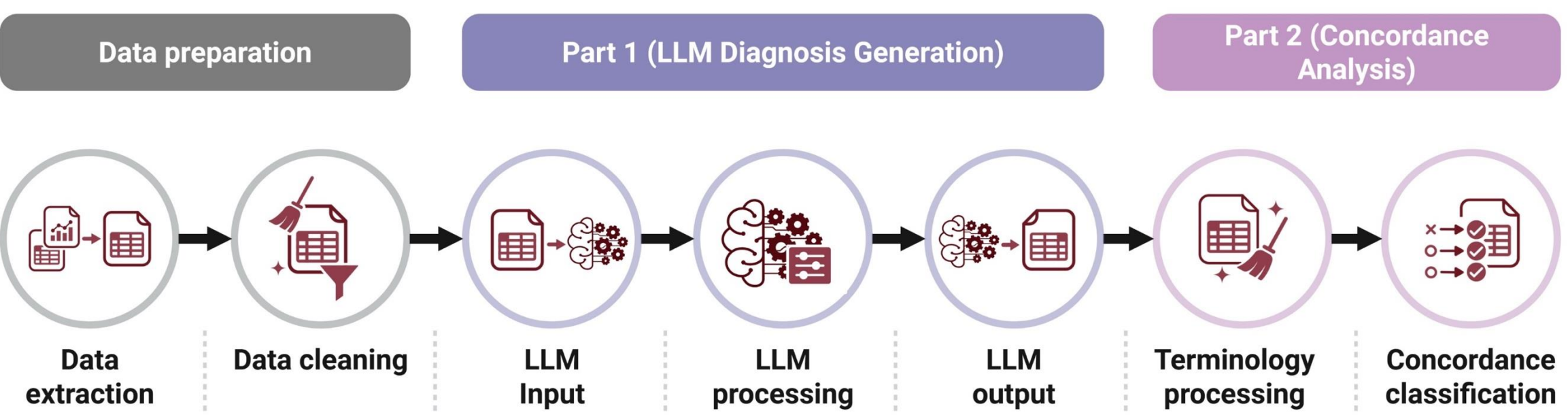


Figure 1. Overview of the two-step analysis workflow, including LLM-based diagnostic generation (Part 1) and diagnostic matching with the reference standard (Part 2).

Main category of diagnosis	n	Top 1 (n, %)	Top 3 (n, %)	Discordance (n, %)
Inflammatory skin diseases	310	135 (43.5%)	195 (62.9%)	115 (37.1%)
Benign neoplasms of the skin	302	13 (4.3%)	59 (19.5%)	243 (80.5%)
Infectious skin diseases	279	33 (11.8%)	96 (34.4%)	183 (65.6%)
Acneiform and follicular disorders	175	40 (22.9%)	102 (58.3%)	73 (41.7%)
Other diagnoses	164	25 (15.2%)	44 (26.8%)	120 (73.2%)
Pre-cancerous lesions and skin cancer	39	8 (20.5%)	16 (41.0%)	23 (59.0%)
Skin appendage	34	13 (38.2%)	15 (44.1%)	19 (55.9%)
Pigmentary disorders	13	1 (7.7%)	3 (23.1%)	10 (76.9%)
Wound and artefacts	6	0 (0.0%)	0 (0.0%)	6 (100%)

Table 1. Category-specific concordance between LLM-derived and teledermatologist diagnoses. n = number, Other diagnoses* = oral diseases, anogenital disease, pruritus, rheumatologic dermatology, vascular disorders, photodermatoses, environmental and sports-related skin diseases, urticaria, erythema, drug reactions, vasculitis and purpura, non-infectious granulomatous disorder, vesiculobullous diseases, atrophies and disorders of dermal connective tissues and disorders of the lymphatic system, other exanthema, efflorescence, other diseases of skin and subcutaneous tissue, other allergic and immunologic diseases

CONCLUSION

This study highlights the diagnostic relevance of patient-written information for LLM-based decision support. Additionally, locally-hosted LLMs may represent a privacy-oriented approach for AI-assisted diagnostics.