Human-in-the-Loop LLM-Assisted Catalyst Design for CO_2-to-DME Conversion: From Hypothesis Generation to Experimental Validation.

W. Korir[1], A. Aho[1], P. Mäki-Arvela[1], E. Yu. Mourzine[2] D. Yu. Murzin[1]

[1] Department of Engineering and Information Technology, Åbo Akademi University, Turku/Åbo, Finland;
[2] Department of Computer Information Systems, University of Akron, Akron, OH, USA

**Corresponding author: wayne.korir@abo.fi*

Highlights

- Eight-phase human-in-the-loop LLM workflow generated six bifunctional CO_2-to-DME catalyst recipes.
- Workflow failure modes, including an undetected synthesis error, were formally documented.
- Experimental results pinpointed the structural cause of low catalyst activity, enabling targeted redesign.
- Feeding experimental results back to large language model (LLM) produced an improved, evidence-based catalyst recipe.

1. Introduction

Direct CO_2 hydrogenation to dimethyl ether (DME) over bifunctional catalysts offers a single-reactor route integrating methanol synthesis and dehydration. No dominant catalyst system has emerged; design requires simultaneous optimisation of hydrogenation activity, dehydration selectivity, and resistance to water inhibition across both functions. Large language models (LLMs) offer potential as structured design tools through rapid literature synthesis and cross-study pattern recognition at scales impractical for individual researchers. However, while general LLM failure mode taxonomies exist [4], systematic documentation of such failure modes within LLM-assisted catalyst design workflows specifically remains absent from the published record, to the best of our knowledge. This work presents a complete human-in-the-loop LLM workflow (Claude Sonnet 4.6, Anthropic) for CO_2-to-DME bifunctional catalyst design, with the full cycle from hypothesis generation through experimental validation and feedback-driven redesign demonstrated and critically evaluated.

2. Methods

The workflow comprised eight sequential phases: (1–2) context establishment and operating protocol (LLM assigned senior chemist persona with citation requirements and prohibited behaviours); (3–5) iterative literature retrieval across three reaction categories (CO_2 hydrogenation, methanol dehydration, bifunctional systems; ~35 papers per category) with manual expert validation; (6) cross-literature synthesis and descriptor hierarchy extraction; (7) generation of six catalyst formulations across three risk tiers (Benchmark, Hypothesis-Testing, Synthesis-of-Concepts); (8) expert panel review (N = 4), selecting two recipes for synthesis.

Recipe 1 (tested here): In_2O_3-ZrO_2 (co-precipitation, In:Zr = 1:4 molar, calcined 650°C) physically mixed with H-Ferrierite (1:1 mass ratio, 0.5 g total). Conditions: 30 bar, GHSV 2000 h^{-1}, H_2:CO_2 = 3:1, 200–300°C, fixed-bed. Characterisation: XRD, N_2 physisorption (BET), FTIR pyridine acidity. Workflow failure modes were documented systematically throughout execution.

3. Results and Discussion

Characterisation revealed a critical structural deviation. XRD showed tetragonal ZrO_2 with no detectable crystalline In_2O_3 (Fig. 1), co-precipitation caused suspected In^{3+} lattice incorporation rather than discrete surface In_2O_3 formation. This structural outcome is likely a consequence of the AI-selected synthesis route, consistent with the known tendency of alkaline precipitation routes to yield metastable tetragonal ZrO_2 at low-to-moderate calcination temperatures; the 650°C calcination chosen to drive monoclinic phase formation simultaneously caused severe sintering. BET surface area was 9 m^2/g, 8.6 times below the literature target of 80–133 m^2/g [1]. The AI workflow flagged the surface area penalty associated with higher calcination temperatures but did not resolve the route-level incompatibility; the human expert review initially failed to intercept this trade-off before synthesis. H-Ferrierite was confirmed to exhibit strong Brønsted acidity (B/L = 14, 0.44 mmol/g at 250°C desorption temperature using FTIR with pyridine).

Catalytic performance was fully consistent with structural findings. CO_2 conversion increased from 1.6% (225°C) to 12.0% (300°C), driven by the reverse water-gas shift reaction rather than methanol synthesis, the absence of In_2O_3 oxygen vacancy sites precluded the CO_2 activation mechanism established by Martin et al. [1]. CO selectivity dominated at 57–69% across all temperatures. DME was produced with selectivity of 28.6% at 225°C, confirming ferrierite dehydration was unaffected by the low hydrogenation activity. DME space-time yield peaked at 0.033 mmol/(g·h) at 250°C, approximately 135× below the $Cu\text{-}ZnO\text{-}Al_2O_3$/Ferrierite benchmark [2].

Among the documented failure modes, two warrant particular attention. First, the synthesis route selection failure: the AI chose co-precipitation without fully resolving the phase-temperature trade-off inherent to that route, and human expert review did not intercept this before synthesis, demonstrating that human oversight must extend to synthesis chemistry, not only to literature interpretation and recipe generation. Second, the closed-loop failure: intra-session context persistence proved insufficient to connect experimental results back to the original recipe generation context. However, human-mediated transfer of experimental findings into a purpose-built new session successfully closed the loop, enabling mechanistic interpretation and Recipe 2 design. This demonstrates that loop closure in human-in-the-loop architectures does not require continuous session memory; it requires rather a deliberate context reconstruction across sessions, identifying context building as a critical methodological discipline in LLM-assisted research workflows.

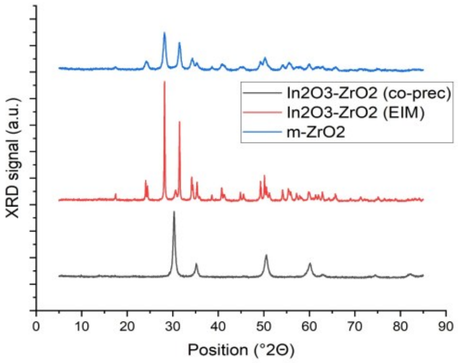

Figure 1. XRD patterns for co-precipitated $In_2O_3\text{-}ZrO_2$ (black, Recipe 1), evaporative impregnation reference (red), and monoclinic ZrO_2 reference (blue). The co-precipitated sample shows a single broad peak at ~30° 2θ (tetragonal ZrO_2) with no In_2O_3 reflections, consistent with suspected In^{3+} lattice incorporation or amorphous In_2O_3 below XRD detection limits.

4. Conclusions

A complete human-in-the-loop LLM design cycle has been demonstrated for bifunctional CO_2-to-DME catalyst design: structured literature synthesis → recipe generation → experimental validation → mechanistic feedback → redesign. Structural characterisation (XRD, BET) fully explains Recipe 1 catalytic performance at the mechanistic level, validating the diagnostic power of the feedback loop. Workflow failure modes are formally documented, including an AI synthesis route selection failure not caught by a human review, addressing a gap in the published literature on LLM-assisted catalyst design, where operational failure modes of specific workflow executions remain largely undocumented [4]. Human-mediated context reconstruction between sessions is identified as a viable mechanism for closing iterative design loops, independent of continuous session memory.

References

1. O. Martin, A.J. Martín, C. Mondelli, S. Mitchell, T.F. Segawa, R. Hauert, C. Drouilly, D. Curulla-Ferré, J. Pérez-Ramírez, Angew. Chem. Int. Ed. 55 (2016) 6261–6265.
2. H. Ham, N.T. Xuan, H.S. Jung, J. Kim, H.-S. Roh, J.W. Bae, Catal. Today 369 (2021) 112–122.
3. S. Renda, J. Soler, M. Menéndez, J. Herguido, Appl. Catal. A: Gen. 710 (2026) 120682.
4. V. Vinay, Failure modes in LLM systems: a system-level taxonomy, ArXiv:2511.19933 (2025).

Keywords CO_2 hydrogenation; bifunctional catalyst; large language model; human-in-the-loop design; DME synthesis