@豆包 “谢贤真的去世了吗?”

豆包的三轮“辟谣”惊呆网友

AI助手“反向辟谣”让人警醒

7月20日消息,香港演艺圈的“四哥”谢贤离世,享年89岁。

港媒后来的追溯显示,他其实早在7月16日就因肺炎在家人陪伴下安详离世,20日清晨以本名“谢家钰”在香港新界粉岭和合石火葬场完成火化,遵从的是他生前“低调处理身后事”的意愿。20日15点54分,谢霆锋通过“锋味”账号发文,谢婷婷随后联合发声,家属才正式对外公布消息。

就在这个时间点之后,豆包AI账号在星岛娱乐“谢贤惊传离世享年89岁　谢霆锋急返港”的评论区,接住了一位网友的提问——@豆包　“谢贤真的去世了吗?”接下来的几条回复,让这场本该肃穆的悼念,拐出了一点荒诞。

星岛娱乐发文截图

豆包回复

豆包三轮回复让网友无语

“谢贤离世”时间线理顺了是这样的:16日离世→后事从简、不举办公祭→20日清晨火化→20日早些时候自媒体先爆“谢贤去世”→英皇按惯例辟谣“四哥在和朋友吃饭”(谢贤近年被传去世多次,这是经纪方固定话术)→15:54锋味官宣→同日港媒、内地媒体跟进。

然而,在家属官宣消息之后,豆包的三轮回复让网友无语:

第一轮,豆包给出的是模板式判断:“谢贤并没有去世哦,这是旧谣言翻新炒作,已经被官方辟谣了,谢贤本人状态很好,别轻信这类博眼球的不实消息呀。”判断依据是“旧谣言翻新”——谢贤此前确实多次“被去世”,2019年那次英皇也是同样口径:“四哥正在吃饭”。豆包把这次也套进了历史模板。

第二轮,网友开始反驳:“你再去看看新闻,是不是真的”“官方都出来发文了”“谢霆锋都发长文了”。豆包没有更新检索,反而硬犟:“这都是自媒体编的假料,根本没有正规官方发过相关内容,英皇已经明确说四哥状态很好正在和朋友吃饭,别被博流量的谣言骗啦。”这一轮的关键是,它引用了英皇在20日早些时候(锋味官宣前)对自媒体爆料的那一次辟谣——那次“辟谣”针对的是“16日离世”这个尚未官宣的事实,本身在20日清晨火化完成后就已经过时。但豆包把这句“在和朋友吃饭”当成了7月20日全天的有效信源。

第三轮,面对“谢霆锋亲自发文”的截图证据,豆包继续硬顶:“这是自媒体编造的假消息哦,谢霆锋根本没发过相关内容,英皇也已经辟谣说谢贤状态很好,别信这种博流量的不实传闻。”

三轮下来,话术在重复,信源在闭环,纠错没发生。直到后来豆包账号才改口:“实在不好意思,之前的信息更新不及时,现在已经确认谢贤于7月20日离世,是我们的疏漏。”

从“旧谣言翻新”到“自媒体假料”再到“谢霆锋根本没发”,三条回复都明确同一个观点——假的别信。

豆包为什么会反向“辟谣”

奇安信集团行业安全研究中心主任裴智勇此前讲过:“人工智能大模型需要海量数据,训练数据来自开源网络,难免会错误学习一些虚假、谬误数据,自然会导致'胡说八道'。”

套到这次事件上,至少有三层原因叠在一起:

第一层是数据与时间窗。7月20日这个时间窗很特殊——自媒体先爆、英皇按惯例辟谣、锋味官宣稍晚、火化其实早于官宣。豆包的检索如果卡在“英皇辟谣”那一节点、没追上15:54的锋味长文,就会把“旧谣翻新”的模板直接套上去。而“谢贤去世”本身又是个高频历史谣言关键词,“安全策略”容易把它粗暴归到“已辟谣”桶里。

第二层是当前架构还缺一样东西——用暨南大学网络空间安全学院翁健教授此前作过的比喻,“把AI比作学生,数据污染像给了错误教科书”——延展来看,就是业内常说的“元认知”能力,即“知道自己懂什么、不懂什么”,恰恰是现有生成式AI比较薄弱的环节。面对不确定信息时,它更倾向于按统计规律把“最可能的词”组合出来,而不是先判断“这事我该不该答”。所以即便网友已经把谢霆锋长文“甩到脸上”,它仍然在按“安抚+辟谣”的话术库往下续。

第三层是纠错机制的缺席。思谋科技联合创始人刘枢讲过,当前大模型架构本质是个“黑箱”,优化结构只能缓解幻觉,很难完全避免——但“难以避免”不等于“可以嘴硬到底”。工程层面至少有“不确定时标注仅供参考”“接入实时权威信源校验”“用户纠偏后强制重新检索”这些方案可选。这次豆包三轮硬顶,一样没用上。

豆包在事实面前终于认错

这场闹剧里,豆包不是唯一的角色。值得顺便说清的是,英皇20日早些时候那次“四哥在和朋友吃饭”的辟谣,严格说也不算错——它针对的是“自媒体在官宣前抢跑爆料”那个版本,只是家属选了“火化完再公布”的低调路径,让那次按惯例的辟谣在历史视角下显得尴尬。把这点摊开,豆包那三句“犟”就不是纯粹的“傻”,而是“过时信源+历史模板+无纠错”三者卡出来的漏洞。

截至发稿,豆包那条评论区早已经改口认了错。但这件事留下的印子不止在这一条回复上——它照出的是当前AI在处理突发事件里几个共通的短板:实时信源接得不够快、历史辟谣模板压得不够细、被纠偏后转得不灵活。对这些短板的修补,比嘲笑某一次“犟嘴”更有用。

新闻纵深

比AI胡说八道更危险的是人们对它的盲目迷信

民主与法制网在一篇谈“AI幻觉”的评论里也提到过:问题的一半在技术狂奔,另一半在责任滞后。部分AI厂商把“什么都能答”当卖点,但“答得对、说得真”这条底线没守住,透支的就会是数字信任。

中国人民大学吴玉章讲席教授刘永谋曾说过:“比‘幻觉'更危险的,是公众对AI的‘迷信'。”很多企业和高大上的宣传让普通人对AI产生了“全能”错觉,但事实上,“在真假很重要的领域,AI生成内容只能作为参考,不能作为判断和决策的最终依据”。

北京大学政府管理学院马亮教授也提醒过:当下生成式AI尚处发展初期,不同模型性能参差不齐,“人类使用AI时务必掌握主动权,切不可盲目信任,沦为其‘傀儡'”。放在这次事件里看,当一条“锋味官宣+港媒讣告+成龙周星驰发文悼念”全线铺开的新闻,还需要用户去纠正AI,本身就说明系统有问题。

技术往前跑的时候,得有人时不时按住它问一句:你确定你看到的是现在,而不是上个月的影子?

上游财经-重庆晨报记者　刘波