

Additional file 1

Supplementary Tables S1–S9

A Configurable LLM System for Automated Data Extraction in Meta-Analysis Using a Minimal Atomic Unit Framework

File contents. This file reports the configured validation inputs, representative records from the three extraction levels, validation-set composition, the exploratory prompt-development comparison, evaluation counts, prompt structures, and input-token lengths. Full machine-readable records are available in the archived project repository cited in the main manuscript.

Supplementary Table S1. Outcome-domain inputs used for validation

Only the outcome-domain input was configured in the independent validation experiments. Intervention, comparator, outcome type, time point, comparison type, and study-design inputs were left unrestricted.

Clinical domain	User-configured outcome-domain input
Breast cancer surgery	Total operative time; nipple necrosis rate; overall wound-related complications; length of hospital stay.
Schizophrenia treatment	Relapse; hospital admission; treatment discontinuation; extrapyramidal symptoms; weight gain; hyperglycemia; hyperprolactinaemia; death; suicide attempts.
Cardiovascular disease prevention	Major adverse cardiovascular events; myocardial infarction; stroke or cerebrovascular events; cardiovascular mortality; all-cause mortality.
Type 2 diabetes management	Myocardial infarction; stroke; retinopathy; photocoagulation; visual deterioration or blindness; microalbuminuria; renal failure; all-cause or cardiovascular death; congestive heart failure; peripheral vascular events; neuropathy; amputation; severe hypoglycemia.
Postoperative orthopedic rehabilitation	Pain intensity (VAS); function indices (DSI, WOMAC, KOOS); balance; postural stability; proprioception; knee range of motion; five-times sit-to-stand; gait; walking speed; LEFS; ABC scale; HSS score; rehabilitation compliance; readmission; patient satisfaction; healthcare costs.

Supplementary Table S2. Representative Study_Characteristics record

The table shows one representative record from the current public JSON output. Empty strings indicate fields not reported or not applicable. Long narrative values are abridged for display; complete values and source-evidence columns are available in the repository.

Field	Extracted value
study_id	text3-5
title	A Randomized Trial of Robotic Mastectomy versus Open Surgery in Women With Breast Cancer or BRCA Mutation
authors	Antonio Toesca; Claudia Sangalli; Patrick Maisonneuve; Giulia Massari; Antonia Girardi; Jennifer L. Baker; et al. (24 authors in the complete output)
year	2024
journal	Annals of Surgery
country	Italy
doi	10.1097/SLA.0000000000004969
study_design	Phase III, open-label, single-center RCT. Randomization used an automated dynamic allocation system.
inclusion_criteria	Adult women with invasive breast cancer, ductal carcinoma in situ, or a pathogenic BRCA1/2 mutation who were candidates for nipple-sparing mastectomy with immediate reconstruction.
exclusion_criteria	Included nodal metastasis, inflammatory or recurrent cancer, skin or nipple involvement, prior thoracic radiotherapy, pregnancy, ASA score above 2, uncontrolled diabetes, heavy smoking, and large breast volume.
study_setting	Single center, European Institute of Oncology, Milan, Italy.
source_of_recruitment	European Institute of Oncology, Milan, Italy.
number_of_participants_randomized	80
number_of_participants_analyzed	80
dropout_reasons	One robotic-arm participant died before one year; one open-arm participant had intraoperative nipple-areola complex removal; some questionnaire items were unanswered.
meets_inclusion_criteria	true
exclusion_reason	Empty.

Supplementary Table S3. Representative Group_Baseline records

Field	Open nipple-sparing mastectomy	Robotic nipple-sparing mastectomy
study_id	text3-5	text3-5
arm_type	control	intervention
timepoint	Baseline, 1 month, 6 months, 12 months	Baseline, 1 month, 6 months, 12 months

Field	Open nipple-sparing mastectomy	Robotic nipple-sparing mastectomy
subgroup	Empty	Empty
total_participants	40	40
gender_male_count	0	0
gender_female_count	40	40
mean_age	45.5	44.5
age_sd	Empty	Empty
intervention_details	Empty	Robotic nipple-sparing mastectomy with an off-breast incision and immediate reconstruction using a permanent implant or tissue expander.
intervention_duration	Empty	Empty
intervention_frequency	Empty	Empty
control_details	Open nipple-sparing mastectomy with a breast incision and immediate reconstruction using a permanent implant or tissue expander.	Empty
adherence_to_control	Empty	Empty
adherence_to_intervention	Empty	Empty
number_of_participants_analyzed	40	40

Supplementary Table S4. Representative Atomic_Outcomes records

This compact view shows the main analytical fields for eight records nested under the two arm-level records in the current public JSON output for text3-5. The complete 31-field records and field-level source evidence are provided in the repository. MD, mean difference; RR, risk ratio.

Arm	Outcome	Type	Unit	Mean (SD)	Events/total	Time point	Effect	<i>P</i> value
Open	Total surgery time	Continuous	Hours	2.3 (0.8)	Empty	Peri-operative	MD 1.3	< 0.0001
Robotic	Total surgery time	Continuous	Hours	3.6 (0.8)	Empty	Peri-operative	MD 1.3	< 0.0001
Open	Nipple necrosis	Dichotomous	Percent	Empty	2/40	Within 3 months	RR, value empty	0.49
Robotic	Nipple necrosis	Dichotomous	Percent	Empty	0/40	Within 3 months	RR 0	Empty
Open	Wound-related complications	Dichotomous	Percent	Empty	20/40	Within 3 months	RR, value empty	0.11
Robotic	Wound-related complications	Dichotomous	Percent	Empty	12/40	Within 3 months	RR 0.6	0.11
Open	Length of hospital stay	Continuous	Days	2.4 (0.6)	Empty	Peri-operative	MD 0.1	0.04
Robotic	Length of hospital stay	Continuous	Days	2.3 (1.2)	Empty	Peri-operative	MD −0.1	0.04

All records had `study_id` = text3-5, comparison type = between-group post-intervention difference, time-point role = post, and `higher_is_better` = false. Comparison targets linked the open and robotic arms. Empty values reproduce null fields and were not imputed. Effect-size confidence intervals were empty for these records.

Supplementary Table S5. Validation-set composition and exclusions

Domain	Planned	Included	Excluded	Reason for exclusion	Outcomes
Breast cancer surgery	5	5	0	None.	Total operative time; nipple necrosis; wound-related complications; length of stay.
Schizophrenia treatment	56	44	12	Full text unavailable for 12 reports.	Relapse; hospital admission; discontinuation; extrapyramidal symptoms; weight gain; hyperglycemia; hyperprolactinaemia; death; suicide attempts.
Cardiovascular disease prevention	21	19	2	Two reports lacked prespecified core outcomes for the simulated meta-analysis.	Major adverse cardiovascular events; myocardial infarction; stroke or cerebrovascular events; cardiovascular mortality; all-cause mortality.
Type 2 diabetes management	13	11	2	Full text unavailable for two reports.	Myocardial infarction; stroke; retinopathy; photocoagulation; visual deterioration or blindness; microalbuminuria; renal failure; death; heart failure; peripheral vascular events; neuropathy; amputation; severe hypoglycemia.
Postoperative orthopedic rehabilitation	9	9	0	None.	Pain; function; balance; postural stability; proprioception; knee range of motion; sit-to-stand; gait; walking speed; LEFS; ABC; HSS; rehabilitation compliance; readmission; satisfaction; costs.
Total	104	88	16		51 outcome measures across five domains.

Supplementary Table S6. Exploratory MAU versus non-MAU comparison

The same three reports were used for prompt development and this comparison. Both prompts used GPT-4.1 at temperature 0, the same extraction scope and fields, and the same source-tracing requirement. The comparison is therefore descriptive of the development dataset rather than an independent estimate of MAU effectiveness.

Count	Non-MAU	MAU
Correctly extracted cells	248	1113
Evaluated cells	1160	1160
Valid model responses	267	1113
Gold-standard targets	1160	1160
Hallucinated responses	0	0
Precision (95% CI)	0.9288 (0.8915–0.9540)	1.0000 (0.9966–1.0000)
Recall (95% CI)	0.2138 (0.1912–0.2383)	0.9595 (0.9465–0.9694)
F1 (95% CI)	0.3476 (0.3159–0.3791)	0.9793 (0.9735–0.9851)
Hallucination rate (95% CI)	0 (0–0.0033)	0 (0–0.0033)

Supplementary Table S7. Evaluation counts overall and by domain

Count	Overall	Breast	Schizophrenia	Cardiovascular	Diabetes	Orthopedic
Correctly extracted (C)	25415	1027	13625	4186	1706	4871
Evaluated cells	25887	1032	13762	4258	1859	4976
Valid model responses ($C + I + H$)	25503	1031	13665	4201	1735	4871
Gold-standard targets ($C + I + M$)	25832	1028	13730	4245	1853	4976
Hallucinated (H)	55	4	32	13	6	0
Missing (M)	384	1	97	57	124	105
Incorrect (I)	33	0	8	2	23	0

Definitions: correct, model value identical to the adjudicated gold standard; incorrect, both cells nonempty but model content wrong; hallucinated, nonempty model value with an empty gold-standard cell; missing, empty model cell with a nonempty gold-standard cell.

Supplementary Table S8. Abridged core prompt structures

This table is a concise English-language comparison rather than a verbatim translation of every production instruction. The repository contains the executable MAU prompts, and de-identified historical prompt templates extracted directly from the surviving non-MAU implementation are reproduced in Additional file 2. The non-MAU comparator requested the same scope, target fields, JSON format, and source evidence but did not decompose the task into MAUs.

Component	MAU setting	Non-MAU setting
Model configuration	GPT-4.1; temperature 0; no fine-tuning.	Identical.
Input	Preprocessed report text plus the same optional user parameters.	Identical.
Extraction scope	Study information, study arms, time points, subgroups, outcomes, descriptive statistics, inferential statistics, and analysis denominators.	Identical.
Target fields	Same fields in Study_Characteristics , Group_Baseline , and Atomic_Outcomes .	Identical.
Source tracing	Each populated field must include a verbatim supporting sentence; unreported values remain null.	Identical.
Output	Valid JSON using the same field names and null-value rules.	Identical.
Task organization	First identify study-level information, then create explicit arm–time point–subgroup units, then extract each outcome within the corresponding unit and comparison context.	Request the same fields from the report without explicit arm–time point–subgroup–outcome MAU decomposition.
Key constraint	Never transfer a value across arms or time points; descriptive and effect-size confidence intervals must remain distinct; comparison target must be explicit.	Same semantic constraints and fields, but no MAU-by-MAU execution structure.

MAU production-prompt outline

1. Extract study-level fields and their supporting source sentences.
2. Identify every eligible study arm. For each arm, record role, time point, subgroup, population, intervention or comparator details, adherence, and analyzed sample size.

3. For each arm-level unit, extract eligible outcomes. Keep arm, time point, subgroup, outcome name, comparison type, and comparison targets explicit.
4. Do not infer missing values or reuse a value from another arm. Keep descriptive mean CIs separate from effect-size CIs.
5. Return valid JSON only; each field contains a **value** and **source**, with null used when the report does not support a value.

Supplementary Table S9. Input-token lengths in the validation dataset

Token counts were calculated with the `o200k_base` tokenizer after preprocessing.

Domain	Reports	Mean	Minimum	Maximum
Breast cancer surgery	5	10805	8479	13625
Schizophrenia treatment	44	8147	2685	14988
Cardiovascular disease prevention	19	10092	6292	14706
Type 2 diabetes management	11	12268	9125	15058
Postoperative orthopedic rehabilitation	9	6604	4710	9538
Overall	88	9075	2685	15058