Assessing Accuracy of Unsupervised Large Language Model Tools for Identifying Key Criteria Supporting Abstract Screening

Jeremy Schneider[1], Kevin Kallmes[1]

https://nested-knowledge.com

Acceptance Code: MSR38

Unsupervised LLMs Summary

Why does it matter? Systematic reviews often require researchers to read hundreds of article summaries (abstracts) to decide which studies to include. This process takes a lot of time. Artificial intelligence (AI) tools can help by identifying key details in abstracts, but most rely on supervised models that require human training data.

What did we do? We tested a new **unsupervised AI tool**, called **Adaptive Smart Tags (ASTs)**, which uses large language models (LLMs) to automatically recognize important information like study type, population, intervention, and outcomes without being trained on specific examples.

What did we find? When tested on 27 clinical trial abstracts about GLP-1 receptor agonists for weight loss, ASTs accurately identified nearly all key information. This suggests that unsupervised AI tools could make systematic review screening much faster and more efficient, while maintaining accuracy.

Background

- This abstract evaluates the accuracy and reliability of ASTs in identifying key study criteria without prior training.
- Systematic literature reviews are essential for summarizing scientific evidence, but screening abstracts is time-intensive and prone to human error. Large language models (LLMs) can help automate parts of this process by recognizing study details in text.
- While supervised AI systems trained with labeled data have been tested for this purpose, unsupervised LLM approaches are less explored. The Adaptive Smart Tags (AST) feature in the Nested Knowledge Platform automatically classifies qualitative information from abstracts using hierarchical tagging system.

Methods

- **Searches and Screening:** An AI-assisted search identified 419 studies published since 2017 evaluating GLP-1 receptor agonists (GLP-1 RAs) for weight loss. The first 50 studies were dual screened by two human researchers to train a machine-learning screening system (Robot Screener, Nested Knowledge). The rest were screened using a hybrid human-AI process where one researcher and Robot Screener served as reviewer-level screeners. Discrepancies were solved by consensus.
- **Study Selection:** 27 randomized controlled trial (RCT) abstracts were included where weight loss was the primary outcome.
- **Tagging Process:** ASTs analyzed each abstract using a tag hierarchy that covered five main dimensions: Study type, Population, Intervention, Comparators, and Outcomes.
- **Evaluation:** Each tag was manually classified as:
  - Correct (True Positive)
  - Missing (False Negative)
  - Incorrect (False Positive)
  - Not in abstract (True Negative)
  - Partially correct (.5 True Positive, .5 True Negative
- AI Performance was measured using Precision, Recall, and F1 scores across seven categories representing critical Population, Intervention, Comparator, and Outcomes categories: Obesity, Study Type, Obesity, Gender, GLP-1 RA, Comparators, Clinical Outcomes, Safety Outcomes.

Sample Studies

Study Title	Study Objective	Study Title	Study Objective	Study Title	Study Objective
New association of bone morphogenetic protein 4 concentrations with fat distribution in obesity and Exenatide intervention on it.	Examine BMP4 levels and fat distribution before/after exenatide.	A 6-month randomized, double-blind, placebo-controlled trial of weekly exenatide in adolescents with obesity.	Test safety and effect of weekly exenatide.	Weight Loss Outcomes Among Early High Responders to Exenatide Treatment: A Randomized, Placebo Controlled Study in Overweight and Obese Women.	Identify predictors of strong early exenatide response.
Effects of liraglutide on weight, satiation, and gastric functions in obesity: a randomized, placebo-controlled pilot trial.	Test liraglutide effects on weight and gastric emptying.	Changes in health-related quality of life with intensive behavioural therapy combined with liraglutide 3.0â€‰mg per day.	Assess HRQOL change with IBT + liraglutide.	Exenatide for weight-loss maintenance in adolescents with severe obesity: A randomized, placebo-controlled trial.	Assess exenatide for weight maintenance.
Patients with Obesity Caused by Melanocortin-4 Receptor Mutations Can Be Treated with a Glucagon-like Peptide-1 Receptor Agonist.	Assess GLP-1 RA efficacy in MC4R obesity.	GLP-1 Analog Modulates Appetite, Taste Preference, Gut Hormones, and Regional Body Fat Stores in Adults with Obesity.	Examine appetite cues and fat distribution changes.	Effects of liraglutide on gastrointestinal functions and weight in obesity: A randomized clinical and pharmacogenomic trial.	Study GI motility and genetic predictors.
Effects of liraglutide on appetite, food preoccupation, and food liking: results of a randomized controlled trial.	Evaluate liraglutide's impact on appetite and food reward.	Liraglutide after diet-induced weight loss for pain and weight control in knee osteoarthritis: a randomized controlled trial.	Evaluate liraglutide post-weight-loss in knee OA.	Two-year effect of semaglutide 2.4Â mg on control of eating in adults with overweight/obesity: STEP 5.	Evaluate long-term eating control and weight outcomes.
Effects of liraglutide plus phentermine in adults with obesity following 1 year of treatment by liraglutide alone: A randomized placebo-controlled pilot trial.	Test whether adding phentermine enhances weight loss.	Effects of Dietary Self-Monitoring, Physical Activity, Liraglutide 3.0 mg, and Placebo on Weight Loss in the SCALE IBT Trial.	Compare weight effects of liraglutide vs placebo during IBT.	Preliminary observations on the administration of a glucagon-like peptide-1 receptor agonist on body weight and select carbohydrate endpoints in persons with spinal cord injury: A controlled case series.	Observe metabolic changes in SCI patients.
Randomized trial comparing effects of weight loss by liraglutide with lifestyle modification in non-alcoholic fatty liver disease.	Compare liraglutide vs. lifestyle modification.	Effectiveness of Combining Anti-obesity Medication With an Employer-Based Weight Management Program for Treatment of Obesity: A Randomized Clinical Trial.	Test medication plus workplace program.	Intragastric injection botulinum toxin A for obesity management with or without liraglutide.	Test benefit of liraglutide added to gastric Botox.

Results

Metric	Mean ± SD
Precision	0.965 ± 0.067
Recall	1.000 ± 0.000
F1-score	0.981 ± 0.036

Across the seven evaluated tagging categories, Adaptive Smart Tags showed **high and consistent performance**. **Precision** remained strong, indicating that tags applied by the system were overwhelmingly correct, while **recall** reached **100%**, meaning ASTs successfully detected all relevant key elements present in the abstracts. The resulting **F1-scores** demonstrated a strong balance between precision and recall, and the overall **accuracy** of tagging decisions was high.

Discussion

- This study demonstrated that Adaptive Smart Tags, an unsupervised large language model (LLM) system, can accurately extract and categorize key study information from clinical trial abstracts. The strong performance across all metrics indicates that ASTs can identify relevant details such as study type, population, interventions, and outcomes without requiring task-specific training.
- Compared with supervised or rule-based models, unsupervised LLM-based systems as practiced here offer distinct advantages:
  - Generalizability: Can be applied to new topics and research areas with minimal setup
  - Efficiency: Reduces human workload by automatically finding key elements across abstracts
  - Transparency and adaptability: ASTs use a defined tagging hierarchy with traceable extractions, making outputs interpretable and easy to validate
- Based on the high accuracy of this approach in early validation, a dedicated tool was integrated also as an automated Screening AI in the Nested Knowledge platform (**Smart Screener**).
- Some limitations were observed: Minor inaccuracies often stemmed from ambiguous phrasing in abstracts or overlapping tag definitions (for example when intervention and comparator terms are similar). Future work could focus on refining the granularity of tags, expanding evaluation across large datasets, and comparing performance across different therapeutic areas.

Conclusion

Unsupervised LLM-based tools like Adaptive Smart Tags show strong promise for automating systematic review screening. In this study, ASTs achieved near perfect recall and high precision in identifying key study features from clinical trial abstracts. These findings suggest that unsupervised AI systems can accelerate abstract screening and data extraction, reduce reviewer burden, and maintain high quality evidence synthesis.

[1]Nested Knowledge St Paul, MN USA
Disclosures: Kevin Kallmes is an employee and equity holder of Nested Knowledge, Jeremy Schneider is a former employee of Nested Knowledge.

Presented at ISPOR Europe 2025 | 9-12 November 2025 | Glasgow, Scotland
Contact: kevin.kallmes@nested-knowledge.com | www.nested-knowledge.com