



スキーマの下で、LLM が公共データと照合した根拠付きの候補を返し、人が採否を決める設計において、都市環境メタデータ 175 行の検査の結果、2モデルは 99% 以上の正解率で動作した。

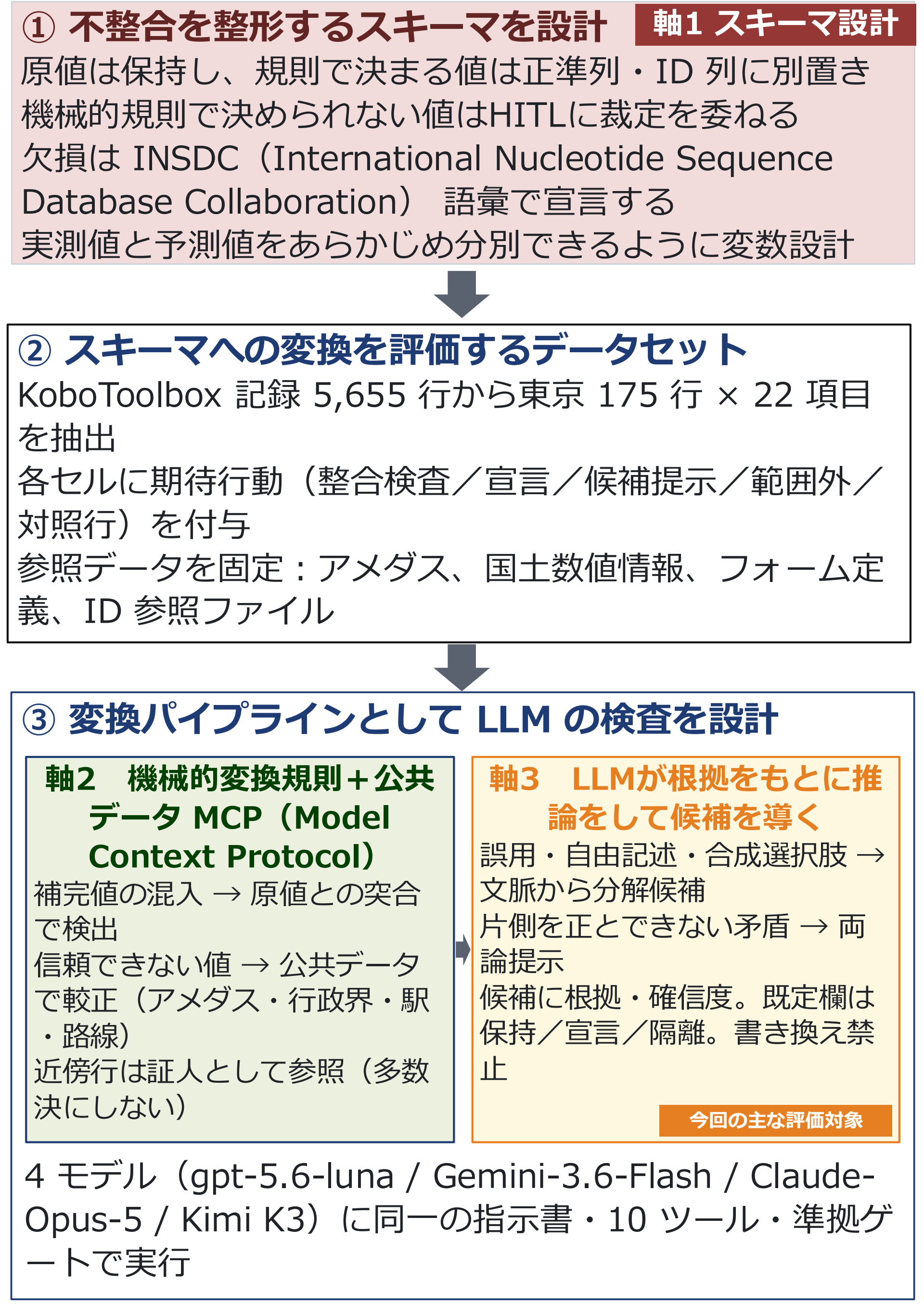
1 背景と目的

都市の人工環境の表面は、ヒトと微生物が絶えず接触する主要な接点である。MetaSUB により都市間比較の基盤が整った一方（Danko et al, 2021）、国内の Urban microbiomes in Japan では、欠損、表記ゆれ、placeholder の混在が Mix S-BE 準拠と DDBJ 登録の障壁になっている（Tsujimoto et al, 2024、Glass et al, 2014）。LLM（Large Language Model）によるメタデータ整備の有効性も報告され始めた（Ikeda et al, 2025）。

現場記録 5,655 行	→	公開BioSample 261 件
KoboToolbox／DDBJ（PRJDB14136）		
既存記録と変換工程の課題		
欠損の表し方が混在	空欄	／ 0.1, 0.1
自由記述の表記揺れ	Keio SFC	と Keio University SFC
扱いが難しい数値	気温欄に	3145
補完の根拠が消える	空欄	→ 代表値

本研究の目的
3軸からなるHITL（Human-in-the-Loop）型のパイプラインによって、既存の現場記録メタデータに残る、簡易な規則だけでは決められない値の品質診断と修正支援を、人が根拠を確認して判断できる形でどこまで実現できるかを明らかにする。

2 HITLの方法



関連検討：実験①

LLM は現場記録だけからスキーマ案を作れるか

上記の4 モデルに、サンプル 175 行（G）か全 5,655 行の分布（F）のどちらかと、共通の国際規格（MixS-BE, 国際MetaSUB）の資料を渡し、35 変数の設計書を計 8 本生成させた。条件Gでは値の事前入力の定義による予防性が支持された一方、条件Fではダミー値や未測位座標、Excel丸め残差、表記揺れの吸収などの異常検出で高い成果が得られた。モデル間でも積極的な導出・再設計を行うOpus、欠損値の回復規則を規約として考案するKimi、値の列挙に消極的なLunaや欠損値を欠損のまま置くGeminiと対応が分かれた。HITL で使ったスキーマはそれらの結果を参考に人が設計した。

6 考察、限界、展望

今回示せたこと

LLMの強みは、文脈に応じて意味を再構成できる点にある。LLM は参照ツールで根拠を集め、規則では決められない値に候補と確信度を付けて返せた。候補と根拠の作成を LLM が担い、値の決定は人が行う分担が進められた。特にOpusでは、高い正解率に加えて、推論による語彙の追加提案や、多様な観点からの問題検出が目立った。

スキーマ依拠性

HITLの実現には適切な出力規約の設計が肝要であり、規約に隙があると、LLMはもっともらしい振る舞いでその隙を埋めてしまう。項目4の未解消は、棄却時の規定値の任意記述を許可したことによる影響で発生したと推察できる。本研究では厳密なスキーマと出力規約の設計によってこうした行動をある程度防いだが、この方法を広範なメタデータへ直ちに適用できるわけではない。

人の判断が必要な範囲

参照できる情報の外にある誤り（材質の現地確認や地下地上の区別など）と、どちらが誤りか決められない矛盾は人が決める。正解は規約が定めた期待行動への一致で、値の真偽と人の作業時間は測っていない。

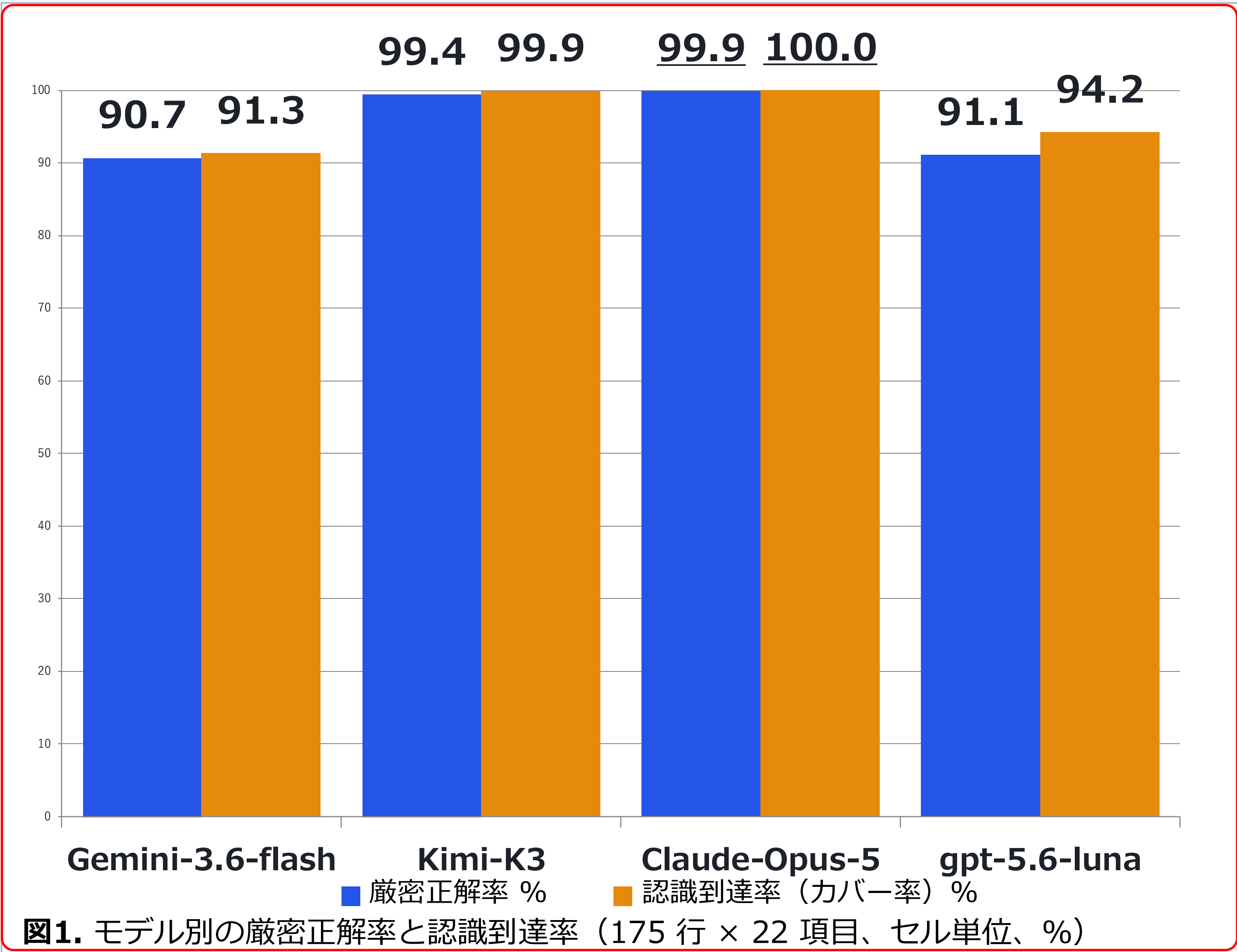
未検証の効果と次の検証

今後は、品質が担保されたメタデータを基に、①来歴付き座標と3層地名による人口密度など公共統計との統合と地域層別化、②MIM S.me.built 形式でのDDBJ実登録へ展開する。また、③フォーム自体の統合を実施し、例外値を自由記述から取得し動的にスキーマを更新するHITLサイクルを実現したい。

2.1 HITLの概念ワークフロー



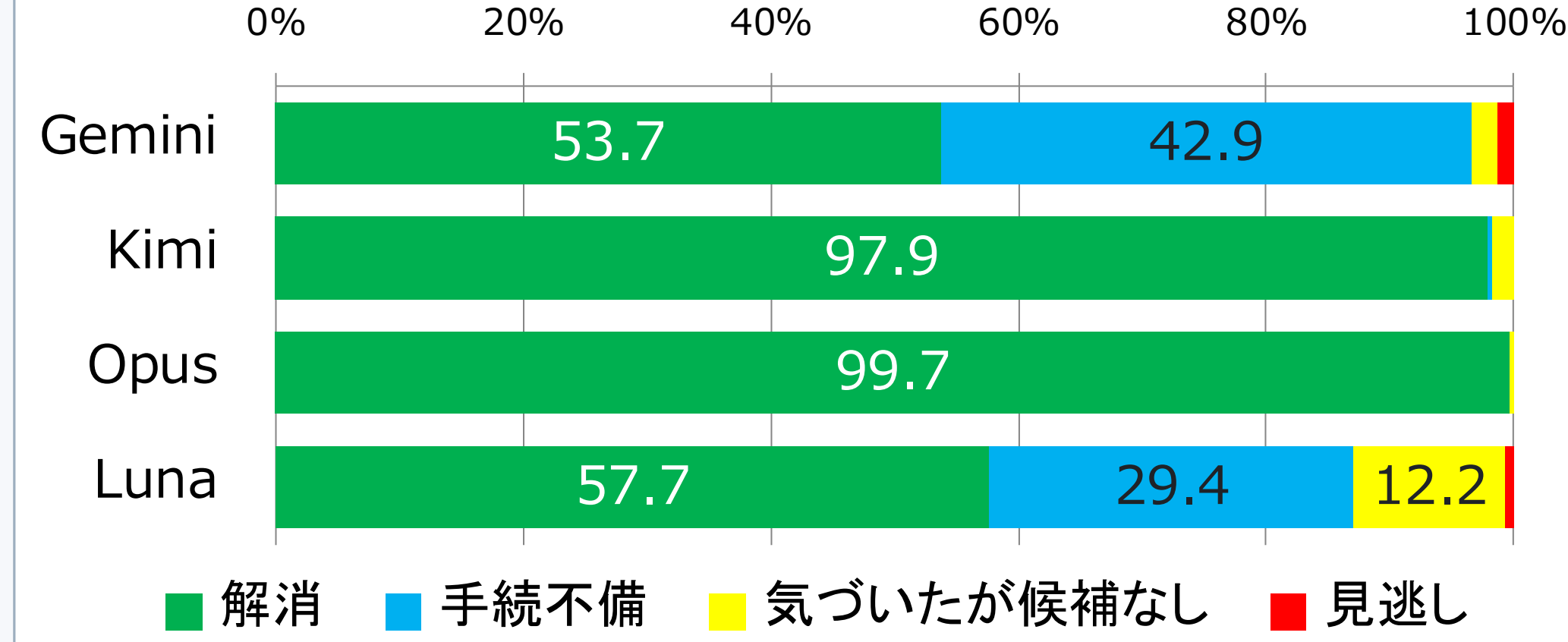
3 HITLの構成と主結果



gpt-Lunaは他モデルと同設定の状態で出力規約を遵守できないと判断できたため、JSON_schema strictにより構造化出力を強制した。その他は4モデルで同じ指示書、同じ参照ツール、同じ準拠ゲートを与えた。出力は候補と根拠に限られ、記録の値を書き換える経路はない。Gemini と Luna の未達の理由は下の 4 に示す。

4 未解消の内訳

HITL固有の介入セル：原値をそのまま受理できないセル（欠損の宣言、候補の提示、範囲外の隔離、対照検体の空欄維持を期待）。3,850 セル中 711。



Gemini と Luna の未解消はほとんどが手続不備で、高度、測位精度、都道府県名 に集中した。判断は正しいが、候補を全て却下されたときの既定欄に placeholder の複写や空欄を書いたため、規約違反になった。
Luna は合成選択肢 Station/busstop に気づいても分解候補を出せず、地点補助情報の 84 セルが残った。見逃しは Gemini 9、Luna 5 セルにとどまる。対照検体の空欄 73 セルは、4 モデルとも値で埋めずに扱えた。

5 具体例で示す支援と限界

候補を提示できた例	
記録の値	
気温 3145 → 31.45	
候補のある座標 0.1 → 代表点（来歴）	
候補のない 0.1 や 0 → 欠損値宣言	
採取時間 空欄（必須欄） → 自由記述(Less than 10 second)から補完	
表面の詳細 Escalator（既存のスキーマにない種別） → スキーマへの追加提案	

人の判断が残る例	
記録の値	LLM の提示
地点名は大井町駅、座標は戸越銀座	地点名を正とする案と座標を正とする案を並記し、断定しなかった。（共通）
りんかい線大井町駅で「地上」	地下駅のため地上への懸念を注記。根拠が内部知識であるため値の確定は求めない（Opus）
提出時刻が入力開始より前の行	現値維持 0.75 と行の隔離 0.15 を並記（Opus）。年を書き換える仮値を出した
床材質「Stone」（実際はコンクリート）	参照面に反証が無く正常と判定

参考文献

Danko D, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. Cell. 2021;184:3376–3393.
Tsujimoto M, et al. Shotgun metagenomic sequencing of swab samples from Japanese university campuses. Microbiol Resour Announc. 2024;13:e00210-24.
Glass EM, et al. MixS-BE: a MixS extension for sequence data from the built environment. ISME J. 2014;8:1–3.
Ikeda S, et al. Extraction of biological terms using LLMs enhances metadata usability in BioSample. GigaScience. 2025;14:giaf070.

謝辞

本研究はJSPS科研費（JP25K03276, JP20K10436）、および JST CREST（JPMJCR20H1）の助成を受けたものである。また、照合には気象庁のアメダス観測値、国土交通省の国土数値情報（行政区画・鉄道）、国土地理院の逆ジオコードを利用した。

